

Molecular Descriptors for Chemoinformatics, Volumes I & II

*Roberto Todeschini
Viviana Consonni*

WILEY-VCH

How to go to your page

This eBook contains two volumes. In the printed version of the book, each volume is paginated separately. To avoid duplicate page numbers in the electronic version, we have inserted a volume number before the page number, separated by a hyphen.

For example, to go to page 5 of Volume I, type I-5 in the “page #” box at the top of the screen and click “Go.” To go to page 5 of Volume II, type II-5... and so forth.

*Roberto Todeschini and
Viviana Consonni*

**Molecular Descriptors for
Chemoinformatics**

Methods and Principles in Medicinal Chemistry

Edited by R. Mannhold, H. Kubinyi, G. Folkers

Editorial Board

H. Timmerman, J. Vacca, H. van de Waterbeemd, T. Wieland

Previous Volumes of this Series:

D. A. Smith, H. van de Waterbeemd,
D. K. Walker

Pharmacokinetics and Metabolism in Drug Design,

2nd Ed.

Vol. 31

2006, ISBN 978-3-527-31368-6

T. Langer, R. D. Hofmann (Eds.)

Pharmacophores and Pharmacophore Searches

Vol. 32

2006, ISBN 978-3-527-31250-4

E. Francotte, W. Lindner (Eds.)

Chirality in Drug Research

Vol. 33

2006, ISBN 978-3-527-31076-0

W. Jahnke, D. A. Erlanson (Eds.)

Fragment-based Approaches in Drug Discovery

Vol. 34

2006, ISBN 978-3-527-31291-7

J. Hüser (Ed.)

High-Throughput Screening in Drug Discovery

Vol. 35

2006, ISBN 978-3-527-31283-2

K. Wanner, G. Höfner (Eds.)

Mass Spectrometry in Medicinal Chemistry

Vol. 36

2007, ISBN 978-3-527-31456-0

R. Mannhold (Ed.)

Molecular Drug Properties

Vol. 37

2008, ISBN 978-3-527-31755-4

R. J. Vaz, T. Klabunde (Eds.)

Antitargets

Vol. 38

2008, ISBN 978-3-527-31821-6

E. Ottow, H. Weinmann (Eds.)

Nuclear Receptors as Drug Targets

Vol. 39

2008, ISBN 978-3-527-31872-8

H. van de Waterbeemd,
B. Testa (Eds.)

Drug Bioavailability,

2nd Ed.

Vol. 40

2009, ISBN 978-3-527-31872-8

Roberto Todeschini and Viviana Consonni

Molecular Descriptors for Chemoinformatics

Volume I: Alphabetical Listing

Second, Revised and Enlarged Edition



**WILEY-
VCH**

WILEY-VCH Verlag GmbH & Co. KGaA

Series Editors

Prof. Dr. Raimund Mannhold

Molecular Drug Research Group
Heinrich-Heine-Universität
Universitätsstrasse 1
40225 Düsseldorf
Germany
mannhold@uni-duesseldorf.de

Prof. Dr. Hugo Kubinyi

Donnersbergstrasse 9
67256 Weisenheim am Sand
Germany
kubinyi@t-online.de

Prof. Dr. Gerd Folkers

Collegium Helveticum
STW/ETH Zurich
8092 Zurich
Switzerland
folkers@collegium.ethz.ch

Volume Authors

Prof. Dr. Roberto Todeschini

Dept. of Environm. Sciences
University Milano-Bicocca
Piazza della Scienza 1
0126 Milano
Italy
roberto.todeschini@unimib.it

Dr. Viviana Consonni

Dept. Environm. Sciences
University Milano-Bicocca
Piazza della Scienza 1
20126 Milano
Italy
viviana.consonni@unimib.it

Cover Description

Background: Front of a tablet fragment
of a middle Assyrian code.
Foreground: Drawing of phenylurea.

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

**Bibliographic information published by
the Deutsche Nationalbibliothek**

Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA,
Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into a machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Cover Design Grafik-Design Schulz, Fußgönheim

Typesetting Thomson Digital, Noida, India

Printing betz-druck GmbH, Darmstadt

Binding Litges & Dopf GmbH, Heppenheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN 978-3-527-31852-0

*This book is dedicated with love to
Alessia, Davide, Edoardo, Marco, Milo, Marilena, and Giovanni*

A good scientist should have the imagination of a child, the determination of a boy, the rationality of a man, and the experience of an old man. The difficulty is to have all these qualities at the same time.

R.T.

Any alternative viewpoint with a different emphasis leads to an inequivalent description. There is only one reality but there are many points of view.

It would be very narrow-minded to use only one context: we have to learn to be able imagining points of view.

*Hans Primas in Chemistry, Quantum Mechanics and
Reductionism
(Springer-Verlag, 1981)*

Contents

Volume I

Dedication	V
The Authors	IX
Acknowledgments	X
Preface	XI
A Personal Foreword	XIII
Introduction	XV
Historical Perspective	XXIII
QSAR/QSPR Modeling	XXVII
How to Learn From This Book	XXXIII
User's Guide	XXXVII
Notations and Symbols	XXXIX

Alphabetical Listing

A	1
B	39
C	77
D	179
E	237
F	311
G	325
H	367
I	395
J	425
K	427
L	433
M	475
N	563

O	567
P	573
Q	613
R	637
S	659
T	799
U	833
V	835
W	875
X	951
Y	953
Z	955
Greek Alphabet Entries	961
Numerical Entries	963

Volume II

Bibliography	1
Appendix A	243
Appendix B	245
Appendix C	251

The Authors



Roberto Todeschini is full professor of chemometrics at the Department of Environmental Sciences of the University of Milano-Bicocca (Milano, Italy), where he constituted the Milano Chemometrics and QSAR Research Group. His main research activities concern chemometrics in all its aspects, QSAR, molecular descriptors, multicriteria decision making, and software development. President of the International Academy of Mathematical Chemistry, president of the Italian Chemometric Society, and “ad honorem”

professor of the University of Azuay (Cuenca, Ecuador), he is author of more than 170 publications in international journals and of the books “The Data Analysis Handbook,” by I.E. Frank and R. Todeschini, 1994, and “Handbook of Molecular Descriptors,” by R. Todeschini and V. Consonni, 2000.



Viviana Consonni received her PhD in chemical sciences from the University of Milano in 2000 and is now full researcher of chemometrics and chemoinformatics at the Department of Environmental Sciences of the University of Milano-Bicocca (Milano, Italy). She is a member of the Milano Chemometrics and QSAR Research Group and has 10 years experience in multivariate analysis, QSAR, molecular descriptors, multicriteria decision making, and software development. She is author of more than 40 pub-

lications in peer-reviewed journals and of the book “Handbook of Molecular Descriptors,” by R. Todeschini and V. Consonni, 2000. In 2006, she obtained the International Academy of Mathematical Chemistry Award for distinguished young researchers and, in June 2009, has been elected as youngest Member of the Academy.

Acknowledgments

The idea of producing the book *Molecular Descriptors for Chemoinformatics* was welcomed by several colleagues whom we warmly thank for their suggestions, revisions, bibliographic information, and moral support; we are particularly grateful to Alexander Balaban, Milan Randić, and several members of the International Academy of Mathematical Chemistry.

Particular thanks go also to Maurizio Bruschi, Ugo Cosentino, Mircea Diudea, and Marco Vighi for their help in revising some topics of the book. The Authors gratefully acknowledge the cooperation with and the support of the editorial staff of Wiley-VCH. In particular, we have to thank Nicola Oberbeckmann-Winter, Frank Weinreich, Carola Schmidt, Susanna Pohl, Claudia Nussbeck, Waltraud Wüst. The Authors also warmly thank Raymund Mannhold, editor of the series, for stimulations and timely pressure to complete this book on time.

Finally, since we have been fully absorbed in writing the book for a long time, we would like to heartily acknowledge Davide Ballabio, Andrea Mauri, Alberto Mangano, and Manuela Pavan of our team not only for their help but also for their patience and assistance during this period.

Preface

In 2000, Roberto Todeschini and Viviana Consonni wrote the highly valuable *Handbook of Molecular Descriptors*, part of our series “Methods and Principles in Medicinal Chemistry.” This volume achieved high acceptance among researchers in the field of drug discovery and design. Now, eight years later, the significant developments in the area of molecular descriptors necessitated a rather comprehensive revision.

All new descriptors, QSAR approaches and chemometric strategies proposed since 2000 have been included in this handbook. Several new topics such as biodescriptors, characteristic polynomial-based descriptors, property filters, scoring functions, and cell-based methods have been added. Other topics, such as substructure descriptors, autocorrelation descriptors, delocalization degree indices, weighted matrices, connectivity indices, and so on, have been completely rewritten.

Attention is also paid to recent methods dedicated to virtual screening of libraries of molecules, such as cell-based methods, property filters, and scoring functions.

Special attention has been paid to strategies for generating families of molecular descriptors based on generalization of classical molecular descriptors; dedicated entries are, for instance, Wiener-type indices, Randić-like indices, Balaban-like indices, connectivity-like indices, and variable descriptors.

Several entries have been joined together in larger entries allowing easier readability and comparability among the different molecular descriptors; for example, entries such as matrices of molecules, weighted matrices, substructure descriptors, and vertex degrees, which were enlarged in order to include a lot of definitions. Moreover, some didactical routes are introduced at the beginning of the book to indicate the main entries concerning a topic.

General entries concerning statistical indices, regression parameters, classification parameters, similarity/diversity were completely rewritten trying to give an exhaustive view of the functions used to characterize data, modeling, and similarity/diversity analysis. For example, more than 50 distance and similarity functions have been reported.

Numerical examples (more than 150) and several tables listing molecular descriptors for two benchmark data sets are added to help students and nonexpert readers to comprehend the algorithms better. Indeed, this new edition has been conceived not

only for experts and professional researchers but also for PhD students and young researchers who wish to enter the field of molecular descriptors and related areas, giving special attention to a didactical use of the book and suggesting some possible routes for didactical purposes.

Molecular descriptors implemented in the most common software for descriptor calculation are discussed and bibliographic references have been extended from 3300 to 6400.

The series editors would like to thank Roberto Todeschini and Viviana Consonni for their brilliant work on this second edition. We also want to express our gratitude to Nicola Oberbeckmann-Winter and Frank Weinreich of Wiley-VCH for their valuable contributions to this project.

May 2009

Raimund Mannhold, Düsseldorf
Hugo Kubinyi, Weisenheim am Sand
Gerd Folkers, Zürich

A Personal Foreword

The first idea to collect into a book all the knowledge about molecular descriptors dates back to September 1997, when we were at a meeting of physical chemistry in Taormina. In this beautiful landscape, we had time to think about the several different ways a molecule can be described and how these often derive from different theories developed in noncommunicating research fields.

At the beginning we collected and studied a lot of papers on the topic driven by childish hope to conclude the job in a few months; however, after a lot of days spent in libraries to search for papers and nights to read them, the initial enthusiasm left place to the awareness of the hugeness of information on this topic and the difficulty of organizing it in a systematic fashion. . . . Finally, two years and half later – working full time – we concluded the *Handbook of Molecular Descriptors*.

The book *Molecular Descriptors for Chemoinformatics* consequently derives from the success of our first book, from the need to update it for the huge number of new molecular descriptors produced from 2000 to 2008, and from the awareness to revise several parts of it and to organize differently the new work to made it also usable for didactical purposes.

Milan, May 2009

Roberto Todeschini
Viviana Consonni

Introduction

The effort being made today to organize Knowledge is a way of participating in the evolution of Knowledge itself. The significance of attempting such organization can be looked for in its ability not only to give information but also to create know-how. Knowledge organization provides not only a collection of facts, a store of information, but also a contribution to the growth of Knowledge, knowledge organization being itself one way of doing research. This is the true end of an encyclopedic guide. In effect, to think that the organization of Knowledge is separated from its production is completely arbitrary.

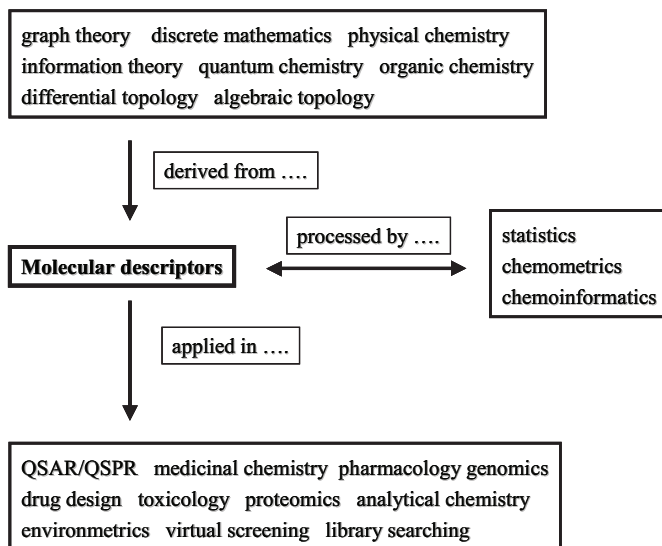
Knowledge should not be considered something given once and for all, based on some final basic theories, but as a *network of models* in progress. This network primarily consists of knots, that is, objects, facts, theories, statements, and models, and the links between the knots are relationships, comparisons, differences, and analogies: such a network is something more than a collection of facts, resulting in a powerful engine for analogical reasoning.

With these purposes in mind, the book *Molecular Descriptors for Chemoinformatics* has been conceived as an encyclopedic guide to molecular descriptors.

Molecular descriptors, tightly connected to the concept of molecular structure, play a fundamental role in scientific research, being the theoretical core of a complex network of knowledge.

Indeed, molecular descriptors are based on several different theories, such as quantum-chemistry, information theory, organic chemistry, graph theory, and so on, and are used to model several different properties of chemicals in scientific fields such as toxicology, analytical chemistry, physical chemistry, and medicinal, pharmaceutical, and environmental chemistry.

Moreover, to obtain reliable estimates of molecular properties, identify the structural features responsible for biological activity, and select candidate structures for new drugs, molecular descriptors are processed by several methods provided by statistics, chemometrics, and chemoinformatics. In particular, chemometrics for about 30 years has been developing classification and regression methods able to provide, although not always, reliable models for both reproducing the known experimental data and predicting the unknown data. The modeling process usually



has not only explanatory purposes but also predictive purposes. The interest in predictive models able to give effective reliable estimates has been largely growing in the last few years as they are more and more considered useful and safer tools for predicting data on chemicals.

It has been nearly 45 years since the QSAR modeling was brought first into the practice of agrochemistry and, successively, in drug design, toxicology, and industrial and environmental chemistry. Its growing importance in the years that followed may be attributed mainly to the rapid and extensive development in methodologies and computational techniques that have allowed to delineate and refine the many variables and approaches used to model molecular properties [Martin, 1979, 1998; Kubinyi, 1993a; Hansch and Leo, 1995; van de Waterbeemd, Testa *et al.*, 1997; Devillers, 1998; Kubinyi, Folkers *et al.*, 1998a, 1998b; Charton and Charton, 2002; Gasteiger, 2003b; Oprea, 2004].

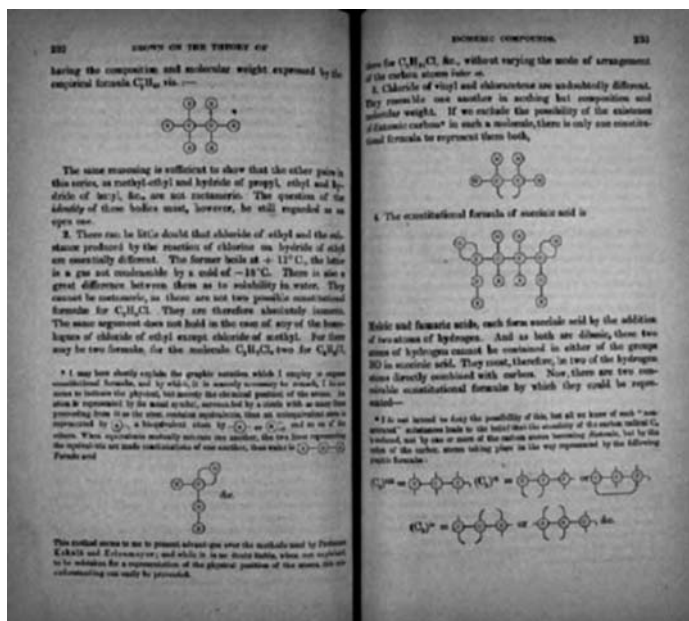
In recent years, "The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization" [Brown, 1998]. In fact, chemoinformatics encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and the use of chemical information [Gasteiger, 2003b; Oprea, 2003]; molecular descriptors play a fundamental role in all these processes being the basic tool to transform chemical information into a numerical code suitable for applying informatic procedures.

Molecular descriptors can be considered as the most important realization of the idea of Crum-Brown. His M.D. Thesis at the University of Edinburgh (1861), entitled "On the Theory of Chemical Combination", shows that he was a pioneer of mathematical chemistry science. In that, he developed a system of graphical

representation of compounds which is basically identical to that used today. His formulae were the first that showed clearly both valency and linking of atoms in organic compounds. Towards the conclusion of his M.D. thesis he wrote:

“It does not seem to me improbable that we may be able to form a mathematical theory of chemistry, applicable to all cases of composition and recomposition.”

In 1864, he published an important study on the “Theory of isomeric compounds” in which, using his graphical formulae, he discussed various types of isomerism [Crum-Brown, 1864] guessing the link between mathematics and chemistry [Crum-Brown, 1867].



The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number or the result of some standardized experiment [Todeschini and Consonni, 2000].

Attention is paid to the term “useful” with its double meaning: it means that the number can give more insight into the interpretation of the molecular properties and/or is able to take part in a model for the prediction of some interesting property of other molecules.

Why must we also accept “or”?

It should not be thought that molecular descriptors are good only if they show an evident link to some information about molecular structure, that is, they are easily interpretable from a structural/chemical point of view.

It often happens that interpretation of molecular descriptors could be weak, provisional, or completely lacking, but their predictive ability or usefulness in application to actual problems should be a strong motive for their use. On the other

hand, descriptors with poor predictive ability may be usefully retained in models when they are theoretically well founded and interpretable due to their ability to encode structural chemical information.

The incompletely realized comprehension of the chemical information provided by molecular descriptors cannot be systematically ascribed to weakness in the descriptors. Actually, our inability to reduce descriptor meaning to well-established chemical concepts is often because newly emergent concepts need new terms in the language and new, hierarchically connected levels for scientific explanation. Thus, what is often considered as scientific failure is sometimes the key to new, useful knowledge.

In any case, all the molecular descriptors must contain, to varying extents, chemical information, must satisfy some basic invariance properties and general requirements, and must be derived from well-established procedures, which enable molecular descriptors to be calculated for any set of molecules. It is obvious – almost trivial – that a single descriptor or a small number of numbers cannot wholly represent the molecular complexity or model all the physico-chemical responses and biological interactions. As a consequence, although we must get used to living with approximate models (*nothing is perfect!*), we have to keep in mind that “approximate” is not a synonym of “useless.”

A molecular descriptor can be thought of as the mythological Dragon on the Babylon Istar Gate (Pergamon Museum of Berlin), which actually is a mixing of several different animals, each corresponding to a different part of the Dragon body; likewise, a molecular descriptor has several different meanings which depend on one's point of view.



Several scientists think that only molecular descriptors derived from quantum-chemistry, which they consider the unique “true” chemical theory, or from simple experimental properties (e.g., partition coefficients or molar refractivity), thought of as the “experimental chemical evidence”, can be legitimately used in QSAR/QSPR modeling. For several years, predictive ability, overfitting, and chance correlation have been not discussed for models derived from those “well-founded” molecular descriptors. On the contrary, a great criticism and skepticism arise against models based on descriptors derived from chemical graph theory, statistics applied to geometrical representations of molecules, and other innovative approaches. In most of the cases, these molecular descriptors give better results than the classical ones but their validity is often brought into question, although it is obvious that chance correlation can be obtained by any kind of descriptor, independent of their interpretability and scientific “nobility.”

The historical development of molecular descriptors reflects some of the distinctive characteristics of the most creative scientists, that is, their capability of being at the same time engaged and/or detached, rational and/or quirky, and serious and/or not so serious. Science is a game and the best players appreciate not only the beauty of a discovery by a precise and logical reasoning but also the taste of making a guess, of proposing eccentric hypotheses, of being doubtful and uncertain when confronted by new and complex problems. Molecular descriptors constitute a research field where the most diverse strategies for scientific discovery can be applied.

Molecular descriptors will probably play an increasing role in science growth. The availability of large numbers of theoretical descriptors that provide diverse sources of chemical information would be useful to better understand relationships between molecular structure and experimental evidence, also taking advantage of more and more powerful methods, computational algorithms, and fast computers. However, as before, deductive reasoning and analogy, theoretical statements and hazardous hypotheses, and determination and perplexity still remain fundamental tools.

The field of molecular descriptors is strongly interdisciplinary and involves a huge number of different theories. For the definition of molecular descriptors, a knowledge of algebra, graph theory, information theory, computational chemistry, and theories of organic reactivity and physical chemistry is usually required, although at different levels. For the use of the molecular descriptors, a knowledge of statistics, chemometrics, chemoinformatics, and the principles of the QSAR/QSPR approaches is necessary in addition to the specific knowledge of the problem. Moreover, programming and sophisticated software and hardware are often inseparable fellow travelers of the researcher in this field.

This book tries to meet the great interest that the scientific community is showing about all the tools that chemoinformatics provides for a quick acquisition and mining of information on chemical compounds and evaluation of their effects on humans and environment in general. Besides the consolidated interest for the quantitative modeling of biological activity, physico-chemical properties, and environmental behavior of compounds, an increasing interest has been shown by the scientific community in recent years in the fields of combinatorial chemistry, high-throughput screening, similarity searching, and database mining, for which several approaches particularly suitable for informatic treatment have been proposed. Thus, several disciplines such as chemistry, pharmacology, environmental protection, drug design, toxicology, and quality control for health and safety, derive great advantages from these methodologies in their scientific and technological development.

The book, *Molecular Descriptors for Chemoinformatics*, collects the definitions, formulas, and short comments of most of the molecular descriptors known in chemical literature. The molecular descriptor definitions, about 3300, are organized in alphabetical order.

The importance of each definition is not related to its length. Only a few old descriptors, abandoned or demonstrated as wrong, were intentionally left out to avoid confusion. An effort was also made to collect appropriate bibliographic information under each definition. We are sorry if any relevant descriptor and/or work has been

missed out; although this has not been done deliberately, we take full responsibility for any omission.

Some molecular descriptors are grouped under a specific topic using a mixed taxonomy based on different points of view, in keeping with the leading idea of the book to promote learning by comparison. These book topics were mainly distinguished according to the *physico-chemical meaning* of molecular descriptors or the specific *mathematical tool* used for their calculation.

Some basic concepts and definitions of statistics, chemometrics, algebra, graph theory, similarity/diversity analysis, which are fundamental tools in the development and application of molecular descriptors, are also discussed in the book in some detail. More attention was paid to information content, multivariate correlation, model complexity, variable selection, applicability domain, and parameters for model quality estimation, as these are the characteristic components of modern QSAR/QSPR modeling.

The book contains nothing about the combinatorial algorithms for the generation and enumeration of chemical graphs, the basic principles of statistics, informatic code for descriptor calculation, or experimental techniques for measuring physico-chemical, technological, and biological responses. Moreover, relevant chemometric methods such as Partial Least Squares regression (PLS) and other regression methods, classification methods, cluster analysis, and artificial neural networks are simply quoted, references are given, but no theoretical aspect is presented. Analogously, computational chemistry methods are quoted only as important tools for theoretical calculations, but no claim is made here to their detailed explanation.

Molecular descriptors on the Web

The authors together with the other members of the Milano Chemometrics and QSAR Research Group of the University of Milano-Bicocca (Milan, Italy) activated in 2007 a web site dedicated to molecular descriptors (<http://www.molecular-descriptors.eu>). This web site aims at promoting information exchange among all the scientists who propose new molecular descriptors and/or apply molecular descriptors in their research.

This web site collects different kinds of information related to molecular descriptors, thus helping researchers in their daily work. *Software*, *books*, *links*, *events*, *tutorials*, and *news* are organized in a systematic way to allow a quick and easy consultation. Moreover, this web site provides a *forum* on molecular descriptors, where experts can initiate discussions on different topics as well as collect lists of bibliographic references about descriptors or discuss their interpretations.

The authors would be grateful to all researchers who would like to send their observations and comments on the book contents, information about new descriptors, and bibliographic references. E-mail submissions can be made at info@molecular-descriptors.eu.

Bibliographic references

The reference list covers a period between 1741 and 2008, lists about 6400 references, for almost 7000 authors and 450 periodicals. Author names are given by the last name, followed by the initials of the first and middle names, if present.

In addition to the cited references, a thematic bibliography with almost 5,000 entries is available. Here, bibliographic references have been collected for some 70 topics of general interest. These references are additional references to those already quoted in the main text of the book. Topics are listed in alphabetic order, from “ADME properties” to “Wiener index”. Selection of the topics was based on the most frequent keywords encountered in publications about molecular descriptors and related research fields.

The thematic bibliography is available as supplementary online material from the book homepage www.wiley-vch.de. Please visit <http://www.wiley-vch.de/publish/en/books/3-527-31852-6> for details.

Historical Perspective

The history of molecular descriptors is closely related to the history of what can be considered one of the most important scientific concepts of the last part of the nineteenth century and the whole twentieth century, that is, the concept of molecular structure.

The years between 1860 and 1880 were characterized by a strong debate on the concept of molecular structure, arising from the studies on substances showing optical isomerism and the studies of Kekulé (1861–1867) on the structure of benzene. The concept of the molecule thought of as a three-dimensional body was first proposed by Butlerov (1861–1865), Wislicenus (1869–1873), Van't Hoff (1874–1875), and Le Bel (1874). The publication in French of the revised edition of *La chimie dans l'espace* by Van't Hoff in 1875 is considered a milestone in the three-dimensional conception of the chemical structures.

QSAR history started a century earlier than the history of molecular descriptors, being closely related to the development of the molecular structure theories.

QSAR modeling was born in toxicology field. Attempts to quantify relationships between chemical structure and acute toxic potency have been part of the toxicological literature for more than 100 years. In the defense of his thesis entitled “Action de l'alcool amylique sur l'organisme” at the Faculty of Medicine, University of Strasbourg, France, on January 9, 1863, Cros noted a relationship existed between the toxicity of primary aliphatic alcohols and their water solubility. This relationship demonstrated the central axiom of structure–toxicity modeling, that is, the toxicity of substances is governed by their properties, which are determined in turn by their chemical structure. Therefore, there are inter-relationships among structure, properties, and toxicity.

Crum-Brown and Fraser (1868–1869) [Crum-Brown, 1864, 1867; Crum-Brown and Fraser, 1868] proposed the existence of a correlation between biological activity of different alkaloids and their molecular constitution. More specifically, the physiological action of a substance in a certain biological system (Φ) was defined as a function (f) of its chemical constitution (C):

$$\Phi = f(C).$$

Thus, an alteration in chemical constitution, ΔC , would be reflected by an effect on biological activity, $\Delta \Phi$. This equation can be considered the first general formulation of a quantitative structure–activity relationship.

The periodic table proposed by Mendeleev [1870] gave relationships between atomic structure and properties; in the following years, the concept of an internal structure of atoms and molecules became more and more relevant and important studies were conducted such as those by G.N. Lewis [Lewis, 1916, 1923].

A hypothesis on the existence of correlations between molecular structure and physico-chemical properties was reported in the work of Körner [1874], which dealt with the synthesis of disubstituted benzenes and the discovery of *ortho*, *meta*, and *para* derivatives: the different colors of disubstituted benzenes were thought to be related to differences in molecular structure and the indicator variables for *ortho*, *meta*, and *para* substitution can be considered as the first three molecular descriptors [Körner, 1869, 1874].

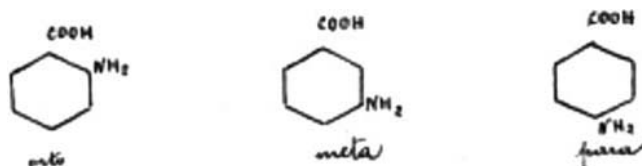


Figure from *Chimica Organica* by W. Körner, 1921, personal document of the Authors.

Ten years later, Mills [Mills, 1884] published in the *Philosophical Magazine* a study “On melting point and boiling point as related to composition.”

The quantitative property–activity models, commonly referred to as those marking the beginning of systematic QSAR/QSPR studies [Richet, 1893], have come out from the search for relationships between the potency of local anesthetics and the oil/water partition coefficient [Meyer, 1899], between narcosis and chain length [Overton, 1901, 1991], and between narcosis and surface tension [Traube, 1904]. In particular, the concepts developed by Meyer and Overton are often referred to as the Meyer–Overton theory of narcotic action [Meyer, 1899; Overton, 1901].

The first theoretical QSAR/QSPR approaches date back to the end of 1940s and are those relating biological activities and physico-chemical properties to theoretical numerical indices derived from the molecular structure.

The $\rightarrow$ *Wiener index* [Wiener, 1947a, 1947b] and the $\rightarrow$ *Platt number* [Platt, 1947], proposed in 1947 to model the boiling point of hydrocarbons, were the first theoretical molecular descriptors based on the graph theory.

In the early 1960s, several other molecular descriptors were proposed, which marked the beginning of systematic studies on molecular descriptors, mainly based on the graph theory [Charton, 1964; Fujita, Iwasa *et al.*, 1964; Gordon and Scantlebury, 1964; Smolenskii, 1964; Spialter, 1964a; Hansch, Deutsch *et al.*, 1965; Reichardt, 1965; Hansch and Anderson, 1967; Balaban and Harary, 1968; Harary, 1969a; Kier, 1971; Cammarata, 1972; Gutman and Trinajstić, 1972; Hosoya, 1972c; Verloop, 1972].

The use of quantum-chemical descriptors in QSAR/QSPR modeling dates back to early 1970s [Kier, 1971], although they actually were conceived several years before to

encode information on relevant properties of molecules in the framework of quantum-chemistry. During 1930–1960, the pioneering studies that signaled the beginning of quantum-chemistry are those of Pauling [Pauling, 1932, 1939] and Coulson [Coulson, 1939] on the chemical bond, of Sanderson on electronegativity [Sanderson, 1952] and of Fukui [Fukui, Yonezawa *et al.*, 1954] and Mulliken on electronic distribution [Mulliken, 1955a].

Once the concept of molecular structure was definitively consolidated by the successes of quantum-chemistry theories and the approaches to the calculation of numerical indices encoding molecular structure information were accepted, all the constitutive elements for the take-off of QSAR strategies were available.

From the Hammett equation [Hammett, 1935, 1937], the seminal work of Hammett gave rise to the “ σ – ρ ” culture in the delineation of substituent effects on organic reactions, whose aim was to search for linear free energy relationships (LFER) [Hammett, 1938]: steric, electronic, and hydrophobic constants were derived for several substituents and used in an additive model to estimate the biological activity of congeneric series of compounds.

In the 1950s, the fundamental works of Taft in physical organic chemistry laid the foundation of relationships between physico-chemical properties and solute–solvent interaction energies (linear solvation energy relationships, LSER), based on steric, polar, and resonance parameters for substituent groups in congeneric compounds [Taft, 1952, 1953a, 1953b].

In the mid-1960s, led by the pioneering works of Hansch [Hansch, Maloney *et al.*, 1962; Hansch, Muir *et al.*, 1963; Fujita, Iwasa *et al.*, 1964], the QSAR/QSPR approach began to assume its modern look.

In 1962, Hansch, Maloney and Fujita [Hansch, Maloney *et al.*, 1962] published their study on the structure–activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity. Using the octanol/water system, a whole series of partition coefficients was measured and, thus, a new hydrophobic scale was introduced for describing the inclination of molecules to move through environments characterized by different degrees of hydrophilicity such as blood and cellular membranes. The delineation of Hansch models led to explosive development in QSAR analysis and related approaches [Hansch and Leo, 1995]. This approach known with the name of *Hansch analysis* became and it still is a basic tool for QSAR modeling.

In the same years, Free and Wilson [Free and Wilson, 1964] developed a model of additive substituent contributions to biological activities, giving a further push to the development of QSAR strategies. They proposed to model a biological response on the basis of the presence/absence of substituent groups on a common molecular skeleton [Free and Wilson, 1964; Kubinyi, 1988b]. This approach, called “de novo approach” when presented in 1964, was based on the assumption that each substituent gives an additive and constant effect to the biological activity regardless of the other substituents in the rest of the molecule.

At the end of 1960s, a lot of structure–property relationships were proposed based not only on substituent effects but also on indices describing the whole molecular structure. These theoretical indices were derived from a topological representation of

molecule, mainly applying the graph theory concepts, and then usually referred to as 2D-descriptors.

The fundamental works of Balaban [Balaban and Harary, 1971; Balaban, 1976a], Randić [Randić, 1974, 1975b], and Kier, Hall *et al.* [Kier, Hall *et al.*, 1975] led to further significant developments in QSAR approaches based on topological indices.

As a natural extension of the topological representation of a molecule, the geometrical aspects of molecular structures were taken into account since the mid-1980s, leading to the development of the 3D-QSAR, which exploits information on the molecular geometry. Geometrical descriptors were derived from the 3D spatial coordinates of a molecule and, among them, there were shadow indices [Rohrbaugh and Jurs, 1987a], charged partial surface area descriptors [Stanton and Jurs, 1990], WHIM descriptors [Todeschini, Lasagni *et al.*, 1994], gravitational indices [Katritzky, Mu *et al.*, 1996b], EVA descriptors [Ferguson, Heritage *et al.*, 1997], 3D-MoRSE descriptors [Schuur, Selzer *et al.*, 1996], EEVA descriptors [Tuppurainen, 1999a], and GETAWAY descriptors [Consonni, Todeschini *et al.*, 2002a].

At the end of 1980s, a new strategy for describing molecule characteristics was proposed, based on molecular interaction fields, which consist of interaction energies between a molecule and probes, at specified spatial points in 3D space. Different probes (such as a water molecule, methyl group, hydrogen, etc.) were used for evaluating the interaction energies in thousands of grid points where the molecule was embedded. As the final result of this approach, a scalar field (a lattice) of interaction energy values characterizing the molecule was obtained. The first formulation of a lattice model to compare molecules by aligning them in 3D space and extracting chemical information from molecular interaction fields was first proposed by Goodford [Goodford, 1985] in the GRID method and then by Cramer, Patterson, Bunce [Cramer III, Patterson *et al.*, 1988] in the Comparative Molecular Field Analysis (CoMFA).

Still based on molecular interaction fields, several other methods were successively proposed and, among them, there were Comparative Molecular Similarity Indices Analysis (CoMSIA) [Klebe, Abraham *et al.*, 1994], Compass method [Jain, Koile *et al.*, 1994], G-WHIM descriptors [Todeschini, Moro *et al.*, 1997], Voronoi field analysis [Chuman, Karasawa *et al.*, 1998], VolSurf descriptors [Cruciani, Pastor *et al.*, 2000], and GRIND descriptors [Pastor, Cruciani *et al.*, 2000].

Finally, the scientific community has been showing an increasing interest in recent years for virtual screening and design of chemical libraries, for which several similarity/diversity approaches, cell-based methods, and scoring functions have been proposed mainly based on *substructure descriptors* such as molecular fingerprints [Gasteiger, 2003b; Kubinyi, 2003b; Oprea, 2004].

QSAR/QSPR Modeling

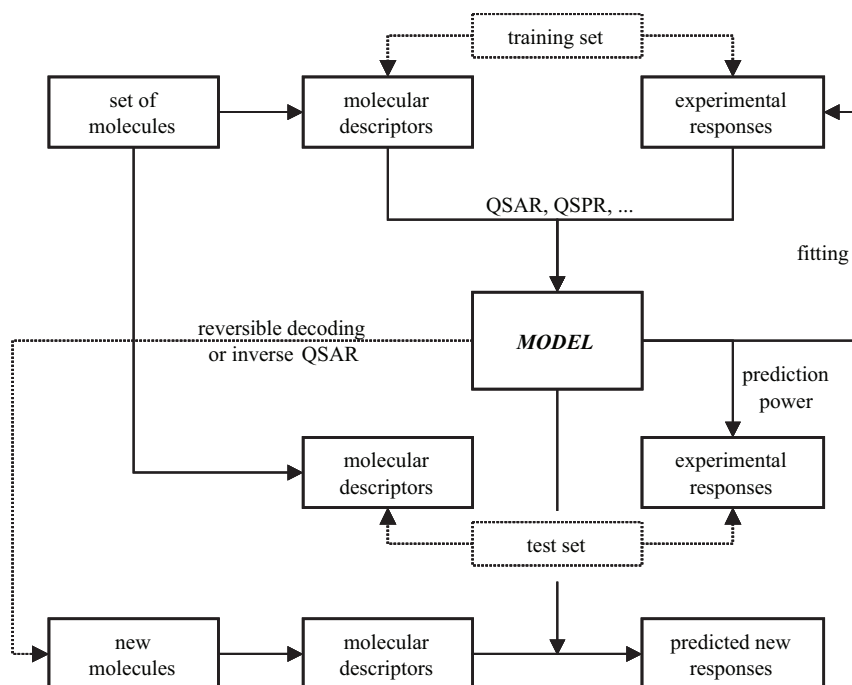
Quantitative Structure–Activity Relationships (QSARs) are the final result of the process that starts with a suitable description of molecular structures and ends with some inference, hypothesis, and prediction on the behavior of molecules in environmental, biological, and physico-chemical systems in analysis.

QSARs are based on the assumption that the structure of a molecule (for example, its geometric, steric, and electronic properties) must contain features responsible for its physical, chemical, and biological properties and on the ability to capture these features into one or more numerical descriptors. By QSAR models, the biological activity (or property, reactivity, etc.) of a new designed or untested chemical can be inferred from the molecular structure of similar compounds whose activities (properties, reactivities, etc.) have already been assessed.

The development of QSAR/QSPR models is a quite complex process.

Once the research goal has been clearly defined, which in most cases means defining the property to be modeled, that is, the end point, the decision to be made concerns how much general the final model should be. This entails the selection of the set of molecules the modeling procedure is applied to. For a long time, QSAR models were developed on sets of congeneric compounds, that is, molecules with a common parental structure and different substituent groups. Later, the interest in producing tools for quick molecular property estimations moved forward more general QSAR models suitable for diverse molecules belonging to different chemical classes, that is, not congeneric sets. The final decision in defining the molecule set mainly depends on the foreseen use of the model and availability of experimental data.

In this phase of the QSAR process, it is of primary concern to gain an exhaustive knowledge about the compounds in analysis with specific regard to the end point of interest. This obviously implies acquisition of reliable experimental data regarding the end point and possibly already existing models. Data of the chemicals can be produced experimentally or retrieved from literature. In both cases, accuracy should be carefully evaluated: the limiting factor in the development of QSAR/QSPR models is the availability of high-quality experimental data, since the accuracy of the property estimated by a model cannot exceed the degree of accuracy of the input data. Moreover, when data are collected from literature to avoid an additional variability



into the data due to different sources of information, data should be taken just from one source or from almost comparable sources.

Another important phase of the QSAR process is the definition of a reliable chemical space or, in other words, the selection of those structural features thought to be the most responsible for modeling the end point in analysis. This implies the selection of proper molecular descriptors but, in most cases, there is no *a priori* knowledge about which molecular descriptors are the best. Then, the tendency is to use a huge number of descriptors, which hopefully include the candidate variables for modeling and later apply a variable selection technique. Two basic strategies can be adopted: (a) the use of algorithms to select the optimal subset(s) of descriptors and (b) the use of chemometric methods (e.g., PCA or PLS) able to condense the large amount of available chemical information into a few principal variables. Before starting to generate quantitative models, relationships between structure and activity of molecules can be qualitatively evaluated by the aid of indices such as SAL index and SAR index, specifically conceived to measure the degree of roughness of the activity landscape in the selected chemical space [Maggiora, 2006]. If there are a number of cliffs, that is, discontinuities, then there are some options available: the chemical space can be changed by selecting a different set of molecular descriptors; nonlinear models can be used instead of the most common linear ones; more compounds need to be sampled in the most discontinuous regions.

Exploratory data analysis is a common preliminary step in all the QSAR/QSPR studies. In particular, Principal Component Analysis (PCA) and clustering methods

(both hierarchical and nonhierarchical) are used most commonly. The clustering approach based on the Kohonen maps (or Self-Organizing Maps), which is an artificial bidimensional neural network providing easy interpretable information on similarity/diversity among objects, has gained wide importance in the last few years [Kohonen, 1989; Zupan, Novič *et al.*, 1995].

By exploratory analysis, the QSAR expert can evaluate if the chosen molecular descriptors are suitable for describing the compounds in analysis and the chemical space is sufficiently represented. Moreover, the tendency observed nowadays is to build a reference chemical space for large categories of chemicals for which molecular properties are known by using methods such as PCA on molecular fingerprints. Then, this chemical space is used to analyze similarities among groups of chemicals showing, for example, different biological activity, and to find which regions in the chemical space require to be more explored by designing new molecular structures. Several methods have been developing to evaluate distributions of compounds in the chemical space; these are mainly cluster-based methods and cell-based methods. Some methods are based on concepts of information theory and statistics useful to derive activity-class specific profiles, scoring functions for selecting the most favorable candidate structures for new drugs, and property filters for screening and designing libraries of compounds.

The majority of the QSAR strategies aimed at building models are based on regression and classification methods, depending on the problem studied. For continuous properties, like most of the biological activities and physico-chemical properties, the typical QSAR/QSPR model is defined as

$$P = f(x_1, x_2, \dots, x_p),$$

where P is the molecular property/activity, $x_1, \dots, x_p$ are the p molecular descriptors, and f is a function representing the relationship between response and descriptors. In most of the cases, the function f is not *a priori* known and needs to be estimated.

Ordinary Least Square regression (OLS), also called Multiple Linear Regression (MLR), is the most common regression technique used to estimate the quantitative relationship between molecular descriptors and the property. Partial Least Squares (PLS) regression is widely applied especially when there are a large number of molecular descriptors with respect to the number of training compounds, as it happens for methods such as GRID and CoMFA.

Regression techniques based on the artificial neural networks are also frequently used [Livingstone and Salt, 1992], often when the relationship between descriptors and activity/property of molecules is not linear.

For discrete molecular properties, such as those defining active/inactive compounds, the typical classification model is defined as

$$C = f(x_1, x_2, \dots, x_p),$$

where C is the class that each object is assigned to under the application of the obtained model, $x_1, \dots, x_p$ are the p molecular descriptors, and f is a function

representing the relationship between class assignment and descriptors. Note also that classification models are quantitative models, only the response C being a qualitative quantity.

Besides the classical Discriminant Analysis (DA) and the k -Nearest Neighbor (k -NN), other classification methods widely used in QSAR/QSPR studies are SIMCA, Linear Vector Quantization (LVQ), Partial Least Squares-Discriminant Analysis (PLS-DA), Classification and Regression Trees (CART), and Cluster Significance Analysis (CSA), specifically proposed for asymmetric classification in QSAR.

In the last few years, ranking methods have also been introduced in the structure–response correlation studies, paying attention to ranking the chemicals instead of reproducing some quantitative property. They are mainly used to build priority list of chemicals [Sabljíć, 1984; Halfon, Galassi *et al.*, 1996; Brüggemann, Pudenz *et al.*, 2001; Carlsen, Sørensen *et al.*, 2001]; however, they were also proposed for modeling purposes [Pavan, Mauri *et al.*, 2004; Pavan, Consonni *et al.*, 2005], for defining new molecular descriptors, and also for numerical characterization of graphical data such as proteomics maps [Randić and Basak, 2002; Todeschini, Ballabio *et al.*, 2007].

A fundamental stage of the QSAR process is model validation. There are a number of validation techniques that allow the evaluation of the effective prediction ability of models. Validation is usually considered the most important requirement for an acceptable QSAR model. The model predictive ability is evaluated dividing the compounds into the training set, that is, the set by which the model is calculated, and the test set, that is, the set of compounds by which the model's predictive ability is evaluated. The partition into training/test sets is performed in different ways, depending on the validation procedure.

Once the model is calculated and properly validated, it can be used to estimate property values for new molecules or obtain information about *mode of action* of a group of compounds or, in general, about which structural features are responsible for a specific behavior of molecules. In the first case, the attention is paid more to obtaining models with the highest predictive ability, regardless of the model interpretability. Indeed, when the aim is to produce data on chemicals, the very important aspect is that the model is as reliable as possible and not the reason why some molecular descriptors were selected in the model.

However, even when the predictive ability of the models was high, the estimated property should be taken carefully because a molecule might be “far” from the model chemical space and, then, the response would be the result of a strong extrapolation, resulting in an unreliable prediction. To cope with this problem, the concept of $\rightarrow$ *applicability domain* of a model came out as a relevant aspect for the evaluation of the prediction reliability.

For some applications, the possibility to obtain information about molecular structure from QSAR/QSPR models is of primary concern. Any procedure capable to reconstruct the molecular structure or fragment starting from selected molecular descriptor values is called $\rightarrow$ *reversible decoding* (or inverse QSAR) [Gordeeva, Molchanova *et al.*, 1990; Kier, Hall *et al.*, 1993; Zefirov, Palyulin *et al.*, 1995; Cho, Zheng *et al.*, 1998; Brüggemann, Pudenz *et al.*, 2001].

Finally, to summarize what is reported above, the development of a QSAR/QSPR model requires three fundamental components: (1) a data set providing experimental measures of a biological activity or property for a group of chemicals (i.e., the dependent variable of the model); (2) molecular descriptors, which encode information about the molecular structures (i.e. the descriptors or the independent variables of the model); and (3) mathematical methods to find the relationships between a molecule property/activity and the molecular structure.

This book focuses on the molecular descriptors, which are discussed in great detail from their beginning up to now; however, some topics concerning model building, drug design, and chemoinformatics in general are also briefly overviewed.

How to Learn From This Book

A great effort was made to give definitions of molecular descriptors by using the same “language” throughout the whole book to make comparison among different molecular descriptors easier. Therefore, a preliminary look at the most important symbols and notations may be very useful. Some standard symbols were chosen for the most important and most occurring quantities; for instance, the symbol δ always indicates the vertex degree, w the weighting scheme for atoms and/or bonds, P an atomic property, and so on. Moreover, the same mathematical entities are reported always using the same type of notation; for example, matrices are denoted by upper-case bold letters and vectors by lower-case bold letters.

This book was conceived not only for experts and professional researchers but also for PhD students and young researchers who wish entering the field of molecular descriptors and related areas.

Then, some **didactical routes** are here proposed to make the comprehension of some important topics easier. Didactical routes include some fundamental entries pertaining to a specific topic whose reading is suggested. Entries are organized into levels of different priority and details.

Molecular descriptors (overview)

first level:

molecular descriptors, constitutional descriptors, count descriptors, graph invariants, vectorial descriptors, geometrical descriptors, ring descriptors, and multiple bond descriptors.

second level:

molecular geometry, molecular graph, size descriptors, steric descriptors, shape descriptors, substituent descriptors, substructure descriptors, grid-based QSAR techniques, autocorrelation descriptors, and variable descriptors.

further levels:

WHIM descriptors, GETAWAY descriptors, EVA descriptors, GRIND descriptors, VolSurf descriptors, quantum-chemical descriptors, molecular transforms, charge descriptors, Charged Partial Surface Area descriptors, delocalization degree indices, chirality descriptors, hydrogen-bonding descriptors, molecular surface, combined descriptors, and symmetry descriptors.

Graph theory and topological indices

first level:

molecular descriptors, graph, molecular graph, algebraic operators, adjacency matrix, distance matrix, incidence matrices, and graph invariants.

second level:

matrices of molecules, local invariants, vertex degree, connectivity indices, walk counts, path counts, self-returning walk counts, Wiener index, Balaban distance connectivity index, Zagreb indices, Laplacian matrix, molecular complexity, equivalence classes, information content, weighted matrices, and variable descriptors.

further levels:

electrotopological state indices, Hosoya Z index, information indices, indices of neighborhood symmetry, spectral indices, characteristic polynomial-based descriptors, detour matrix, Harary indices, resistance matrix, edge adjacency matrix, edge distance matrix, Wiener matrix, Cluj matrices, Szeged matrices, layer matrices, sequence matrices, determinant-based descriptors, TOMOCOMD descriptors, topological charge indices, expanded distance matrices, and ID numbers.

Classical QSAR/QSPR

first level:

Structure/Response Correlations, Hansch analysis, Hammett equation, Free–Wilson analysis, Linear Solvation Energy Relationships, Linear Free Energy Relationships, group contribution methods, substituent descriptors, extrathermodynamic approach, and biological activity indices.

second level:

electronic substituent constants, steric descriptors, lipophilicity descriptors, Sterimol parameters, BC(DEF) parameters, hydrogen-bonding descriptors, Minimal Topological Difference, DARC-PELCO analysis, Distance Geometry.

further levels:

classification parameters, regression parameters.

Drug design

first level:

drug design, similarity/diversity, substructure descriptors, and grid-based QSAR techniques.

second level:

molecular interaction fields, property filters, cell-based methods, scoring functions, and computational chemistry.

further levels:

Molecular Field Topology Analysis, Molecular Shape Analysis, quantum-similarity, Shannon Entropy Descriptors, Electronic-Topological method, Compass method, Comparative

Molecular Moment Analysis, Comparative Receptor Surface Analysis, and topological feature maps.

Model building

first level:

Structure/Response Correlations, data set, chemometrics, statistical indices, Principal Component Analysis, similarity/diversity, validation, variable selection, and variable reduction

second level:

classification parameters, regression parameters, applicability domain, and consensus analysis

further levels:

self-organizing maps and Hasse diagram.

Experimental properties

first level:

experimental measurements, physico-chemical properties, atomic properties, spectra descriptors, chromatographic descriptors, lipophilicity descriptors, Linear Free Energy Relationships, and Hammett equation.

second level:

electric polarization descriptors, electronegativity, electronic descriptors, hydrogen-bonding descriptors, biological activity indices, and environmental indices.

further levels:

volume descriptors, molecular surface, and technological properties.

Recent advances in QSAR strategies

first level:

biodescriptors, chirality descriptors, polymer descriptors, validation techniques, applicability domain, and consensus analysis.

second level:

Membrane Interaction QSAR analysis, weighted matrices, 4D-Molecular Similarity Analysis, cell-based methods, Graph of Atomic Orbitals, 3D-VAIF descriptors, and TAE descriptor methodology.

further levels:

regression parameters, classification parameters, matrices of molecules, and ranking methods.

User's Guide

This book consists of definitions of technical terms in alphabetical order, each technical term being an *entry* of the book and being related to a specific topic.

Each topic is organized in a hierarchical fashion. By following cross-references ($\rightarrow$ and typeset in *italics*), one can easily find all the entries pertaining to a topic even if they are not located together. Starting from the topic name itself, one is referred to more and more specific entries in a top-down manner.

Each entry begins with an entry line. There are three different kinds of entries: *regular*, *referenced*, and *synonym*.

A **regular entry** has its definition immediately after the entry line. A regular entry is typeset in bold face; it is followed by its (ACRONYM and/or SYMBOL), if any, and by its ($\equiv$ *synonym*), if any. For example:

■ **Wiener index** (*W*) ($\equiv$ *Wiener number*, *Wiener path number*)

A **referenced entry** has its definition in the text of another entry. Each referenced entry begins with the bookmark $\triangleright$ and the symbol $\rightarrow$ precedes the name of the regular entry where the definition of the referenced entry is located. For example:

$\triangleright$ **Wiener operator** $\rightarrow$ Wiener-type indices

A **synonym entry** is followed by the symbol " $\equiv$ " and its synonym typeset in *italics*. To find the definition of a synonym entry, if the synonym is a regular entry, one goes directly to the text under the entry line of the synonym word; otherwise, if the synonym is a referenced entry, one goes to the text of the entry indicated by $\rightarrow$. For example:

$\triangleright$ **Wiener number** $\equiv$ *Wiener index*

$\triangleright$ **walk number** $\equiv$ *molecular walk count* $\rightarrow$ walk counts

In the former case, *Wiener number* is the synonym of the *Wiener index*, which is a regular entry while, in the latter, *walk number* is the synonym of *molecular walk count* whose definition is under the entry *walk counts*.

The text of a regular entry may include the definition of one or more referenced entries highlighted in bold face. When there are many referenced entries listed under one regular entry, called a “main” entry, they are often organized in hierarchical fashion, denoting them by the symbol •. The sub-entries can be in either alphabetic or logical order. For example, in the mega entry “steric descriptors,” the first subentries, each followed by the corresponding text, are

■ **steric descriptors** (*main entry*)

- **Taft steric constant** (*subentry*)
- **Charton steric constant** (*subentry*)

Finally, a referenced entry within a subentry has its definition in the text of the subentry. Its entry line contains the regular entry preceded by the symbol → and the subentry denoted by the symbol (⊙ ...). For example:

➤ **WHIM shape** → WHIM descriptors (⊙ global WHIM descriptors)

indicates that the topic “WHIM shape” is defined in the subentry “global WHIM descriptors” of the regular entry “WHIM descriptors.”

In the text of a regular entry, one is referred to other relevant terms by words in italics indicated by →. To reach a complete view of a topic, we highly recommend reading also the definitions of these words in conjunction with the original entry. For example:

■ **count descriptors**

These are simple molecular descriptors based on counting the defined elements of a compound. The most common chemical count descriptors are → *atom number A*, → *bond number B*, → *cyclomatic number C*, → *hydrogen-bond acceptor number*, and → *hydrogen-bond donor number*, → *distance-counting descriptors*, → *path counts*, → *walk counts*, → *atom pairs*, and other related → *substructure descriptors*.

Finally, words in italics not indicated by → in the text of a regular entry (or subentry) denote relevant terms for the topic that are not further explained or whose definition is reported in a successive part of the same entry.

The symbol  at the end of each entry denotes a list of further readings.

We have made a special effort to keep mathematical notation simple and uniform. A collection of the most often appearing symbols is provided below. Moreover, a list of acronyms, provided in Vol. II – Appendix B, helps decipher and locate the all terminologies given in the book.

Notations and Symbols

The notations and symbols used in the book are listed below. In some cases, slightly different notations with respect to those proposed by the authors are used to avoid confusion with other descriptors and quantities.

Basic symbols

$\mathcal{D}$	molecular descriptor
$\mathcal{M}$	molecule, compound
$\mathcal{G}$	graph, molecular graph
$M\mathcal{G}$	multigraph, molecular multigraph
$\mathcal{A}$	atom type
$P, \mathcal{P}$	property
Φ	physico-chemical property
$\mathcal{L}_i$	local vertex invariant
$\mathcal{L}_{ij}$	local edge invariant
m	atomic mass
Z	atomic number
Z^v	valence electron number
L	principal quantum number
R	radius (covalent, atomic, etc.)

Sets

V	set of vertices of a graph
E	set of edges of a graph
$\mathcal{F}$	set of fragments of a molecule partition
${}^m\mathcal{P}_{ij}$	set of paths of order m from the i th to the j th atom
${}^m\mathcal{P}$	set of paths of order m
$\{a, b, \dots\}$	set of elements

Counts

A	number of the atoms of a molecule
B	number of the bonds of a molecule
C	number of the cycles of a molecule
C^+	number of the cycles with overlapping of a molecule
G	number of equivalence classes
n	number of objects, data, molecules
p	number of variables
M	number of significant principal components or latent variables
N	generic number of elements
N_X	number of atoms, groups, fragments of X-type
n_x	number of elements with an x -value
h_i	number of hydrogens bonded to the i th atom
${}^mP$	total number of paths of length m in the graph
${}^kf$	total number of topological distances equal to k in the graph
${}^kf_i$	number of topological distances of k from the i th vertex to any other vertex in the graph

Matrix operators

VS	row sum operator
CS	column sum operator
IB	Ivanciuc–Balaban operator
Wi	Wiener operator
$Wi^\perp$	orthogonal Wiener operator
$H_Y Wi$	hyper-Wiener operator

Indices and characteristic symbols

b	index for a bond
g	index for equivalence classes
i, j, k, l, f, m, s, t	generic indices
x, y, z	spatial coordinates
d	topological distances
r	geometric distances
d_{st}	distance between objects s and t
s_{st}	similarity between objects s and t
D	dimension (0,1,2,3,4)
D	graph diameter
R	graph radius
π	bond order
q	atomic charge
w	statistical weights, weighting scheme
p	probability

λ	eigenvalue
$\lambda_m(\mathbf{M})$	m th eigenvalue from the matrix $\mathbf{M}$
ℓ_{jm}	PCA loadings of the m th component for the j th variable
t_{im}	i th score of the m th component from PCA or PLS
$[\mathbf{M}]_{ij}, m_{ij}$	i - j element of the matrix $\mathbf{M}$
$Ch(\mathbf{M}, x)$	characteristic polynomial of matrix $\mathbf{M}$
$\mathbf{I}$	identity matrix
$\mathbf{U}$	unit matrix
I	binary or indicator variable
I	information content
δ	vertex degree
ε	edge degree
η	atom eccentricity
σ	vertex distance degree
$\mathbf{a}, \mathbf{b}, \dots$	column vectors
$\mathbf{a}^T, \mathbf{b}^T, \dots$	row vectors

A

- A_{xi} eigenvalue indices → spectral indices
- **Abraham–Klamt descriptors** → Linear Solvation Energy Relationships
- **Abraham's general equation** → Linear Solvation Energy Relationships
- **absolute hardness** → quantum-chemical descriptors (⊙ hardness indices)
- **absorption parameter** → property filters (⊙ drug-like indices)
- **Acceptable Daily Intake** → biological activity indices (⊙ toxicological indices)
- **acceptor superdelocalizability** $\equiv$ *electrophilic superdelocalizability* → quantum-chemical descriptors
- **ACCS** $\equiv$ *Activity Class Characteristic Substructures* → substructure descriptors (⊙ structural keys)
- **ACC transforms** $\equiv$ *Auto-Cross-Covariance transforms* → autocorrelation descriptors
- **ACD/log P** → lipophilicity descriptors
- **ACGD index** → charged partial surface area descriptors
- **acid dissociation constant** → physico-chemical properties (⊙ equilibrium constants)
- **activation energy index** → quantum-chemical descriptors (⊙ highest occupied molecular orbital energy)
- **activation hardness** → quantum-chemical descriptors (⊙ hardness indices)
- **Activity Class Characteristic Substructures** → substructure descriptors (⊙ structural keys)
- **acyclic graph** $\equiv$ *tree* → graph
- **acyclic polynomial** $\equiv$ *matching polynomial* → Hosoya Z-index

■ ADAPT descriptors

ADAPT descriptors [Jurs, Chou *et al.*, 1979; Jurs, Hasan *et al.*, 1988], implemented in the homonymous software ADAPT (Automated Data Analysis and Pattern Recognition Toolkit), fall into three general categories: → *topological indices*, → *geometrical descriptors* (including → *principal moments of inertia*, → *volume descriptors*, and → *shadow indices*), and → *electronic descriptors* (including partial atomic charges and the → *dipole moment*); moreover, → *molecular weight*, → *count descriptors*, and a large number of → *substructure descriptors* are also generated. In addition, the → *charged partial surface area descriptors* constitute a fourth class of descriptors derived by combining electronic and geometrical information.

ADAPT software allows (a) molecular descriptor generation; (b) objective feature selection to discard descriptors that contain redundant or minimal information; and (c) multiple regression

analysis by genetic algorithm or simulated annealing variable selection, or computational $\rightarrow$ *artificial neural networks*.

Several molecular properties have been modeled by ADAPT descriptors, such as *biological activities* [Henry, Jurs *et al.*, 1982; Jurs, Hasan *et al.*, 1983; Jurs, Stouch *et al.*, 1985; Walsh and Claxton, 1987; Wessel, Jurs *et al.*, 1998; Eldred, Weikel *et al.*, 1999; Eldred and Jurs, 1999; Patankar and Jurs, 2000, 2002; He, Jurs *et al.*, 2003, He *et al.*, 2005; Benigni, 2005]; *boiling point* [Smeeks and Jurs, 1990; Stanton, Jurs *et al.*, 1991; Stanton, Egolf *et al.*, 1992; Egolf and Jurs, 1993a; Egolf, Wessel *et al.*, 1994; Wessel and Jurs, 1995a, 1995b; Goll and Jurs, 1999a; Stanton, 2000]; *chromatographic indices* [Anker, Jurs *et al.*, 1990; Sutter, Peterson *et al.*, 1997]; *aqueous solubilities* [Dunnivant, Elzerman *et al.*, 1992; Nelson and Jurs, 1994; Sutter and Jurs, 1996; Mitchell and Jurs, 1998a], *critical temperature and pressure* [Turner, Costello *et al.*, 1998]; *ion mobility constants* [Wessel and Jurs, 1994; Wessel, Sutter *et al.*, 1996]; *reaction rate constants* [Bakken and Jurs, 1999a, 1999b]; and other various properties [Egolf and Jurs, 1992; Russell, Dixon *et al.*, 1992; Stanton and Jurs, 1992; Egolf and Jurs, 1993b; Engelhardt and Jurs, 1997; Mitchell and Jurs, 1997; Goll and Jurs, 1999b; Johnson and Jurs, 1999; Kauffman and Jurs, 2000, 2001a; Mattioni and Jurs, 2003].

- **additivity model** $\equiv$ *Free–Wilson model* $\rightarrow$ Free–Wilson analysis
- **additive adjacency matrix** $\rightarrow$ adjacency matrix
- **additive chemical adjacency matrix** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **additive–constitutive models** $\rightarrow$ group contribution methods
- **additive model of inductive effect** $\rightarrow$ electronic substituent constants ($\odot$ inductive electronic constants)
- **ADI** $\equiv$ *Acceptable Daily Intake* $\rightarrow$ biological activity indices ($\odot$ toxicological indices)
- **adjacencies** $\rightarrow$ graph

■ **adjacency matrix (A)** ($\equiv$ *vertex adjacency matrix*)

The adjacency matrix **A** is one of the fundamental $\rightarrow$ *graph theoretical matrices*; it represents the whole set of connections between adjacent pairs of atoms [Trinajstić, 1992]. The entries a_{ij} of the matrix equal 1 if vertices v_i and v_j are adjacent (i.e., the atoms i and j are bonded) and zero otherwise:

$$[A]_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

where $E(G)$ is the set of the graph edges.

This is the classical definition of the adjacency matrix, which refers to a $\rightarrow$ *simple graph*, where multiple bonds are not accounted for. The adjacency matrix is symmetric with dimension $A \times A$, where A is the number of atoms and it is usually derived from a $\rightarrow$ *H-depleted molecular graph*.

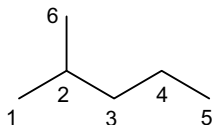
The i th row sum of the adjacency matrix is called $\rightarrow$ *vertex degree*, denoted by δ_i and defined as

$$\delta_i \equiv VS_i(A) = \sum_{j=1}^A a_{ij}$$

where VS is the $\rightarrow$ *vertex sum operator*. The vertex degree represents the number of σ bonds of the i th atom.

Example A1

Adjacency matrix **A** and vertex degrees δ_i of 2-methylpentane.



A =

Atom	1	2	3	4	5	6	δ_i
1	0	1	0	0	0	0	1
2	1	0	1	0	0	1	3
3	0	1	0	1	0	0	2
4	0	0	1	0	1	0	2
5	0	0	0	1	0	0	1
6	0	1	0	0	0	0	1

The **total adjacency index** A_V is a measure of the graph connectedness and is calculated as the sum of all the entries of the adjacency matrix of a molecular graph, which is twice the number B of graph edges [Harary, 1969a; Rouvray, 1983]:

$$A_V = \sum_{i=1}^A \sum_{j=1}^A a_{ij} = \sum_{i=1}^A \delta_i = 2 \cdot B$$

For example, the total adjacency index of 2-methylpentane is $A_V = 1 + 3 + 2 + 2 + 1 + 1 = 10$, which is twice the number of edges equal to five in the H-depleted molecular graph of this molecule. Therefore, the number of entries equal to 1 in the adjacency matrix is $2B$, while the number of entries equal to zero is $A^2 - 2B$; in particular, for acyclic graphs the total number of entries equal to 1 is $2(A - 1)$ and the number of entries equal to zero is $A^2 - 2(A - 1)$; for monocyclic graphs, the values are $2A$ and $A^2 - 2A$, respectively. The total adjacency index is sometimes calculated as the half sum of the adjacency matrix elements.

The global connectivity of a graph can also be characterized by the average of the total adjacency index as [Bonchev and Buck, 2007]

$$\bar{\delta} = \frac{A_V}{A} = \frac{2B}{A}$$

where A is the number of graph vertices. If calculated from the $\rightarrow$ H-filled molecular graph, this average index is one of the two $\rightarrow$ Schafli indices, called connectivity.

The doubly normalized total adjacency index is called **density index** and is defined as

$$\bar{\bar{\delta}} = \frac{A_V}{A \cdot (A-1)} = \frac{2B}{A \cdot (A-1)} \quad \text{or} \quad \bar{\bar{\delta}} = \frac{A_V}{A^2} = \frac{2B}{A^2}$$

The adjacency matrix is one important source of molecular descriptors. Simple $\rightarrow$ topological information indices can be calculated on both the equality and the magnitude of adjacency matrix elements. $\rightarrow$ Walk counts and $\rightarrow$ self-returning walk counts that coincide with the spectral moments of the adjacency matrix are calculated by the increasing powers of the adjacency matrix [McKay, 1977; Jiang, Tang *et al.*, 1984; Hall, 1986; Kiang and Tang, 1986; Jiang and Zhang, 1989, 1990; Marković and Gutman, 1991; Jiang, Qian *et al.*, 1995; Marković and Stajkovic, 1997; Marković, 1999].

$\rightarrow$ Spectral indices, $\rightarrow$ determinant-based descriptors, and $\rightarrow$ characteristic polynomial-based descriptors of the adjacency matrix are largely used in QSAR modeling.

The **clustering coefficient of a vertex**, denoted as C_i , is a local vertex invariant derived from the adjacency matrix by considering the first-neighbor interconnectivity [Bonchev and Buck, 2007]. It was proposed as a measure of the clustered structure of a graph around a vertex and is defined as

$$C_i = \frac{2 \cdot b_i}{\delta_i \cdot (\delta_i - 1)} \quad 0 \leq C_i \leq 1$$

where b_i is number of edges between the first neighbors of the i th vertex, measuring to what extent the first neighbors of the i th vertex are linked between themselves:

$$b_i = \frac{1}{2} \cdot \sum_{j=1}^A a_{ij} \cdot \sum_{m=1}^A a_{jm} \cdot a_{mi} \quad m \neq i$$

where A is the number of vertices, and a_{ij} , a_{jm} , and a_{mi} are the elements of the adjacency matrix. Then, the terms a_{ij} of the first summation are equal to 1 only for the vertices v_j , which are connected to the i th vertex, while terms $a_{jm} \cdot a_{mi}$ in the second summation are equal to 1 for those pairs of vertices v_j and v_m , which are contemporarily neighbors of the i th vertex and are bonded to each other, and are zero otherwise.

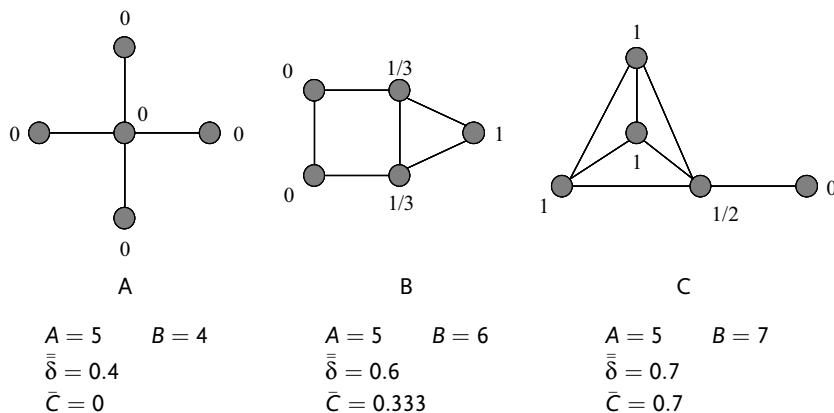
The **overall degree of clustering of a graph** is given by [Bonchev and Buck, 2007]

$$\bar{C} = \frac{1}{A} \cdot \sum_{i=1}^A C_i$$

It should be noted that clustering around a vertex is possible only in trimembered cycles; in all other structures, there are no edges between the first-neighbor vertices, for example, $b_i = 0$.

Example A2

Calculation of the density index $\bar{\delta}$ and overall degree of clustering $\bar{C}$ for graphs A, B, and C. Each vertex is labeled with its clustering coefficient.



To account for multiple bonds, $\rightarrow$ *atom connectivity matrix*, $\rightarrow$ *adjacency matrix of a multigraph*, and $\rightarrow$ *adjacency matrix of a general graph* can be used instead of the adjacency matrix of a simple graph. To account for heteroatoms, different $\rightarrow$ *weighted adjacency matrices* were proposed, such as the $\rightarrow$ *augmented adjacency matrix* and $\rightarrow$ *chemical adjacency matrix*.

The **additive adjacency matrix** is derived from the adjacency matrix substituting row elements equal to 1, corresponding to pairs of adjacent vertices, with the vertex degrees of the connected vertices as

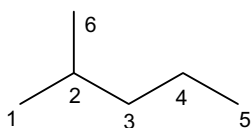
$$[\delta \mathbf{A}]_{ij} = \begin{cases} \delta_j & \text{if } (i,j) \in \mathbf{E}(G) \\ 0 & \text{otherwise} \end{cases}$$

where δ_j is the vertex degree of the j th vertex connected to the i th vertex.

This matrix is a special case of $\rightarrow$ *distance degree matrices* obtained by the parameter combination $\alpha = 0$, $\beta = 0$, $\gamma = 1$. The row sum of the additive adjacency matrix is the $\rightarrow$ *extended connectivity* of first-order EC^1 defined by Morgan. This local invariant was used to calculate the $\rightarrow$ *eccentric adjacency index*. A modification of this matrix, which accounts for heteroatoms, is the $\rightarrow$ *additive chemical adjacency matrix*.

Example A3

Additive adjacency matrix $\delta \mathbf{A}$ and extended connectivities EC^1 of 2-methylpentane.



$\delta \mathbf{A} =$

Atom	1	2	3	4	5	6	EC_i^1
1	0	3	0	0	0	0	3
2	1	0	2	0	0	1	4
3	0	3	0	2	0	0	5
4	0	0	2	0	1	0	3
5	0	0	0	2	0	0	2
6	0	3	0	0	0	0	3

Other topological matrices are derived from the adjacency matrix, such as $\rightarrow$ *Laplacian matrix* and the powers of the adjacency matrix used to obtain walk counts and the corresponding molecular descriptors.

The **fragmental adjacency matrix** ${}^m \mathbf{A}_F$ of m th order is a generalization of the adjacency matrix $\mathbf{A}$, which encodes information about adjacencies of the K fragments of the same m th order (i.e., the same number m of edges) contained in the molecular graph instead of adjacencies between vertices [Guevara, 1999]. This matrix is a square symmetric ($K \times K$) matrix whose elements are different from zero only if two fragments i and j are adjacent, that is, they have $m - 1$ edges in common. The **fragmental degree** is defined in the same way as the vertex degree, that is, the row sum of the fragmental adjacency matrix, and, therefore, represents the number of fragments adjacent to the fragment considered. Then, by using the fragmental degree in place of the vertex degree, the **fragmental connectivity index** can be calculated in the same way as the $\rightarrow$ *Randić connectivity index*.

Moreover, the adjacency matrix can be transformed into a **decimal adjacency vector**, denoted by $\mathbf{a}^{10}$, of A elements each being a local vertex invariant obtained by the following expression [Schultz and Schultz, 1991]:

$$a_i^{10} = (2 \cdot a_{i1})^{A-1} + (2 \cdot a_{i2})^{A-2} + \cdots + (2 \cdot a_{iA})^0$$

where a_{ij} is the j th column element of the i th row of the adjacency matrix $\mathbf{A}$. In this way, the information contained in the adjacency matrix is compressed into an A -dimensional vector. For

example, a row of the adjacency matrix equal to [0 1 1 1 0] gives a value of 14, obtained as

$$a_i^{10} = (2 \cdot 0)^{5-1} + (2 \cdot 1)^{5-2} + (2 \cdot 1)^{5-3} + (2 \cdot 1)^{5-4} + (2 \cdot 0)^0 = 14$$

The elements of the decimal adjacency vector are integers that were used for $\rightarrow$ *canonical numbering* of molecular graphs [Randić, 1974].

From the decimal adjacency vector, three different indices were proposed as molecular descriptors:

- (a) the sum of the elements of the vector $\mathbf{a}^{10}$, that is,

$$A1 = \sum_{i=1}^A a_i^{10}$$

- (b) the sum of the linear combination of vertex degrees δ_i , each weighted by the corresponding decimal adjacency vector element a_i^{10} , that is,

$$A2 = \sum_{i=1}^A \delta_i \cdot a_i^{10}$$

- (c) the sum of the elements of the A -dimensional vector $\mathbf{d}$ obtained by multiplying the topological $\rightarrow$ *distance matrix* $\mathbf{D}$ by the decimal adjacency vector, that is,

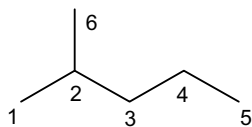
$$A3 = \sum_{i=1}^A [\mathbf{d}]_i$$

where the vector $\mathbf{d}$ is calculated as

$$\mathbf{d} = \mathbf{D} \cdot \mathbf{a}^{10}$$

Example A4

Decimal adjacency vector of 2-methylpentane and related molecular descriptors.



Atom	1	2	3	4	5	6		a_i^{10}		d_i
1	0	1	2	3	4	2		16		159
2	1	0	1	2	3	1		41		84
3	2	1	0	1	2	2	$\times$	20	$=$	123
4	3	2	1	0	1	3		10		202
5	4	3	2	1	0	4		4		301
6	2	1	2	3	4	0		16		159

$$a_1^{10} = (2 \times 1)^{6-2} = 16$$

$$a_2^{10} = (2 \times 1)^{6-1} + (2 \times 1)^{6-3} + (2 \times 1)^{6-6} = 41$$

$$a_3^{10} = (2 \times 1)^{6-2} + (2 \times 1)^{6-4} = 20$$

$$a_4^{10} = (2 \times 1)^{6-3} + (2 \times 1)^{6-5} = 10$$

$$a_5^{10} = (2 \times 1)^{6-4} = 4$$

$$a_6^{10} = (2 \times 1)^{6-2} = 16$$

$$A1 = 16 + 41 + 20 + 10 + 4 + 16 = 107$$

$$A2 = 1 \times 16 + 3 \times 41 + 2 \times 20 + 2 \times 10 + 1 \times 4 + 1 \times 16 = 219$$

$$A3 = 159 + 84 + 123 + 202 + 301 + 159 = 1028$$

- **adjacency matrix of a general graph** → weighted matrices (⊙ weighted adjacency matrices)
- **adjacency matrix of a multigraph** → weighted matrices (⊙ weighted adjacency matrices)
- **adjacency plus distance matrix** → Schultz molecular topological index
- **adjacent eccentric distance sum index** → eccentricity-based Madan indices (⊙ Table E1)
- **adjusted R^2** → regression parameters
- **adjusted retention time** → chromatographic descriptors (⊙ retention time)
- **admittance matrix** $\equiv$ *Laplacian matrix*
- **ADME properties** → drug design

■ adsorbability index (AI)

An empirical molecular descriptor derived from a → *group contribution method* based on molecular refractivity to predict activated carbon adsorption of 157 compounds [Abe, Tatsumoto *et al.*, 1986]. This index was also applied to predict the → *soil sorption partition coefficient* of the same 157 compounds [Okouchi and Saegusa, 1989; Okouchi, Saegusa *et al.*, 1992].

The adsorbability index is calculated by the expression

$$AI = \sum_i f_i \cdot N_i + \sum_j c_j$$

where the summations run over atomic and functional groups; f_i indicates the contribution to activated carbon adsorption of the i th atom- or group type and N_i the number of atoms or groups of type i ; c_j represents a special correction factor accounting for functional group effects.

Atomic and group contributions and correction factors are reported in Table A1.

Table A1 Values of f and c factors proposed by Abe, *et al.* [Abe, Tatsumoto *et al.*, 1986].

Atom/group	f	Group	c
C	0.26	Aliphatic:	
H	0.12	–OH (alcohols)	–0.53
N	0.26	–O– (esters)	–0.36
O	0.17	–CHO (aldehydes)	–0.25
S	0.54	N (amines)	–0.58
Cl	0.59	–COOR (esters)	–0.28
Br	0.86	>C=O (ketones)	–0.30
NO ₂	0.21	–COOH (fatty acids)	–0.03
–C=C–	0.19		
Iso	–0.12	α-Amino acids	–1.55
Tert	–0.32		
Cyclo	–0.28	All groups in aromatics	0

For example, for benzene

$$AI = 6 \times f_C + 6 \times f_H + 3 \times f_{C=C} = 6 \times 0.26 + 6 \times 0.12 + 3 \times 0.19 = 2.85;$$

$$\text{for 1,1,2-trichloroethane } AI = 2 \times f_C + 3 \times f_H + 3 \times f_{Cl} = 2 \times 0.26 + 3 \times 0.12 + 3 \times 0.59 = 2.65$$

- **AEI indices** → spectral indices (⊙ A_{xi} eigenvalue indices)
- **AFC method** $\equiv$ *KOWWIN* → lipophilicity descriptors

■ affinity fingerprints

Affinity fingerprints are → *vectorial descriptors* of molecules either comprising their binding affinities and docking scores or superpositioning pseudoenergies against a reference panel of uncorrelated proteins or small drug molecules [Briem and Lessel, 2000]. These molecular descriptors can be used both for high-throughput compound screening and the → *similarity/diversity* analysis and for the prediction of biological activities of compounds.

In contrast to most other molecular descriptors, affinity fingerprints are not directly derived from molecular structures.

In vitro affinity fingerprints are based on binding affinities, experimentally determined, and can be used to estimate general cross-reactivity and, then, possible toxicity in the drug design process [Weinstein, Kohn *et al.*, 1992; Kauvar, Higgins *et al.*, 1995; Weinstein, Myers *et al.*, 1997; Dixon and Villar, 1998]. The underlying assumption is that compounds binding similarly to all the proteins in the reference panel are likely also to have similar affinity to their target receptor.

Virtual affinity fingerprints (or **in silico affinity fingerprints**) are derived by computational methods and, thus, are vectorial descriptors where experimentally determined binding affinities of molecules are replaced by some calculated scores with respect to the reference panel [Briem and Lessel, 2000].

Some *virtual* affinity fingerprints are explained below.

DOCKSIM fingerprints are derived by computational docking of the molecules into binding pockets of protein structures solved by X-ray crystallography [Briem and Kuntz, 1996]. Therefore, they are 3D vectorial descriptors collecting the docking scores (**DOCK scores**) with respect to the protein-binding site of the reference panel; the scores are obtained by rigid docking of the molecules, and the reference panel contains eight uncorrelated and arbitrarily selected protein structures.

Flexsim-X fingerprints are vectors of docking scores as the DOCKSIM fingerprints, but the scores are obtained by flexible docking of molecules, and the reference panel contains around 40 protein structures, optimized by systematic and genetic algorithm (GA)-based procedures [Lessel and Briem, 2000].

Flexsim-S fingerprints are *virtual* affinity fingerprints where the docking scores are replaced by the superpositioning pseudoenergies, which measure the alignment quality of ligands onto a set of small reference molecules [Lemmen, Lengauer *et al.*, 1998]. Also for Flexsim-S fingerprints, size and composition of the reference panel should be properly optimized.

Flexsim-R fingerprints are *virtual* affinity fingerprints specifically designed for similarity assessments of small fragments, such as R-groups of combinatorial libraries [Weber, Teckentrup *et al.*, 2002].

Molecular hashkeys are another kind of *virtual* affinity fingerprints derived from surface-based comparisons of ligands with a reference panel comprising small, drug-like molecules instead of proteins [Ghuloum, Sage *et al.*, 1999]. A molecular hashkey is a vectorial descriptor of fixed dimension that captures information about the surface properties of a molecule. The elements of the hashkey of a molecule are values of its molecular surface similarity to a set of basis molecules in low-energy-fixed conformations. The molecule is flexibly aligned to each of the set of basis molecules to maximize molecular surface similarity.

- **AH weighting scheme** → weighting schemes
- **Aihara resonance energy** → delocalization degree indices
- **AI indices** → atom-type-based topological indices

■ AIM theory ($\equiv$ Atoms-in-Molecules theory)

Bader's Atoms-in-Molecules (AIM) quantum theory [Bader, 1990] provides a bridge between quantum chemistry and chemical concepts and the framework for reconstructing large molecules from a number of small electron density fragments. In the AIM theory, the electron density of a molecule is partitioned into distinct electron density basins, that is, regions occupied by the corresponding atoms, each containing an atomic nucleus. These electron density atomic fragments are essentially bounded by surfaces of zero net flux in the electron density.

An atomic property P can be then expressed as the integral of the corresponding property density ρ over an atomic region Ω as

$$P(\Omega) = \int_{\Omega} \rho_P d\tau$$

These atomic properties possess a high degree of transferability from the electronic environment in one molecule to another molecule with similar environments. Consequently, the properties of a whole molecule or a functional group can be obtained by adding the atomic properties as

$$P(\text{molecule}) = \sum_{\Omega} P(\Omega).$$

Based on the AIM theory are $\rightarrow$ TAE descriptors and the $\rightarrow$ delocalization index DI.

📖 [Song, Breneman *et al.*, 2002; Lamarche and Platts, 2003; Chaudry and Popelier, 2004; Krygowski, Ejsmont *et al.*, 2004]

- **Akaike Information Criterion** $\rightarrow$ regression parameters
- **alert indices** $\rightarrow$ property filters

■ algebraic operators

Algebraic operators play a meaningful role in the framework of $\rightarrow$ molecular descriptors, since they represent the fundamental mathematical tool used to transform into single numerical quantities the information encoded in $\rightarrow$ matrices of molecules.

Let $\mathbf{M}$ be a generic matrix with n rows and p columns, denoted as

$$\mathbf{M} \equiv [m_{ij}] = \begin{vmatrix} m_{11} & m_{12} & \dots & \dots & m_{1p} \\ \vdots & & & & \vdots \\ m_{n1} & m_{n2} & \dots & \dots & m_{np} \end{vmatrix}$$

The matrix elements m_{ij} are commonly denoted as

$$m_{ij} \equiv [\mathbf{M}]_{ij} \equiv (i, j)$$

A column vector $\mathbf{v}$ is a special case of matrix having n rows and one column; the row vector $\mathbf{v}^T$ is a special case of matrix having one row and p columns.

Some definitions of matrix algebra [Ledermann and Vajda, 1980; Golub and van Loan, 1983; Mardia, Kent *et al.*, 1988], algebraic operators and set theory are given below.

- **characteristic polynomial**

Let $\mathbf{M}$ be a square matrix ($n \times n$) and x a scalar variable, the characteristic polynomial Ch is defined as

$$Ch(\mathbf{M}, x) = \det(\mathbf{M} - x\mathbf{I}) = \sum_{i=0}^n a_i \cdot x^{n-i}$$

where $\mathbf{I}$ is the identity matrix, that is, a matrix having the diagonal elements equal to 1 and all the off-diagonal elements equal to zero, and a_i the polynomial coefficients. The characteristic polynomial is obtained by expanding the determinant and, then, collecting terms with equal powers of x .

The **eigenvalues** λ of the matrix $\mathbf{M}$ are the n roots of its characteristic polynomial, and the set of the eigenvalues is called **spectrum of a matrix**, denoted as $\Lambda(\mathbf{M})$.

Determinant and trace of $\mathbf{M}$ are given by the following expressions:

$$\det(\mathbf{M}) = a_n = \prod_{i=1}^n \lambda_i \quad \text{tr}(\mathbf{M}) = a_1 = \sum_{i=1}^n \lambda_i$$

respectively, where a_n and a_1 are the characteristic polynomial coefficients corresponding to i equal to n and 1, respectively.

For each eigenvalue λ_i , there exists a nonzero vector $\mathbf{v}_i$ satisfying the following relationship:

$$\mathbf{M} \cdot \mathbf{v}_i = \lambda_i \cdot \mathbf{v}_i.$$

The n -dimensional vectors $\mathbf{v}_i$ are called **eigenvectors** of $\mathbf{M}$.

A large number of $\rightarrow$ *characteristic polynomial-based descriptors* and $\rightarrow$ *spectral indices* are defined in literature, to study both molecular graphs and model physico-chemical properties of molecules.

- **cardinality of a set**

The cardinality of a set S is the number of elements in S and is indicated as $|S|$.

- **column sum operator**

This operator, denoted as CS_j , performs the sum of the elements of the j th matrix column:

$$CS_j(\mathbf{M}) \equiv \sum_{i=1}^n m_{ij}$$

The **column sum vector**, denoted by $\mathbf{cs}$, is a p -dimensional vector collecting the results obtained by applying the column sum operator to all the p columns of the matrix.

- **determinant**

The determinant of a $n \times n$ square matrix $\mathbf{M}$, denoted by $\det(\mathbf{M})$, is a scalar quantity and is defined as

$$\det(\mathbf{M}) = \sum_{\pi} s(\pi) \cdot m_{1, i_1} \cdot m_{2, i_2} \cdot \dots \cdot m_{n, i_n}$$

where the summation ranges over all $n!$ permutations π of the symbols 1, 2, ..., n . Each permutation π of degree n is given by

$$\pi = \begin{pmatrix} 1 & 2 & \dots & n \\ i_1 & i_2 & \dots & i_n \end{pmatrix}$$

where $i_1, i_2, \dots, i_n$ are the symbols $1, 2, \dots, n$ in some order. The sign function $s(\pi)$ is defined as

$$s(\pi) = \begin{cases} +1 & \text{if } \pi \text{ is even} \\ -1 & \text{if } \pi \text{ is odd} \end{cases}$$

Related to the definition of determinant are permanent, pfaffian, and hafnian.

The **permanent**, denoted by $\text{per}(\mathbf{M})$, also referred to as the positive determinant, is defined by omitting the sign function $s(\pi)$ [Kasum, Trinajstić *et al.*, 1981; Schultz, Schultz *et al.*, 1992, 1995; Cash, 1995a, 1998; Jiang, Liang *et al.*, 2006] as

$$\text{per}(\mathbf{M}) = \sum_{\pi} m_{1, i_1} \cdot m_{2, i_2} \cdot \dots \cdot m_{n, i_n}$$

where π runs over the $n!$ permutations.

From the permanent, the corresponding permanent polynomial was also defined [Kasum, Trinajstić *et al.*, 1981; Cash, 2000b].

The **immanant**, denoted by $d_{\lambda}(\mathbf{M})$, is defined as

$$d_{\lambda}(\mathbf{M}) = \sum_{\pi} \chi_{\lambda}(\pi) m_{1, i_1} \cdot m_{2, i_2} \cdot \dots \cdot m_{n, i_n}$$

where π runs over the $n!$ permutations. $\chi_{\lambda}(\pi)$ is an irreducible character of the symmetric group indexed by a partition λ of n .

The **pfaffian**, denoted by $\text{pfa}(\mathbf{M})$, and the **hafnian**, denoted by $\text{haf}(\mathbf{M})$, are analogous to the determinant except for the summation that goes over all the permutations $\pi (i_1, i_2, \dots, i_n)$ and must also satisfy the limitations

$$i_1 < i_2, i_3 < i_4, \dots, i_{n-1} < i_n; \quad i_1 < i_3 < i_5 < \dots < i_{n-1}$$

The entries of the main diagonal are excluded from the calculation of the pfaffian and hafnian [Caianiello, 1953, 1956]. Hafnians and pfaffians differ in the sign function $s(\pi)$ that is included in the definition of pfaffian only.

The hafnian calculated considering only the entries above the main diagonal is called **short hafnian**, $\text{shaf}(\mathbf{M})$, whereas the hafnian calculated considering both entries above and below the main diagonal can also be referred to as **long hafnian**, $\text{lhaf}(\mathbf{M})$ [Schultz and Schultz, 1992; Schultz, Schultz *et al.*, 1995].

For example, for a matrix $\mathbf{M}$ of order 4, pfaffian, long hafnian, and short hafnian are the following:

$$\begin{aligned} \text{pfa} &= m_{12} \cdot m_{34} - m_{13} \cdot m_{24} + m_{14} \cdot m_{23} \\ \text{shaf} &= m_{12} \cdot m_{34} + m_{13} \cdot m_{24} + m_{14} \cdot m_{23} \\ \text{lhaf} &= m_{12} \cdot m_{21} \cdot m_{34} \cdot m_{43} + m_{13} \cdot m_{31} \cdot m_{24} \cdot m_{42} + m_{14} \cdot m_{41} \cdot m_{23} \cdot m_{32} \end{aligned}$$

Some molecular descriptors, called $\rightarrow$ *determinant-based descriptors*, are calculated as the determinant of $\rightarrow$ *matrices of molecules*. Moreover, permanents, short and long hafnians, calculated on the topological $\rightarrow$ *distance matrix D*, were used as graph invariants by Schultz and called **per(D) index**, **shaf(D) index**, and **lhaf(D) index** [Schultz and Schultz, 1992; Schultz, Schultz *et al.*, 1992].

☞ [Schultz and Schultz, 1993; Schultz, Schultz *et al.*, 1993, 1994, 1995, 1996; Chan, Lam *et al.*, 1997; Gutman, 1998; Cash, 2000a, 2002a, 2003]

- **diagonal matrix**

A diagonal matrix is a square matrix whose diagonal terms m_{ii} are the only nonzero elements. The **diagonal operator** $\mathcal{D}(\mathbf{M})$ is an operator that transforms a generic square matrix $\mathbf{M}$ into a diagonal matrix:

$$\mathcal{D}(\mathbf{M}) = \begin{vmatrix} m_{11} & \dots & 0 & \dots & 0 \\ 0 & \dots & m_{ii} & \dots & 0 \\ 0 & \dots & 0 & \dots & m_{nn} \end{vmatrix}$$

- **Hadamard matrix product**

The Hadamard product of two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension is denoted as $\otimes$ and defined as

$$[\mathbf{A} \otimes \mathbf{B}]_{ij} = [\mathbf{A}]_{ij} \times [\mathbf{B}]_{ij}$$

that is, the elements of the resulting matrix are obtained by the scalar product of the corresponding elements of $\mathbf{A}$ and $\mathbf{B}$ matrices.

- **identity matrix (I)**

The identity matrix is a square diagonal matrix defined as

$$\mathbf{I} = \begin{vmatrix} 1 & \dots & 0 & \dots & 0 \\ 0 & \dots & 1 & \dots & 0 \\ 0 & \dots & 0 & \dots & 1 \end{vmatrix}$$

- **polynomial**

A polynomial $\mathcal{P}(x)$ of the x variable is a linear combination of its powers, usually written as

$$\mathcal{P}(x) = a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n$$

where n is the order of the polynomial. The values of x , for which $\mathcal{P}(x)$ is zero, are the *roots* of the polynomial.

Several polynomials associated with graphs were defined, such as the $\rightarrow$ *characteristic polynomials*, $\rightarrow$ *counting polynomials*, $\rightarrow$ *matching polynomial*, chromatic polynomial, and Tutte polynomial [Noy, 2003].

- **product of matrices**

Let $\mathbf{A}$ (n, m) and $\mathbf{B}$ (m, p) be two matrices. The product of the two matrices is defined as

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{C}$$

where the resulting product matrix $\mathbf{C}$ has n rows and p columns. Each scalar element c_{ij} of the $\mathbf{C}$ matrix is obtained by the scalar product between the i th row of the $\mathbf{A}$ matrix and the j th column of the $\mathbf{B}$ matrix. The row vector is represented as $\mathbf{a}_i^T$ and the column vector as $\mathbf{b}_j$; the resulting element c_{ij} is then calculated as

$$\mathbf{a}_i^T \cdot \mathbf{b}_j \equiv c_{ij} = \sum_{k=1}^m a_{ik} \cdot b_{kj}$$

A basic condition for the product of two matrices is that the number of columns of the left matrix and the number of rows of the right matrix are equal (m).

The k th **power matrix** $\mathbf{A}^k$ is a special case of the matrix product:

$$\mathbf{A}^k = \mathbf{A} \cdot \mathbf{A}^{k-1}$$

The main properties of the product of two matrices are

$$(a) \mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}, \quad (b) (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}), \quad (c) (\mathbf{A} \cdot \mathbf{B})^T = \mathbf{B}^T \cdot \mathbf{A}^T$$

- **row sum operator** ($\equiv$ vertex sum operator)

This operator, denoted as VS_i , performs the sum of the elements of the i th matrix row:

$$VS_i(\mathbf{M}) = \sum_{j=1}^p m_{ij}$$

where p is the number of columns of the $\mathbf{M}$ matrix.

The **row sum vector**, $\mathbf{rs}$, is an n -dimensional vector collecting the results obtained by applying the row sum operator to all the n rows of the matrix.

This operator is used to derive $\rightarrow$ Local Vertex Invariants from $\rightarrow$ graph theoretical matrices. For symmetric matrices, the local vertex invariants obtained by applying this operator on the transposed matrix coincide with those obtained by applying the operator on the original matrix.

- **scalar product of vectors**

Let $\mathbf{a}$ and $\mathbf{b}$ be two column vectors with the same dimension n . The scalar product between the two vectors is defined as the sum of the products of the corresponding elements of the row vector $\mathbf{a}^T$ and the column vector $\mathbf{b}$ or, vice versa, of the row vector $\mathbf{b}^T$ and the column vector $\mathbf{a}$:

$$\mathbf{a}^T \cdot \mathbf{b} = \mathbf{b}^T \cdot \mathbf{a} = \sum_{k=1}^n a_k \cdot b_k$$

- **sparse matrices**

These are matrices with relatively few nonzero elements. A **binary sparse matrix** $\mathbf{B}$ is a sparse matrix comprised of elements equal to zero or 1. The **geodesic matrix** is a binary sparse matrix ${}^m\mathbf{B}$ defined as [Harary, 1969a]

$$[{}^m\mathbf{B}]_{ij} = \begin{cases} 1 & \text{if } d_{ij} = m \\ 0 & \text{otherwise} \end{cases}$$

where m defines the order of the matrix and d_{ij} is the $\rightarrow$ topological distance between vertices v_i and v_j . The geodesic matrix is largely applied to calculate $\rightarrow$ autocorrelation descriptors, $\rightarrow$ Estrada generalized topological indices, $\rightarrow$ higher order Wiener numbers, $\rightarrow$ interaction geodesic matrices.

Let $\mathbf{M}$ be a matrix $A \times A$ representing a $\rightarrow$ molecular graph G , where A is the number of vertices. To obtain a **m th order sparse matrix** ${}^m\mathbf{M}$ from any matrix $\mathbf{M}$, the $\rightarrow$ Hadamard matrix product is performed as the following:

$${}^m\mathbf{M} = \mathbf{M} \otimes {}^m\mathbf{B}$$

where the superscript “ m ” of the sparse matrix ${}^m\mathbf{M}$ means that all of the $\mathbf{M}$ matrix elements are taken as zero but those corresponding to pairs of vertices v_i and v_j at topological distance m and ${}^m\mathbf{B}$ constitute the geodesic matrix defined above.

The $\rightarrow$ adjacency matrix $\mathbf{A}$ of a molecular graph G is an example of binary sparse matrix, only the off-diagonal entries i - j corresponding to pairs of adjacent vertices v_i and v_j , that is, vertices connected by a bond, being equal to one. Using the adjacency matrix as the multiplier in the Hadamard product, it follows

$${}^1\mathbf{M} \equiv \mathbf{M}_e = \mathbf{M} \otimes \mathbf{A}$$

where ${}^1\mathbf{M}$ is a **first-order sparse matrix**, also called edge-matrix, denoted as $\mathbf{M}_e$.

Opposite to sparse matrices are **dense matrices**, that is, matrices with several nonzero entries [Randić and DeAlba, 1997].

- **stochastic matrices** ($\equiv$ probability matrices, transition matrices)

These are square matrices $\mathbf{M}$ for which each row sum, *right stochastic matrices*, or each column sum, *left stochastic matrices*, is equal to 1, that is, the row elements or the column elements consist of nonnegative real numbers that can be interpreted as probabilities:

$$VS_i(\mathbf{M}) = 1 \quad \text{or} \quad CS_j(\mathbf{M}) = 1$$

where VS is the $\rightarrow$ row sum operator and CS the $\rightarrow$ column sum operator.

Stochastic matrices for which both row and column sums are equal to 1 are called *double stochastic matrices*. Stochastic matrices are defined in the framework of the $\rightarrow$ MARCH-INSIDE descriptors, $\rightarrow$ TOMOCOMD descriptors, and $\rightarrow$ walk counts.

- **sum of matrices**

Let $\mathbf{A}$ (n, p) and $\mathbf{B}$ (n, p) be two equal-sized matrices. The sum of the two matrices is defined as

$$\mathbf{A} + \mathbf{B} = \mathbf{C}$$

Each scalar element c_{ij} of the matrix $\mathbf{C}$ is obtained by summing up the corresponding elements of the two matrices, that is,

$$c_{ij} = a_{ij} + b_{ij}$$

A basic condition for the sum of two matrices is that the two matrices have the same dimension.

The main properties of the sum of two matrices are

$$(a) \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \quad (b) (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \quad (c) \alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$$

where α is a scalar value.

- **total sum operator**

This operator S performs the sum of all of the elements of a matrix $\mathbf{M}$ of size $n \times p$:

$$S(\mathbf{M}) \equiv \sum_{i=1}^n \sum_{j=1}^p m_{ij} = \sum_{i=1}^n VS_i(\mathbf{M}) = \sum_{j=1}^p CS_j(\mathbf{M})$$

where VS_i and CS_j are the row sum operator and the column sum operator, respectively.

- **trace**

The trace of a square matrix $\mathbf{M}$ (i.e., $n = p$), denoted by $tr(\mathbf{M})$, is the sum of the diagonal elements:

$$tr(\mathbf{M}) \equiv \sum_{i=1}^n m_{ii}$$

- **transposition of a matrix**

The matrix $\mathbf{M}^T$ is the transposed matrix of $\mathbf{M}$ if its elements are

$$[\mathbf{M}^T]_{ij} = [\mathbf{M}]_{ji}$$

If the dimension of $\mathbf{M}$ is $n \times p$, the transposed matrix $\mathbf{M}^T$ has dimension $p \times n$.

- **unit matrix ($\mathbf{U}$)**

The unit matrix is a square matrix defined as

$$\mathbf{U} = \begin{vmatrix} 1 & \dots & 1 & \dots & 1 \\ 1 & \dots & 1 & \dots & 1 \\ 1 & \dots & 1 & \dots & 1 \end{vmatrix}$$

- **algebraic semisum charge transfer index** → topological charge indices
- **algebraic structure count** → Kekulé number

■ alignment rules

In most → *grid-based QSAR techniques*, which use as the molecular descriptors energy values of → *molecular interaction fields* (steric, hydrophobic, coulombic, etc.), rules for alignment of all the molecules in the data set are required for comparability purposes. In effect, the energy value at each grid point $\mathbf{p}$ depends on the relative orientation of the compound with respect to the grid. As a consequence, the use of the grid points as molecular descriptors requires the mandatory step of aligning the molecules of the considered → *data set* in such a way that each of the thousands of grid points represents, for all the molecules, the same kind of information, and not spurious information due to lack of invariance in the rotation of the molecules in the grid.

Therefore, in applying grid-based QSAR techniques there are, in most cases, two closely related problems: the selection of a suitable molecular conformation for each compound and the relative alignment of the compounds, either among themselves or with respect to any → *receptor*, if its structure is known.

The ideal choice of conformers for QSAR would be the bioactive one. Wherever experimental structural data (e.g., X-ray data) on ligands bound to targets exist, the bioactive ligand conformation is available and should be used to derive an alignment rule.

When no structural data are available for the receptor, methods that explore conformational space may find the best relative match among the different ligands. During this process, low-energy conformations are selected to obtain the best match from all the different conformations. The solution is usually not unique because other conformations may bind to the unknown receptor, and multiple alignment rules, based on different starting hypotheses, should be considered when no structural information and no rigid compounds are available.

The success or failure of the grid-based methods to find acceptable $\rightarrow$ *quantitative structure–activity relationships* strongly depends on how the molecules are aligned in the grid on which the molecular interaction fields are sampled. In effect, problems may be mainly due to (a) an alignment that leads to a resulting common structure, that is, the pharmacophore, not reliable, and (b) the same grid points in different molecules represent chance variation in model geometry.

To avoid the drawbacks of the molecule alignment, several approaches based on different criteria were proposed; two basic alignment techniques are explained below.

- **point-by-point alignment**

For a set of congeneric compounds, the atoms of each compound are superimposed on their common backbone, aligning as much of each structure as possible.

For structurally diverse compounds, hypotheses on the $\rightarrow$ *pharmacophore* can provide an approach to overcome ambiguities in atom superimposition and identify a suitable alignment.

- **field-fitting alignment**

In this approach, the molecules are aligned by maximizing the degree of similarity between their molecular interaction fields. Different types of probes result in different fields as well as different molecular alignments. Therefore, the selection of suitable fields (and how to weight them) depends on external considerations.

Moreover, a difficulty in field-based alignment is that molecular regions not relevant, that is, parts of the molecule not involved in ligand–receptor interactions, may distort the alignment.

📖 [Kato, Itai *et al.*, 1987; Mayer, Naylor *et al.*, 1987; Kearsley and Smith, 1990; Manaut, Sanz *et al.*, 1991; Cramer III, DePriest *et al.*, 1993; Dean, 1993; Klebe, 1993, 1998; Waller and Marshall, 1993; Waller, Oprea *et al.*, 1993; Klebe, Mietzner *et al.*, 1994; Cramer III, Clark *et al.*, 1996; Petitjean, 1996; Greco, Novellino *et al.*, 1997; Handschuh, Wagener *et al.*, 1998; Langer and Hoffmann, 1998b; Norinder, 1998; Bernard, Kireev *et al.*, 1999; Robinson, Lyne *et al.*, 1999; Lemmen and Lengauer, 2000; Vedani, McMasters *et al.*, 2000; Jewell, Turner *et al.*, 2001; Makhija and Kulkarni, 2001b; Nissink, Verdonk *et al.*, 2001; Pitman, Huber *et al.*, 2001; Wildman and Crippen, 2001; Zhu, Hou *et al.*, 2001; Bringmann and Rummey, 2003; Bultinck, Kuppens *et al.*, 2003; Bultinck, Carbó-Dorca *et al.*, 2003; Hasegawa, Arakawa *et al.*, 2003; Marialke, Körner *et al.*, 2007]

- **Alikhanidi vertex degree** $\rightarrow$ vertex degree
- **all-path matrix** $\rightarrow$ path counts
- **all-path Wiener index** $\rightarrow$ path counts
- **all possible models** $\rightarrow$ variable selection
- **ALOGP** $\rightarrow$ lipophilicity descriptors (© Ghose–Crippen hydrophobic atomic constants)

■ ALPHA descriptor

This is a vectorial molecular descriptor derived from the trajectories obtained by molecular dynamic simulation applying a technique of Gaussian smoothing [Tuppurainen, Viisas *et al.*, 2004]. For each trajectory coordinate x , the *ALPHA* descriptor is defined as

$$ALPHA(x) = \sum_{i=1}^N \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-[(x-\alpha_i)^2/2 \cdot \sigma^2]}$$

where α and σ are the mean and standard deviation of the Gaussian function and the summation denotes over N -overlaid Gaussian functions; α values are first transformed to a bounded range (e.g., 0.5–3). Then, a Gaussian kernel of fixed standard deviation σ (a parameter to be optimized) is placed over each α value. Finally, the quantity *ALPHA*(x) is calculated at intervals of L (usually L is set at $\sigma/2$) resulting into a (pseudo) spectrum, which can be used as a molecular descriptor for QSAR modeling.

The dimensionality of the *ALPHA* descriptor is high (depending strongly on the value of σ) and, thus, the PLS method is suggested to compress the data.

- **Altenburg polynomial** → counting polynomials
- **altered Wiener indices** → Wiener index
- **amino acid descriptors** → biodescriptors
- **amino acid sequences** $\equiv$ *peptide sequences*
- **Amoore shape indices** → shape descriptors

■ amphiphilic moments

The amphiphilicity of a compound is defined as the difference between the free energy of transfer of a compound from the aqueous phase to the air–water interface and the free energy of micelle formation and is quantified by means of surface tension measurements.

Amphiphilic moments are defined as vectors pointing from the center of the hydrophobic domain to the center of the hydrophilic domain of a molecule. It is defined as [Fischer, Gottschlich *et al.*, 1998; Fischer, Kansy *et al.*, 2001]:

$$\bar{A} = \sum_{i=1}^A d_i \cdot \alpha_i$$

where d is the distance of an identified charged residue from the farthest hydrophobic/hydrophilic residues. Each atom is weighted by its hydrophobic/hydrophilic property α on the basis of an atom contribution method [Meylan and Howard, 1995].

The vector length is proportional to the strength of the amphiphilic moment and it may determine the ability of a compound to permeate a membrane [Cruciani, Crivori *et al.*, 2000].

- **AMSP** $\equiv$ *Autocorrelation of Molecular Surface Properties* → autocorrelation descriptors
- **Andrews' curves** → molecular descriptors ($\odot$ transformations of molecular descriptors)
- **Andrews descriptors** → count descriptors
- **angular distance** → similarity/diversity
- **angular separation** → similarity/diversity ($\odot$ Table S7)
- **a_N -index** → determinant-based descriptors ($\odot$ general a_N -index)

- **anisometry** → shape descriptors
- **anisotropy of the polarizability** → electric polarization descriptors
- **Ant Colony fitness function** → regression parameters
- **antibonding orbital information index** → information theoretic topological index
- **anticonnectivity indices** → variable descriptors

■ applicability domain

The concept of the applicability domain concerns the predictive use of QSAR/QSPR models and, then, is closely related to the concept of model validation (→ *validation techniques*). In other words, the applicability domain is a concept related to the quality of the QSAR/QSPR model predictions and prevention of the potential misuse of model's results. A key component of the prediction quality is indeed to define when a QSAR/QSPR model is suitable to predict a property/activity of a new compound [Tropsha, Gramatica *et al.*, 2003; Jaworska, Nikolova-Jeliazkova *et al.*, 2004; Dimitrov, Dimitrova *et al.*, 2005; Jaworska, Nikolova-Jeliazkova *et al.*, 2005; Netzeva, Worth *et al.*, 2005; Nikolova-Jeliazkova and Jaworska, 2005].

A model will yield reliable predictions when model assumptions are fulfilled and unreliable predictions when they are violated. In particular, for QSAR/QSPR models, based on statistical mining techniques, the → *training set* and the model prediction space are the basis for the estimation of space where predictions are reliable.

Two basic approaches were proposed for evaluating the applicability domain.

The first approach to applicability domain evaluation is the statistical analysis of the training set, trying to define the best conditions for interpolated prediction that is usually more reliable than extrapolation. Extrapolation is not a problem in principle, because extrapolated results from theoretically well-founded models can often be reliable. However, QSAR/QSPR models are usually based on empirical, and limited experimental evidence and/or are only locally valid; therefore, extrapolation usually results in high uncertainty and not reliable predictions.

Different approaches to estimate interpolation regions in a multivariate space were evaluated by Jaworska [Jaworska, Nikolova-Jeliazkova *et al.*, 2005], based on (a) ranges of the descriptor space; (b) distance-based methods, using Euclidean, Manhattan, and Mahalanobis distances, Hotelling T^2 method and leverage values; and (c) probability density distribution methods based on parametric and nonparametric approaches. Both ranges and distance-based methods were also evaluated in the principal component space by → *Principal Component Analysis*.

Another approach to applicability domain evaluation is based on the → *similarity/diversity* of the compound considered with respect to those belonging to the training set; a QSAR/QSPR prediction should be reliable if the compound is, in some way, similar to one or more compounds present in the training set [Nikolova and Jaworska, 2003]. High similarity is simply another way to use the interpolation ability of the model in place of the extrapolation.

A stepwise procedure was also proposed [Dimitrov, Dimitrova *et al.*, 2005] based on a four-stage procedure: (1) a study of the variations of molecular parameters that may affect the quality of the measured endpoint significantly (e.g., molecular weight, absorption, water solubility volatility, etc.); (2) an analysis of the structural domain based on a set of → *atom-centered fragment descriptors* that could be used to characterize the structural domain of the atoms present in the training set; (3) an analysis of the mechanistic domain focused on functional groups whose reactivity modulates the endpoint studied or structural fragments used in group contribution models; and (4) an analysis of the metabolic domain by simulators, although the metabolic aspects are usually not included in QSAR models.

☞ [Martin, Kofron *et al.*, 2002; Eriksson, Jaworska *et al.*, 2003; Papa, Villa *et al.*, 2005; Tetko, Bruneau *et al.*, 2006; Zhang, Golbraikh *et al.*, 2006; Stanforth, Kolossov *et al.*, 2007]

- **arcs** → graph
- **arithmetic mean** → statistical indices (⊙ indices of central tendency)
- **arithmetic topological index** → vertex degree
- **aromatic bond count** → multiple bond descriptors
- **aromaticity** → delocalization degree indices
- **aromaticity indices** → delocalization degree indices
- **artificial neural networks** → chemometrics
- **aryl electronic constants** → electronic substituent constants
- **ASIIg index** → charge descriptors (⊙ charge-related indices)
- **asphericity** → shape descriptors
- **association coefficients** → similarity/diversity
- **asymptotic Q^2 rule** → regression parameters
- **ATAC** $\equiv$ *Atom-Type AutoCorrelation* → autocorrelation descriptors
- **atom–atom polarizability** → electric polarization descriptors
- **atom-centered fragment descriptors** $\equiv$ *Augmented Atoms* → substructure descriptors
- **atom connectivity matrices** → weighted matrices (⊙ weighted adjacency matrices)
- **atom count** $\equiv$ *atom number*
- **atom detour eccentricity** → detour matrix
- **atom eccentricity** → distance matrix
- **atom electronegativity** → atomic properties
- **Atom Environment descriptors** → substructure descriptors (⊙ fingerprints)
- **atomic charges** → quantum-chemical descriptors
- **atomic charge-weighted negative surface area** → charged partial surface area descriptors
- **atomic charge-weighted positive surface area** → charged partial surface area descriptors

■ **atomic composition indices** ($\equiv$ *composition indices*)

Molecular → *0D descriptors* with high degeneracy, derived from the chemical formula of compounds and defined as → *information indices* of the elemental composition of the molecule. They can be considered → *molecular complexity indices* that take into account the molecular diversity in terms of different atom types.

→ *Average molecular weight* and → *relative atom-type count* are simple molecular descriptors that encode information on atomic composition. Other important descriptors of the atomic composition are based on the → *total information content* and the → *mean information content*, defined as

• **total information index on atomic composition** (I_{AC})

The total information content on atomic composition of the molecule is calculated from the complete molecular formula, hydrogen included, as

$$I_{AC} = A^h \cdot \log_2 A^h - \sum_g A_g \cdot \log_2 A_g$$

where A^h is the total number of atoms (hydrogen included) and A_g is the number of atoms of chemical element of type g [Dancoff and Quastler, 1953].

For example, benzene has 6 carbon and 6 hydrogen atoms; then, as $A^h = 12$ and $A_g = 6$ for both equivalence classes, $I_{AC} = 12$.

• **mean information index on atomic composition ($\bar{I}_{AC}$)**

The mean information content on atomic composition is the mean value of the total information content and is calculated as

$$\bar{I}_{AC} = -\sum_g \frac{A_g}{A^h} \cdot \log_2 \frac{A_g}{A^h} = -\sum_g p_g \cdot \log_2 p_g$$

where A^h is the total number of atoms (hydrogen included), A_g is the number of atoms of type g and p_g is the probability to randomly select a g th atom type [Dancoff and Quastler, 1953]. For example, for benzene $\bar{I}_{AC} = 1$.

- **atomic connectivity indices** $\equiv$ *local connectivity indices* $\rightarrow$ connectivity indices
- **atomic dispersion coefficient** $\rightarrow$ hydration free energy density
- **Atomic Environment Autocorrelations** $\rightarrow$ autocorrelation descriptors
- **atomic ID numbers** $\rightarrow$ ID numbers
- **atomic information content** $\rightarrow$ atomic information indices

■ **atomic information indices**

Atomic descriptors related to the internal composition of atoms [Bonchev, 1983].

The **atomic information content** I_{at} is the $\rightarrow$ *total information content* of an atom viewed as a system whose structural elements, that is, protons p , neutrons n , and electrons el , are partitioned into nucleons, $p + n$, and electrons el :

$$I_{at} = (N_n + N_p + N_{el}) \cdot \log_2 (N_n + N_p + N_{el}) - N_{el} \cdot \log_2 N_{el} - (N_n + N_p) \cdot \log_2 (N_n + N_p)$$

where N_n , N_p , and N_{el} are the numbers of neutrons, protons, and electrons, respectively [Bonchev and Peev, 1973].

To account for the different isotopes of a given chemical element, the **information index on isotopic composition** was defined as

$$I_{IC} = \sum_k (I_{at})_k \cdot f_k$$

where the sum runs over all isotopes of the considered chemical element, $(I_{at})_k$ is the atomic information content of the k th isotope and f_k is its relative amount [Bonchev and Peev, 1973].

The **information index on proton–neutron composition** is an atomic descriptor defined as total information content of the atomic nucleus:

$$I^{n,p} = (N_n + N_p) \cdot \log_2 (N_n + N_p) - N_n \cdot \log_2 N_n - N_p \cdot \log_2 N_p$$

where N_n and N_p are the numbers of neutrons and protons, respectively [Bonchev, Peev *et al.*, 1976].

The **nuclear information content** I_{NUCL} is a molecular descriptor calculated as the sum of information indices on the proton–neutron composition of all the nuclei of a molecule:

$$I_{NUCL} = \sum_{i=1}^A I_i^{n,p}$$

where A is the number of atoms and $I_i^{n,p}$ is the information index on proton–neutron composition of the nucleus of the i th atom. This index also accounts for molecular size by means of the number of atomic nuclei.

- **atomic molecular connectivity index** → connectivity indices
- **atomic moments of energy** → self-returning walk counts
- **atomic multigraph factor** → bond order indices (⊙ conventional bond order)
- **atomic path count** → path counts
- **atomic path count sum** → path counts
- **atomic path number** $\equiv$ *atomic path count* → path counts
- **atomic path/walk indices** → shape descriptors (⊙ path/walk shape indices)
- **atomic polarization** → electric polarization descriptors

■ atomic properties

“Most atomic properties are a consequence of atomic structure, which in turn must be related to the inherent nature of the component electrons and nuclei. Therefore it is almost inevitable that such properties be related to one another, if only because of their common origin. It should not be surprising that a particular property, here the electronegativity, can be derived from or correlated with a wide variety of other properties, with reasonable agreement among the several results” [Sanderson, 1988].

Atomic properties P are physics and chemical observables characterizing each chemical element. They play a fundamental role in the definition of most of the molecular descriptors, being physico-chemical properties, as well as biological, toxicological, and environmental properties, deeply determined by the chemical elements constituting the molecule itself.

In Table A2, some important atomic properties are listed for the most common chemical elements.

Table A2 Atomic properties for some chemical elements.

Atom	Z	L	Z'	R^{vdw}	R^{cov}	m	V^{vdw}	χ^{SA}	α	IP	EA
H	1	1	1	1.17	0.37	1.01	6.71	2.59	0.67	13.598	0.754
Li	3	2	1	1.82	1.34	6.94	25.25	0.89	24.3	5.392	0.618
Be	4	2	2	–	0.90	9.01	–	1.81	5.60	9.323	–
B	5	2	3	1.62	0.82	10.81	17.88	2.28	3.03	8.298	0.277
C	6	2	4	1.75	0.77	12.01	22.45	2.75	1.76	11.260	1.263
N	7	2	5	1.55	0.75	14.01	15.60	3.19	1.10	14.534	–
O	8	2	6	1.40	0.73	16.00	11.49	3.65	0.80	13.618	1.461
F	9	2	7	1.30	0.71	19.00	9.20	4.00	0.56	17.423	3.401
Na	11	3	1	2.27	1.54	22.99	49.00	0.56	23.6	5.139	0.548
Mg	12	3	2	1.73	1.30	24.31	21.69	1.32	10.6	7.646	–
Al	13	3	3	2.06	1.18	26.98	36.51	1.71	6.80	5.986	0.441
Si	14	3	4	1.97	1.11	28.09	31.98	2.14	5.38	8.152	1.385
P	15	3	5	1.85	1.06	30.97	26.52	2.52	3.63	10.487	0.747
S	16	3	6	1.80	1.02	32.07	24.43	2.96	2.90	10.360	2.077
Cl	17	3	7	1.75	0.99	35.45	22.45	3.48	2.18	12.968	3.613

(Continued)

Table A2 (Continued)

Atom	Z	L	Z ^v	R ^{vdw}	R ^{cov}	m	V ^{vdw}	χ ^{SA}	α	IP	EA
K	19	4	1	2.75	1.96	39.10	87.11	0.45	43.4	4.341	0.501
Ca	20	4	2	—	1.74	40.08	—	0.95	22.8	6.113	0.018
Cr	24	4	6	2.20	1.27	52.00	44.60	1.66	11.60	6.767	0.666
Mn	25	4	7	2.18	1.39	54.94	43.40	2.20	9.40	7.434	—
Fe	26	4	8	2.14	1.25	55.85	41.05	2.20	8.40	7.902	0.151
Co	27	4	9	2.03	1.26	58.93	35.04	2.56	7.50	7.881	0.662
Ni	28	4	10	1.60	1.21	58.69	17.16	1.94	6.80	7.640	1.156
Cu	29	4	11	1.40	1.38	63.55	11.49	1.98	6.10	7.723	1.235
Zn	30	4	12	1.39	1.31	65.39	11.25	2.23	7.10	9.394	—
Ga	31	4	3	1.87	1.26	69.72	27.39	2.42	8.12	5.999	0.300
Ge	32	4	4	1.90	1.22	72.61	28.73	2.62	6.07	7.900	1.233
As	33	4	5	1.85	1.19	74.92	26.52	2.82	4.31	9.815	0.810
Se	34	4	6	1.90	1.16	78.96	28.73	3.01	3.77	9.752	2.021
Br	35	4	7	1.95	1.14	79.90	31.06	3.22	3.05	11.814	3.364
Rb	37	5	1	—	2.11	85.47	—	0.31	47.3	4.177	0.486
Sr	38	5	2	—	1.92	87.62	—	0.72	27.6	5.695	0.110
Mo	42	5	6	2.00	1.45	95.94	33.51	1.15	12.80	7.092	0.746
Ag	47	5	11	1.72	1.53	107.87	21.31	1.83	7.20	7.576	1.302
Cd	48	5	12	1.58	1.48	112.41	16.52	1.98	7.20	8.994	—
In	49	5	3	1.93	1.44	114.82	30.11	2.14	10.20	5.786	0.300
Sn	50	5	4	2.22	1.41	118.71	45.83	2.30	7.70	7.344	1.112
Sb	51	5	5	2.10	1.38	121.76	38.79	2.46	6.60	8.640	1.070
Te	52	5	6	2.06	1.35	127.60	36.62	2.62	5.50	9.010	1.971
I	53	5	7	2.10	1.33	126.90	38.79	2.78	5.35	10.451	3.059
Gd	64	6	10	2.59	1.79	157.25	72.78	2.00	23.50	6.150	0.500
Pt	78	6	10	1.75	1.28	195.08	22.45	2.28	6.50	9.000	2.128
Au	79	6	11	1.66	1.44	196.97	19.16	2.54	5.80	9.226	2.309
Hg	80	6	12	1.55	1.49	200.59	15.60	2.20	5.70	10.438	—
Tl	81	6	3	1.96	1.48	204.38	31.54	2.25	7.60	6.108	0.200
Pb	82	6	4	2.02	1.47	207.20	34.53	2.29	6.80	7.417	0.364
Bi	83	6	5	2.10	1.46	208.98	38.79	2.34	7.40	7.289	0.946

Z, atomic number; L, principal quantum number; Z^v, number of valence electrons; R^{vdw}, van der Waals atomic radius; R^{cov}, covalent radius; m, atomic mass; V^{vdw}, van der Waals volume; χ^{SA}, Sanderson electronegativity; α, atomic polarizability (10^{−24} cm³); IP, ionization potential (eV); EA, electron affinity (eV).

For atomic mass, van der Waals volume, Sanderson electronegativity, and atom polarizability, the scaled values with respect to the carbon atom are listed in Table A3.

Table A3 Atomic mass (m), van der Waals volume (V^{vdw}), Sanderson electronegativity (χ^{SA}), and polarizability (α): original values and scaled values with respect to the carbon atom value.

ID	Atomic mass		Volume		Electronegativity		Polarizability	
	m	m/m _C	V ^{vdw}	V ^{vdw} /V _C ^{vdw}	χ ^{SA}	χ ^{SA} /χ _C ^{SA}	α	α/α _C
H	1.01	0.084	6.709	0.299	2.592	0.944	0.667	0.379
B	10.81	0.900	17.875	0.796	2.275	0.828	3.030	1.722

(Continued)

Table A3 (Continued)

ID	Atomic mass		Volume		Electronegativity		Polarizability	
	m	m/m _C	V ^{vdw}	V ^{vdw} /V _C ^{vdw}	χ ^{SA}	χ ^{SA} /χ _C ^{SA}	α	α/α _C
C	12.01	1.000	22.449	1.000	2.746	1.000	1.760	1.000
N	14.01	1.166	15.599	0.695	3.194	1.163	1.100	0.625
O	16.00	1.332	11.494	0.512	3.654	1.331	0.802	0.456
F	19.00	1.582	9.203	0.410	4.000	1.457	0.557	0.316
Al	26.98	2.246	36.511	1.626	1.714	0.624	6.800	3.864
Si	28.09	2.339	31.976	1.424	2.138	0.779	5.380	3.057
P	30.97	2.579	26.522	1.181	2.515	0.916	3.630	2.063
S	32.07	2.670	24.429	1.088	2.957	1.077	2.900	1.648
Cl	35.45	2.952	23.228	1.035	3.475	1.265	2.180	1.239
Fe	55.85	4.650	41.052	1.829	2.000	0.728	8.400	4.773
Co	58.93	4.907	35.041	1.561	2.000	0.728	7.500	4.261
Ni	58.69	4.887	17.157	0.764	2.000	0.728	6.800	3.864
Cu	63.55	5.291	11.494	0.512	2.033	0.740	6.100	3.466
Zn	65.39	5.445	38.351	1.708	2.223	0.810	7.100	4.034
Br	79.90	6.653	31.059	1.384	3.219	1.172	3.050	1.733
Sn	118.71	9.884	45.830	2.042	2.298	0.837	7.700	4.375
I	126.90	10.566	38.792	1.728	2.778	1.012	5.350	3.040

Other atomic electronegativity scales are reported elsewhere (→ *electronegativity*, Table E7).

- **atomic refractivity** → physico-chemical properties (⊙ molar refractivity)
- **atomic self-returning walk count** → self-returning walk counts
- **atomic sequence count** → sequence matrices

■ **atomic solvation parameter ($\Delta\sigma$)**

An empirical atomic descriptor $\Delta\sigma$ proposed to calculate solvation free energy of a group X in terms of atomic contributions by the following equation:

$$\Delta G_X = \sum_i \Delta\sigma_i \cdot SA_i$$

where the sum runs over all the nonhydrogen atoms of the X group, SA is the → *solvent accessible surface area* of the i th atom, and $\Delta\sigma$ denotes the corresponding atomic contribution to solvation energy [Eisenberg and McLachlan, 1986]. The atomic contributions $\Delta\sigma_i \cdot SA_i$ are the free energy of transfer of each atom to the solution; note that the areas SA depend on molecule conformation. Proposed to study protein folding and binding, the estimated values for the atomic solvation parameters $\Delta\sigma$ (in cal Å⁻² mol⁻¹) are $\Delta\sigma_C = 16$, $\Delta\sigma_N = -6$, $\Delta\sigma_O = -6$, $\Delta\sigma_{N^+} = -50$, $\Delta\sigma_{O^-} = 24$, and $\Delta\sigma_S = 21$ for carbon, nitrogen, oxygen, nitrogen cation, oxygen anion, and sulfur, respectively.

- **atomic valency index** → quantum-chemical descriptors
- **atomic walk count** → walk counts
- **atomic walk count sum** → walk counts

- **atomic weight-weighted adjacency matrix** → weighted matrices ($\odot$ weighted adjacency matrices)
- **atomic weight-weighted distance matrix** → weighted matrices ($\odot$ weighted distance matrices)
- **Atom-in-Molecules theory** $\equiv$ AIM theory
- **atom-in-structure invariant index** → charge descriptors ($\odot$ charge-related indices)
- **atomistic topological indices** → count descriptors
- **atom-level composite ETA index** → ETA indices
- **atom leverage-based center** → center of a molecule

■ atom number (A) ($\equiv$ atom count)

This is the simplest measure with regard to molecular size, defined as the total number of atoms in a molecule. It is a global, zero dimensional, descriptor with a high degeneracy. In several applications for the calculation of → *molecular descriptors*, the atom number A refers only to nonhydrogen atoms.

The **information index on size** is the → *total information content* on the atom number, defined as

$$I_{\text{SIZE}} = A^h \log_2 A^h$$

where the atom number A^h also takes hydrogen atoms into account [Bertz, 1981]. This index can also be calculated without considering hydrogen atoms.

Other related molecular descriptors are → *atomic composition indices*, several → *information indices* and → *graph invariants*.

- **atom-pair matching function** → molecular shape analysis
- **atom pairs** → substructure descriptors
- **atom polarizability** → electric polarization descriptors
- **Atom-Type AutoCorrelation** → autocorrelation descriptors
- **atom-type autocorrelation matrix** → weighted matrices ($\odot$ weighted distance matrices)

■ atom type-based topological indices

Atom-type topological indices are used to describe a molecule by information related to different atom types in the molecule. An atom-type index is usually derived from some properties of all the atoms of the same type and their structural environment. → *Atom-type E-state indices* of Kier and Hall, → *perturbation connectivity indices*, → *atom-type path counts*, and → *atom-type autocorrelation descriptors* are examples of these molecular descriptors.

AI indices are atom-type topological indices derived from the → *Xu index*, whose formula is applied to single atom types [Ren, 2002a, 2002b, 2002c, 2003a, 2003c, 2003d]. For any i th atom in the molecular graph, first a local vertex invariant, denoted as AI_i , is calculated as

$$AI_i = 1 + \phi_i = 1 + \frac{\delta_i^m \cdot \sigma_i^2}{\sum_{i=1}^A \delta_i^m \cdot \sigma_i}$$

where ϕ is a perturbation term reflecting the effects of the structural environment of the i th atom on the topological index and σ the → *vertex distance sum*. The expression defined above is based

on the $\rightarrow$ Ren vertex degree δ_i^m derived from the Kier–Hall valence vertex degree and defined as

$$\delta_i^m = \delta_i + \left[\left(\frac{2}{L_i} \right)^2 \cdot \delta_i^v + 1 \right]^{-1} = \delta_i + (I_i \cdot \delta_i)^{-1}$$

where δ_i is the $\rightarrow$ vertex degree of the i th atom, L_i its principal quantum number, δ_i^v its $\rightarrow$ valence vertex degree, and I_i denotes the $\rightarrow$ intrinsic state. This formula is applied only to heteroatoms or carbon atoms with multiple bonds and/or bonded to heteroatoms; otherwise, the Ren vertex degree coincides with the simple vertex degree δ_i .

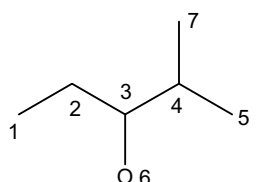
According to this definition of local vertex invariants, the AI index for the k th atom type is derived by adding the values of the local AI indices for all the atoms of type k .

$$AI(k) = \sum_{i=1}^{n_k} AI_i = n_k + \sum_{i=1}^{n_k} \phi_i = n_k + \frac{\sum_{i=1}^{n_k} \delta_i^m \cdot \sigma_i^2}{\sum_{i=1}^A \delta_i^m \cdot \sigma_i}$$

where n_k is the number of atoms of type k .

Example A5

Calculation of AI indices for 4-methyl-3-pentanol. σ is the vertex distance sum and δ^m the Ren vertex degree; **D** is the distance matrix.

	Atom	1	2	3	4	5	6	7	σ_i	δ_i^m
	1	0	1	2	3	4	3	4	17	1.000
	2	1	0	1	2	3	2	3	12	2.000
	3	2	1	0	1	2	1	2	9	3.250
	4	3	2	1	0	1	2	1	10	3.000
	5	4	3	2	1	0	3	2	15	1.000
	6	3	2	1	2	3	0	3	14	1.167
	7	4	3	2	1	2	3	0	15	1.000

$$AI(-CH_3) = AI_1 + AI_5 + AI_7 = \left(1 + \frac{1 \times 17^2}{146.588} \right) + 2 \left(1 + \frac{1 \times 15^2}{146.588} \right) = 8.041$$

$$AI(-CH_2-) = AI_2 = 1 + \frac{2 \times 12^2}{146.588} = 2.965$$

$$AI(-CH<) = AI_3 + AI_4 = \left(1 + \frac{3 \times 9^2}{146.588} \right) + \left(1 + \frac{3 \times 10^2}{146.588} \right) = 5.704$$

$$AI(-OH) = AI_6 = 1 + \frac{1.167 \times 14^2}{146.588} = 2.560$$

Based on the same approach as the AI indices but derived from the formula of the $\rightarrow$ Lu index, the **DAI indices** are atom-type topological indices that exploit bond length-weighted

interatomic distances calculated by adding the relative bond lengths of the edges along the shortest path [Lu, Guo *et al.*, 2006b, 2006c, 2006d; Lu, Wang *et al.*, 2006]. For any i th atom in the molecular graph, the local vertex invariant DAI_i is calculated as

$$DAI_i = 1 + \phi_i = 1 + A \frac{\sum_{j=1}^A [\mathbf{D}(r^*)]_{ij}}{\sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}(r^*)]_{ij}}$$

where ϕ is the perturbation term relative to the atom environment, A is the number of atoms, and $[\mathbf{D}(r^*)]_{ij}$ are the elements of the $\rightarrow$ bond length-weighted distance matrix.

The DAI index for the k th atom type is calculated by adding contributions of all atoms of the considered type:

$$DAI(k) = \sum_{i=1}^{n_k} DAI_i = n_k + \sum_{i=1}^{n_k} \phi_i$$

where n_k is the number of atoms of type k .

Example A6

Calculation of DAI indices for 4-methyl-3-pentanol. VS_i indicates the matrix row sums; $\mathbf{D}(r^*)$ is the bond length-weighted distance matrix.

Atom	1	2	3	4	5	6	7	VS_i
1	0	1	2	3	4	2.928	4	16.928
2	1	0	1	2	3	1.928	3	11.928
3	2	1	0	1	2	0.928	2	8.928
4	3	2	1	0	1	1.928	1	9.928
5	4	3	3	1	0	2.928	2	14.928
6	2.928	1.928	0.928	1.928	2.928	0	2.928	13.571
7	4	3	2	1	2	2.928	0	14.928

$$DAI(-CH_3) = DAI_1 + DAI_5 + DAI_7 = \left(1 + 7 \times \frac{16.928}{91.143}\right) + 2 \left(1 + 7 \times \frac{14.928}{91.143}\right) = 6.593$$

$$DAI(-CH_2-) = DAI_2 = 1 + 7 \times \frac{11.928}{91.143} = 1.916$$

$$DAI(-CH<) = DAI_3 + DAI_4 = \left(1 + 7 \times \frac{8.928}{91.143}\right) + \left(1 + 7 \times \frac{9.928}{91.143}\right) = 3.448$$

$$DAI(-OH) = DAI_6 = 1 + 7 \times \frac{13.571}{91.143} = 2.042$$

- atom-type count $\rightarrow$ count descriptors
- atom-type *E*-state counts $\rightarrow$ electrotopological state indices
- atom-type *E*-state indices $\rightarrow$ electrotopological state indices
- atom-type *HE*-state indices $\rightarrow$ electrotopological state indices

- **atom-type interaction matrices** → weighted matrices (⊙ weighted distance matrices)
- **ATS descriptor** → autocorrelation descriptors (⊙ Moreau–Broto autocorrelation)
- **attractive steric effects** → minimal topological difference
- **augmented adjacency matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **Augmented Atom keys** → substructure descriptors
- **Augmented Atoms** → substructure descriptors
- **augmented connectivity** → eccentricity-based Madan indices
- **augmented distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **augmented eccentric connectivity index** → eccentricity-based Madan indices (⊙ Table E1)
- **augmented edge adjacency matrix** → edge adjacency matrix
- **augmented matrices** → matrices of molecules
- **augmented pair descriptors** → substructure descriptors
- **augmented valence** → vertex degree
- **augmented vertex degree** → weighted matrices (⊙ weighted adjacency matrices)
- **augmented vertex degree matrix** → weighted matrices (⊙ weighted distance matrices)
- **Austel branching index** → steric descriptors

■ autocorrelation descriptors

→ *Molecular descriptors* based on the autocorrelation function AC_k , defined as

$$AC_k = \int_a^b f(x) \cdot f(x+k) \cdot dx$$

where $f(x)$ is any function of the variable x and k is the lag representing an interval of x , and a and b define the total studied interval of the function. The function $f(x)$ is usually a time-dependent function such as a time-dependent electrical signal or a spatial-dependent function such as the population density in space. Then, autocorrelation measures the strength of a relationship between observations as a function of the time or space separation between them [Moreau and Turpin, 1996].

The autocorrelation function AC_k is the integration of the products of the function values calculated at x and $x+k$. This function expresses how numerical values of the function at intervals equal to the lag are correlated.

Autocorrelation functions AC_k can also be calculated for any ordered discrete sequence of n values $f(x_i)$ by summing the products of the i th value and the $(i+k)$ th value as

$$AC_k = \frac{1}{(n-k) \cdot \sigma^2} \cdot \sum_{i=1}^{n-k} [(f(x_i) - \mu) \cdot (f(x_{i+k}) - \mu)]$$

where k is the lag, σ^2 is the variance of the function values, and μ is their mean. The lag assumes values between 1 and K , where the maximum value K can be $n-1$; however, in several applications, K is chosen equal to a small number ($K < 8$). A lag value of zero corresponds to the sum of the square-centered values of the function.

Note that it is common practice in many disciplines to use the term *autocorrelation* even if the standardization by σ^2 is not applied; in this case, the correct term should be *autocovariance*.

Autocorrelation descriptors of chemical compounds are calculated by using various molecular properties that can be represented at the atomic level or molecular surface level or else.

A property of the autocorrelation function is that it does not change when the origin of the x variable is shifted. In effect, autocorrelation descriptors are considered $\rightarrow$ *TRI descriptors*, meaning that they have translational and rotational invariance.

Based on the same principles as the autocorrelation descriptors, but calculated contemporarily on two different properties $f(x)$ and $g(x)$, **cross-correlation descriptors** are calculated to measure the strength of relationships between the two considered properties. For any two ordered sequences comprised of a number of discrete values, the cross-correlation is calculated by summing the products of the i th value of the first sequence and the $(i + k)$ th value of the second sequence as

$$CC_k = \frac{1}{(n-k) \cdot \sigma_{f(x)} \cdot \sigma_{g(x)}} \cdot \sum_{i=1}^{n-k} \left[(f(x_i) - \mu_{f(x)}) \cdot (g(x_{i+k}) - \mu_{g(x)}) \right]$$

where n is the lowest cardinality of the two sets. For the autocorrelation, cross-correlation is usually calculated without the standardization by the two standard deviations $\sigma_{f(x)}$ and $\sigma_{g(x)}$; in this case, the correct term should be *cross-covariance*.

The most common spatial autocorrelation molecular descriptors are obtained by taking the molecule atoms as the set of discrete points in space and an atomic property as the function evaluated at those points.

Common weighting schemes w used to describe atoms in the molecule are $\rightarrow$ *physico-chemical properties* such as atomic masses, $\rightarrow$ *van der Waals volumes*, $\rightarrow$ *atomic electronegativities*, $\rightarrow$ *atomic polarizabilities*, covalent radius, and so on. Alternatively, the weighting scheme for atoms can be based on $\rightarrow$ *local vertex invariants* such as the topological $\rightarrow$ *vertex degrees*, Kier–Hall $\rightarrow$ *intrinsic states* or $\rightarrow$ *E-state indices*, $\rightarrow$ *normalized distance complexity index*, and related indices. Most of these $\rightarrow$ *weighting schemes* are implemented in DRAGON software [DRAGON – Talete s. r.l., 2007; Mauri, Consonni *et al.*, 2006] allowing calculation of different types of autocorrelation descriptors. A comparison of QSARs based on autocorrelation descriptors derived from different weighting schemes is reported in [Kabankin and Gabrielyan, 2005].

For spatial autocorrelation molecular descriptors calculated on a molecular graph, the lag k coincides with the $\rightarrow$ *topological distance* between any pair of vertices.

Autocorrelation descriptors can also be calculated from 3D spatial molecular geometry. In this case, the distribution of a molecular property can be evaluated by a mathematical function $f(x, y, z)$, x , y , and z being the spatial coordinates, defined either for each point of molecular space or molecular surface (i.e., a continuous property such as electronic density or molecular interaction energy) or only for points occupied by atoms (i.e., atomic properties) [Wagener, Sadowski *et al.*, 1995].

The plot of an ordered sequence of autocorrelation descriptors from lag 0 to lag K is called **autocorrelogram** and is usually used to describe a chemical compound in $\rightarrow$ *similarity/diversity* analysis.

Maximum Auto-Cross-Correlation descriptors (or **MACC descriptors**) are autocorrelation and cross-correlation descriptors calculated by taking into account only the maximum product of molecular properties for each lag k :

$$MACC_k = \max_i (f(x_i) \cdot g(x_{i+k}))$$

This function was applied to derive $\rightarrow$ *maximal R indices* and, with the name of MACC-2 transform, to calculate the $\rightarrow$ *GRIND descriptors*.

Moreover, a special case of autocorrelation descriptors is the **Atom-Type AutoCorrelation** (ATAC), which is calculated by summing property values only of atoms of given types. The simplest atom-type autocorrelation is given by

$$\text{ATAC}_k(u, v) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \delta_{ij}(u, v) \cdot \delta(d_{ij}; k)$$

where u and v denote two atom types. $\delta_{ij}(u, v)$ is a Kronecker delta function assuming a value equal to 1 if the atoms i and j form a pair of types u and v or, equivalently, of types v and u ; $\delta(d_{ij}; k)$ is a Kronecker delta function equal to 1 if the interatomic distance d_{ij} is equal to the lag k , and zero otherwise.

This descriptor is defined for each pair of atom types and simply encodes the occurrence numbers of the given atom type pair at different distance values. It can be normalized by using two different procedures: the first one consists in dividing each ATAC_k value by the total number of atom pairs at distance k independently of their types; the second one consists in dividing each ATAC_k value by a constant, which can be equal to the total number of atoms in the molecule or, alternatively, to the total number of (u, v) atom type pairs in the molecule.

Atom types can be defined in different ways; they can be defined in terms of the simple chemical elements or may account also for atom connectivity, hybridization states, and pharmacophoric features. Atom-type autocorrelations can be viewed as a special case of the $\rightarrow$ *atom-type interaction matrices*, from which other kinds of descriptors can also be derived.

Atom-type autocorrelations have been used to derive some $\rightarrow$ *substructure descriptors* such as $\rightarrow$ *atom pairs*, $\rightarrow$ *CATS descriptors*, and related descriptors.

Examples of autocorrelation descriptors, which are derived from the molecular graph but exploit 3D spatial information, are $\rightarrow$ *GETAWAY descriptors*, $\rightarrow$ *PEST Autocorrelation Descriptors* and $\rightarrow$ *SWM signals*. Other autocorrelation descriptors were derived from $\rightarrow$ *molecular shape field*.

A collection of other auto- and cross-correlation descriptors is discussed in the following sections.

• **Moreau–Broto autocorrelation** ($\equiv$ *Autocorrelation of a Topological Structure, ATS*)

This is the most known spatial autocorrelation defined on a molecular graph G as [Moreau and Broto, 1980a, 1980b; Broto, Moreau *et al.*, 1984a]

$$\text{ATS}_k = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A w_i \cdot w_j \cdot \delta(d_{ij}; k) = \frac{1}{2} \cdot (\mathbf{w}^T \cdot {}^k\mathbf{B} \cdot \mathbf{w})$$

where w is any atomic property, A is the number of atoms in a molecule, k is the lag, and d_{ij} is the topological distance between i th and j th atoms; $\delta(d_{ij}; k)$ is a Kronecker delta function equal to 1 if $d_{ij} = k$, zero otherwise. ${}^k\mathbf{B}$ is the k th order $\rightarrow$ *geodesic matrix*, whose elements are equal to 1 only for vertices v_i and v_j at topological distance k , and zero otherwise; $\mathbf{w}$ is the A -dimensional vector of atomic properties. The autocorrelation ATS_0 defined for the path of length zero is calculated as

$$\text{ATS}_0 = \sum_{i=1}^A w_i^2$$

that is, the sum of the squares of the atomic properties. Typical atomic properties are atomic masses, polarizabilities, charges, and electronegativities.

Moreau–Broto autocorrelations can be viewed as a special case of the $\rightarrow$ *interaction geodesic matrices*, from which other kinds of descriptors can also be derived.

It has to be noted that atomic properties w should be centered by subtracting the average property value in the molecule to obtain proper autocorrelation values. Hollas demonstrated that only if properties are centered, all autocorrelation descriptors are uncorrelated thus becoming more suitable for subsequent statistical analysis [Hollas, 2002].

For each atomic property w , the set of the autocorrelation terms defined for all existing topological distances in the graph is the **ATS descriptor** defined as

$$\{ATS_0, ATS_1, ATS_2, \dots, ATS_D\}_w$$

where D is the $\rightarrow$ *topological diameter*, that is, the maximum distance in the graph. The plot of the ATS descriptor is the corresponding autocorrelogram.

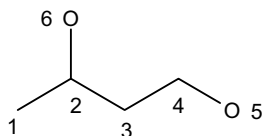
Average spatial autocorrelation descriptors are obtained by dividing each term by the corresponding number of contributions, thus avoiding any dependence on molecular size:

$$\overline{ATS}_k = \frac{1}{2\Delta_k} \cdot \sum_{i=1}^A \sum_{j=1}^A w_i \cdot w_j \cdot \delta(d_{ij}; k)$$

where Δ_k is the sum of the Kronecker delta, that is, the total number of vertex pairs at distance equal to k [Wagener, Sadowski *et al.*, 1995].

Example A7

Moreau–Broto autocorrelation descriptors calculated from the H-depleted molecular graph of 4-hydroxy-2-butanone.



$w_i = m_i$ atomic masses

$$ATS_0 = w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6^2$$

$$ATS_1 = w_1 \cdot w_2 + w_2 \cdot w_3 + w_3 \cdot w_4 + w_4 \cdot w_5 + w_2 \cdot w_6$$

$$ATS_2 = w_1 \cdot w_3 + w_1 \cdot w_6 + w_2 \cdot w_4 + w_3 \cdot w_5 + w_3 \cdot w_6$$

$$ATS_3 = w_1 \cdot w_4 + w_2 \cdot w_5 + w_4 \cdot w_6$$

$$ATS_0 = 12^2 + 12^2 + 12^2 + 12^2 + 16^2 + 16^2 = 1088$$

$$ATS_1 = 12 \cdot 12 + 12 \cdot 12 + 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 16 = 816$$

$$ATS_2 = 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 16 = 864$$

$$ATS_3 = 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 16 = 528$$

$$\overline{ATS}_0 = 1088/6 = 181.3$$

$$\overline{ATS}_1 = 816/5 = 163.2$$

$$\overline{ATS}_2 = 864/5 = 172.8$$

$$\overline{ATS}_3 = 528/3 = 176.0.$$

The ATS descriptor is a graph invariant describing how the property considered is distributed along the topological structure. Assuming an additive scheme, the ATS descriptor corresponds to a decomposition of the square molecular property Φ in different atomic contributions:

$$\Phi^2 = \left(\sum_{i=1}^A w_i \right)^2 = \sum_{i=1}^A w_i^2 + \sum_{i \neq j} 2 \cdot w_i \cdot w_j = ATS_0 + 2 \cdot \sum_{k=1}^D ATS_k$$

where ATS_0 contains all atomic contributions to the square molecular property and ATS_k the interactions between each pair of atoms.

• 3D molecular autocorrelation

Autocorrelation descriptors calculated for 3D spatial molecular geometry are based on interatomic distances collected in the $\rightarrow$ *geometry matrix* $\mathbf{G}$ instead of topological distances and the property function is still defined by the set of atomic properties.

The interatomic distance r is divided into elementary distance intervals of equal width (e.g., 0.5 Å). Each distance interval is defined by a lower and upper value of interatomic distance r_{ij} . All interatomic distances falling in the same interval are considered identical. For each distance interval, the autocorrelation function AC_k is obtained by summing all the products of the property values of atoms i and j whose interatomic distance r_{ij} falls within the considered interval $[r_u, r_v]_k$:

$$AC_k(r_u, r_v) = \sum_{i,j} w_i \cdot w_j \quad (r_u \leq r_{ij} \leq r_v)$$

 [Broto, Moreau *et al.*, 1984c; Broto and Devillers, 1990; Zakarya, Belkhadir *et al.*, 1993]

• Moran coefficient (I_k)

This is a general index of spatial autocorrelation that, if applied to a molecular graph, can be defined as

$$I_k = \frac{\frac{1}{\Delta_k} \cdot \sum_{i=1}^A \sum_{j=1}^A (w_i - \bar{w}) \cdot (w_j - \bar{w}) \cdot \delta(d_{ij}; k)}{\frac{1}{A} \cdot \sum_{i=1}^A (w_i - \bar{w})^2}$$

where w_i is any atomic property, $\bar{w}$ is its average value on the molecule, A is the number of atoms, k is the considered lag, d_{ij} is the topological distance between i th and j th atoms, and $\delta(d_{ij}; k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. Δ_k is the number of vertex pairs at distance equal to k [Moran, 1950].

Moran coefficient usually takes value in the interval $[-1, +1]$. Positive autocorrelation corresponds to positive values of the coefficient whereas negative autocorrelation produces negative values.

• Geary coefficient (c_k)

This is a general index of spatial autocorrelation that, if applied to a molecular graph, can be defined as

$$c_k = \frac{\frac{1}{2\Delta_k} \cdot \sum_{i=1}^A \sum_{j=1}^A (w_i - w_j)^2 \cdot \delta(d_{ij}; k)}{\frac{1}{(A-1)} \cdot \sum_{i=1}^A (w_i - \bar{w})^2}$$

where w_i is any atomic property, $\bar{w}$ is its average value on the molecule, A is the number of atoms, k is the lag considered, d_{ij} is the topological distance between i th and j th atoms, and $\delta(d_{ij}; k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. Δ_k is the number of vertex pairs at distance equal to k [Geary, 1954].

Geary coefficient is a distance-type function varying from zero to infinite. Strong autocorrelation produces low values of this index; moreover, positive autocorrelation translates in values between 0 and 1 whereas negative autocorrelation produces values larger than 1; therefore, the reference “no correlation” is $c_k = 1$.

Table A4 Some autocorrelation descriptors for the data set of phenethylamines (Appendix C – Set 2).

Mol.	X	Y	ATS ₁	ATS ₂	ATS ₃	ATS ₄	I ₁	I ₂	I ₃	I ₄	c ₁	c ₂	c ₃	c ₄
1	H	H	2.952	3.313	3.473	3.554	−0.006	−0.056	−0.139	−0.319	0.504	0.804	1.364	2.193
2	H	F	3.032	3.422	3.567	3.598	−0.006	−0.055	−0.134	−0.322	0.512	0.782	1.299	2.198
3	H	Cl	3.096	3.508	3.641	3.635	−0.006	−0.077	−0.177	−0.368	0.531	0.811	1.306	2.114
4	H	Br	3.251	3.708	3.819	3.728	−0.005	−0.098	−0.203	−0.320	0.549	0.838	1.174	1.485
5	H	I	3.392	3.884	3.977	3.818	−0.004	−0.090	−0.174	−0.223	0.542	0.828	1.053	1.042
6	H	Me	3.003	3.383	3.533	3.582	−0.005	−0.045	−0.112	−0.290	0.503	0.768	1.280	2.176
7	F	H	3.032	3.422	3.567	3.640	−0.006	−0.055	−0.134	−0.299	0.512	0.782	1.299	2.034
8	Cl	H	3.096	3.508	3.641	3.710	−0.006	−0.077	−0.177	−0.366	0.531	0.811	1.306	2.008
9	Br	H	3.251	3.708	3.819	3.876	−0.005	−0.098	−0.203	−0.374	0.549	0.838	1.174	1.645
10	I	H	3.392	3.884	3.977	4.027	−0.004	−0.090	−0.174	−0.300	0.542	0.828	1.053	1.363
11	Me	H	3.003	3.383	3.533	3.609	−0.005	−0.045	−0.112	−0.261	0.503	0.768	1.280	2.009
12	Cl	F	3.165	3.598	3.828	3.748	−0.006	−0.073	−0.159	−0.365	0.539	0.793	1.208	2.025
13	Br	F	3.310	3.783	4.081	3.909	−0.005	−0.089	−0.194	−0.364	0.554	0.816	1.235	1.655
14	Me	F	3.079	3.485	3.663	3.651	−0.005	−0.045	−0.106	−0.266	0.511	0.752	1.168	2.025
15	Cl	Cl	3.221	3.671	3.966	3.780	−0.007	−0.095	−0.159	−0.403	0.561	0.825	1.201	1.969
16	Br	Cl	3.359	3.843	4.264	3.936	−0.006	−0.109	−0.141	−0.398	0.574	0.845	1.180	1.655
17	Me	Cl	3.140	3.566	3.763	3.686	−0.006	−0.063	−0.157	−0.302	0.529	0.778	1.210	1.942
18	Cl	Br	3.359	3.843	4.264	3.861	−0.006	−0.109	−0.141	−0.333	0.574	0.845	1.180	1.417
19	Br	Br	3.480	3.991	4.636	4.006	−0.005	−0.130	−0.029	−0.372	0.594	0.875	1.069	1.386
20	Me	Br	3.289	3.756	3.993	3.775	−0.005	−0.080	−0.214	−0.257	0.545	0.802	1.257	1.361
21	Me	Me	3.052	3.449	3.617	3.636	−0.005	−0.037	−0.083	−0.241	0.503	0.740	1.149	2.008
22	Br	Me	3.289	3.756	3.993	3.897	−0.005	−0.080	−0.214	−0.342	0.545	0.802	1.257	1.633

ATS, Moreau–Broto autocorrelations; I, Moran coefficient; c, Geary coefficient. Calculations are based on the carbon-scaled atomic mass as the weighting scheme for atoms (see Table A3).

• Auto-Cross-Covariance transforms ($\equiv$ ACC transforms)

These are autocovariances and cross-covariances calculated from sequential data with the aim of transforming them into $\rightarrow$ *uniform-length descriptors* suitable for QSAR modeling. ACC transforms were originally proposed to describe peptide sequences [Wold, Jonsson *et al.*, 1993; Sjöström, Rännar *et al.*, 1995; Andersson, Sjöström *et al.*, 1998; Nyström, Andersson *et al.*, 2000]. To calculate ACC transforms, each amino acid position in the peptide sequence is defined in terms of three orthogonal $\rightarrow$ *z-scores*, derived from a $\rightarrow$ *Principal Component Analysis* (PCA) of 29 physico-chemical properties of the 20 coded amino acids.

Then, for each peptide sequence, auto- and cross-covariances with lags $k = 1, 2, \dots, K$, are calculated as

$$\text{ACC}_k(j, j) = \sum_{i=1}^{n-k} \frac{z_i(j) \cdot z_{i+k}(j)}{n-k} \quad \text{ACC}_k(j, m) = \sum_{i=1}^{n-k} \frac{z_i(j) \cdot z_{i+k}(m)}{n-k}$$

where j and m indicate two different z-scores, n is the number of amino acids in the sequence, and index i refers to amino acid position in the sequence. z-score values, being derived from PCA, are used directly because they are already mean centered.

ACC transforms were also used to encode information contained into $\rightarrow$ *CoMFA fields* (steric and electrostatic fields) using as the lag the distance between grid points along each coordinate axis, along the diagonal, or along any intermediate direction. The cross-correlation terms were calculated by the products of the $\rightarrow$ *interaction energy values* for steric and electrostatic fields in grid points at distances equal to the lag. Different kinds of interactions, namely, positive–positive, negative–negative, and positive–negative, were kept separated, thus resulting in 10 ACC terms for each lag. The major drawback of these ACC transforms is that their values depend on molecule orientation along the axes [Clementi, Cruciani *et al.*, 1993b; van de Waterbeemd, Clementi *et al.*, 1993].

• **TMACC descriptors** ($\equiv$ *Topological MAXimum Cross-Correlation descriptors*)

These are cross-correlation descriptors [Melville and Hirts, 2007] calculated by taking into account the topological distance d_{ij} between atoms i and j and four basic atomic properties: (1) Gasteiger–Marsili partial charges, accounting for electrostatic properties [Gasteiger and Marsili, 1980]; (2) Wildman–Crippen molar refractivity parameters, accounting for steric properties and polarizabilities [Wildman and Crippen, 1999]; (3) Wildman–Crippen log P parameters, accounting for hydrophobicity [Wildman and Crippen, 1999]; and (4) log S parameters, accounting for solubility and solvation phenomena [Hou, Xia *et al.*, 2004].

The general formula for the calculation of TMACC descriptors is

$$TMACC(P, P'; k) = \frac{1}{\Delta_k} \cdot \sum_{i=1}^A \sum_{j=1}^A P_i \cdot P'_j \cdot \delta(d_{ij}; k)$$

where P and P' are two atomic properties, A is the number of atoms in the molecule, k is the lag, d_{ij} is the topological distance between i th and j th atoms, Δ_k is the number of atom pairs located at topological distance k , and $\delta(d_{ij}; k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. If only one property is considered, that is, $P = P'$, autocorrelations are obtained.

Moreover, because all the selected properties, except for molar refractivity, contain both positive and negative values, these are treated as different properties and cross-correlation terms are also calculated between positive and negative values of each property. Therefore, 7 autocorrelation terms and 12 cross-correlation terms constitute the final TMACC descriptor vector.

• **DZ^K descriptors**

These are a modification of the Moreau–Broto autocorrelation descriptors defined by using the topological distance in conjunction with the properties of the atoms [Zakarya, Nohair *et al.*, 2000]:

$$DZ_w^K = \sum_{k=1}^K \left[k \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A (w_i \cdot w_j)^\alpha \cdot \delta(d_{ij}; k) \right]$$

where w is the selected atomic property, K the maximum considered topological distance, and α an exponent taking values 1 or 0.5. In particular, for $\alpha = 1$, the following expression holds:

$$DZ_w^K = \sum_{k=1}^K k \cdot ATS_k(w)$$

where ATS_k is the Moreau–Broto autocorrelation relative to lag k .

The use of atomic properties such as atom connectivity, electronegativity, van der Waals volume, and molar refraction was suggested.

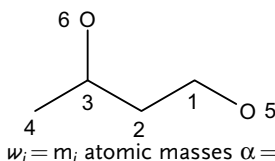
An extended form has been also proposed defined as

$${}^eDZ_w^K = \sum_{i=1}^A w_i + \sum_{k=1}^K \left[k \sum_{i=1}^{A-1} \sum_{j=i+1}^A (w_i \cdot w_j)^\alpha \cdot \delta(d_{ij}; k) \right] = \sum_{i=1}^A w_i + DZ_w^K$$

where the sum of the atomic properties is added to the autocorrelation term.

Example A8

DZ^k autocorrelation descriptors from the H-depleted molecular graph of 4-hydroxy-2-butanone.



$$DZ_m^1 = 1 \cdot ATS_1 = 816$$

$$DZ_m^2 = 1 \cdot ATS_1 + 2 \cdot ATS_2 = 816 + 2 \cdot 864 = 2544$$

$$DZ_m^3 = 1 \cdot ATS_1 + 2 \cdot ATS_2 + 3 \cdot ATS_3 = 816 + 2 \cdot 864 + 3 \cdot 528 = 4128$$

• Molecular Electronegativity Edge Vector (VMEE)

This is a modification of the Moreau–Broto autocorrelation defined by using reciprocal topological distances in conjunction with the Pauling atom electronegativities [Li, Fu *et al.*, 2001]. The autocorrelation value for each k th lag is calculated as

$$VMEE_k \equiv v_k = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{\chi_i^{\text{PA}} \cdot \chi_j^{\text{PA}}}{d_{ij}} \cdot \delta(d_{ij}; k), \quad k = 1, 2, 3, \dots$$

where χ^{PA} is the atomic electronegativity and d_{ij} is the topological distance between i th and j th atoms. This autocorrelation vector was used in modeling biological activities of dipeptides.

• 3D topological distance-based descriptors (S_k , X_k , I_k)

These are autocorrelation descriptors contemporarily based on topological and geometric distances, also called **3D TDB descriptors** [Klein, Kaiser *et al.*, 2004].

For each k th lag, steric descriptors, namely, **TDB-steric descriptors** S , are defined as

$$S_k = \frac{1}{\Delta_k} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A (R_i^{\text{cov}} \cdot r_{ij} \cdot R_j^{\text{cov}}) \cdot \delta(d_{ij}; k)$$

where Δ_k is the number of atom pairs located at a topological distance d_{ij} equal to k , r_{ij} is the geometric distance between i th and j th atoms, R^{cov} is the covalent radius of the atoms and δ is the Kronecker delta, which is equal to 1 when d_{ij} is equal to k and zero otherwise. In a similar way, electronic descriptors, namely, **TDB-electronic descriptors** X , are defined as

$$X_k = \frac{1}{\Delta_k} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A (\chi_i \cdot r_{ij} \cdot \chi_j) \cdot \delta(d_{ij}; k)$$

where χ is the sigma orbital electronegativity.

Together with steric and electronic descriptors, atom-type autocorrelation descriptors, namely, **TDB-atom type descriptors** I , are defined as

$$I_k(u, u) \equiv ATAC_k(u, u) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \delta_{ij}(u, u) \cdot \delta(d_{ij}; k)$$

where u denotes an atom type and $\delta_{ij}(u, u)$ is a Kronecker delta equal to 1 if both atoms i and j are of type u . These atom-type autocorrelations are calculated only for pairs of atoms of the same type. Moreover, unlike the previous two TDB descriptors (S_k and X_k), this autocorrelation descriptor does not account for 3D information.

- **Atomic Environment Autocorrelations (AEA)**

Aimed at characterizing the local environment of atoms, these descriptors are calculated by applying the autocorrelation function to encode spatial information relative to each single i th atom in a molecule as [Nohair, Zakarya *et al.*, 2002; Nohair and Zakarya, 2003]

$$AEA_{ik} = \sum_{j=1}^A (w_i \cdot w_j)^\alpha \cdot \delta(d_{ij}; k)$$

where w is any atomic property, α an adjustable parameter, and δ is the Kronecker δ function, which is equal to 1 when the topological distance d_{ij} between focused i th atom and any j th neighbor atom is equal to k . Atom connectivity, atomic van der Waals volume, and surface are among the suggested properties. Moreover, to take properties of the atoms along the i - j path of a topological distance d_{ij} equal to k also into account, a modified autocorrelation function was proposed as

$$AEA'_{ik} = \sum_{j=1}^A \left[w_i \cdot \left(\sum_m w_m \right)_{ij} \cdot w_j \right]^{1/(k+1)} \cdot \delta(d_{ij}; k)$$

where the exponent is the reciprocal of the number of atoms along the shortest path connecting vertices i and j ; w_i and w_j are properties of the two terminal vertices of the path, whereas w_m is the property of a vertex along the path.

- **Autocorrelation of Molecular Surface Properties (AMSP)**

This is a general approach for the description of property measures on the molecular surface by using uniform-length descriptors that consist of the same number of elements regardless of the

size of the molecule [Gasteiger, 2003a; Sadowski, Wagener *et al.*, 1995; Wagener, Sadowski *et al.*, 1995].

First, a number of points are randomly distributed on the molecular surface with a user-defined density and in an orderly manner to ensure a continuous surface. Then, the **Surface Autocorrelation Vector** (SAV) is derived by calculating for each lag k the sum of the products of the property values at two surface points located at a distance falling into the k th distance interval. This value is then normalized by the number Δ_k of the geometrical distances r_{ij} in the interval:

$$A(k) = \frac{1}{\Delta_k} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^N w_i \cdot w_j \cdot \delta(r_{ij}; k)$$

where N is the number of surface points and k represents a distance interval defined by a lower and upper bound.

It was demonstrated that to obtain the best surface autocorrelation vectors for QSAR modeling, the van der Waals surface is better than other molecular surfaces. Then, surface should have no fewer than five grid points per $\text{\AA}^2$, and a distance interval not more than 1 $\text{\AA}$ should be used in the distance binning scheme (Figure A1).

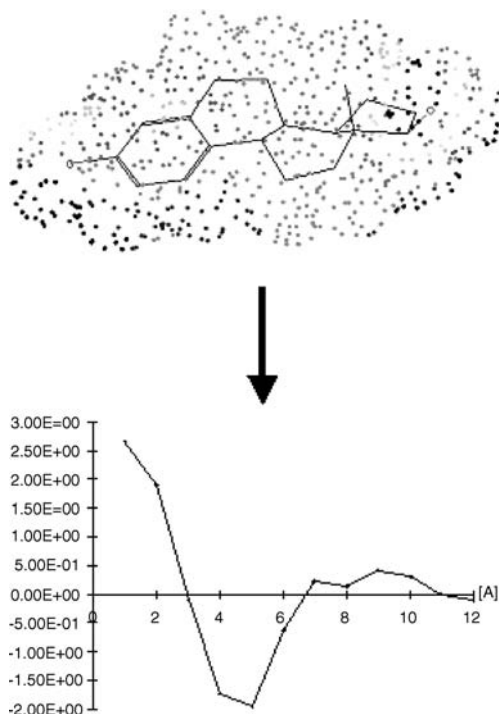


Figure A1 Surface autocorrelation vector of estradiol calculated by using MEP as the surface property.

📖 [Chastrette, Zakarya *et al.*, 1986; Devillers, Chambon *et al.*, 1986; Grassy and Lahana, 1993; Zakarya, Tiyal *et al.*, 1993; Clementi, Cruciani *et al.*, 1993b; van de Waterbeemd, Clementi *et al.*, 1993; Blin, Federici *et al.*, 1995; Sadowski, Wagener *et al.*, 1995; Bauknecht, Zell *et al.*, 1996; Patterson, Cramer III *et al.*, 1996; Huang, Song *et al.*, 1997; Anzali, Gasteiger *et al.*, 1998a; Legendre and Legendre, 1998; Devillers, 2000; Gancia, Bravi *et al.*, 2000; Gasteiger, 2003a; Moon, Song *et al.*, 2003; Cruciani, Baroni *et al.*, 2004]

- **Autocorrelation of a Topological Structure** $\equiv$ *Moreau–Broto autocorrelation* $\rightarrow$ autocorrelation descriptors
- **Autocorrelation of Molecular Surface Properties** $\rightarrow$ autocorrelation descriptors
- **autocorrelogram** $\rightarrow$ autocorrelation descriptors
- **Auto-Cross-Covariance transforms** $\rightarrow$ autocorrelation descriptors
- **autoignition temperature** $\rightarrow$ physico-chemical properties ($\odot$ flash point)
- **autometricity class** $\rightarrow$ topological information indices ($\odot$ autometricity index)
- **autometricity index** $\rightarrow$ topological information indices
- **automorphism group** $\rightarrow$ graph
- **Avalon fingerprints** $\rightarrow$ substructure descriptors ($\odot$ fingerprints)
- **average atom charge density** $\rightarrow$ quantum-chemical descriptors
- **average atom eccentricity** $\rightarrow$ distance matrix
- **average binding energy** $\rightarrow$ scoring functions
- **average bond charge density** $\rightarrow$ quantum-chemical descriptors
- **average cyclicity index** $\rightarrow$ detour matrix
- **average distance between pairs of bases** $\rightarrow$ biodescriptors ($\odot$ DNA sequences)
- **average distance/distance degree** $\rightarrow$ molecular geometry
- **average distance sum connectivity** $\equiv$ *Balaban distance connectivity index*
- **average electrophilic superdelocalizability** $\rightarrow$ quantum-chemical descriptors ($\odot$ electrophilic superdelocalizability)
- **average Fukui function** $\rightarrow$ quantum-chemical descriptors ($\odot$ Fukui functions)
- **average geometric distance degree** $\rightarrow$ molecular geometry
- **average graph distance degree** $\rightarrow$ distance matrix
- **average information content based on center** $\rightarrow$ centric indices
- **average local ionization energy** $\rightarrow$ quantum-chemical descriptors ($\odot$ electron density)
- **average molecular weight** $\rightarrow$ physico-chemical properties ($\odot$ molecular weight)
- **average nucleophilic superdelocalizability** $\rightarrow$ quantum-chemical descriptors ($\odot$ nucleophilic superdelocalizability)

■ average quasivalence number (AQVN)

It is a molecular descriptor calculated as average of the atomic numbers Z of the molecule atoms as [Veljković, Mouscadet *et al.*, 2007]

$$Z^* \equiv \text{AQVN} = \frac{\sum_{i=1}^A Z_i}{A}$$

where A is the number of atoms. It is used in the definition of the **electron-ion interaction potential** (EIIP), proposed to estimate long-range properties of biological molecules [Veljković, 1980] and defined as

$$\text{EIIP} = \frac{0.25 \cdot Z^* \cdot \sin(1.04 \cdot \pi \cdot Z^*)}{2\pi}$$

where Z^* is the average quasivalence number. Moreover, the ratio EIIP/AQVN was proposed as a $\rightarrow$ *drug-like index* for compounds.

- **average radical superdelocalizability** $\rightarrow$ quantum-chemical descriptors ($\odot$ radical superdelocalizability)
- **average row sum of the influence/distance matrix** $\rightarrow$ GETAWAY descriptors
- **average span** $\rightarrow$ size descriptors ($\odot$ span)
- **average vertex distance degree** $\rightarrow$ Balaban distance connectivity index
- **average writhe** $\rightarrow$ polymer descriptors
- **A weighting scheme** $\rightarrow$ weighting schemes
- **AH weighting scheme** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **AZV descriptors** $\rightarrow$ MPR approach
- **azzoo similarity coefficient** $\rightarrow$ similarity/diversity

B

- **backward Fukui function** → quantum-chemical descriptors (⊙ Fukui functions)
- **Balaban centric indices** → centric indices
- **Balaban DJ index** → Balaban distance connectivity index

■ Balaban distance connectivity index

The Balaban distance connectivity index (also called **distance connectivity index** or **average distance sum connectivity**), denoted as J , is one of the most known graph invariant. It is a very discriminating → *molecular descriptor* and its values do not increase substantially with molecule size or number of rings; it is defined in terms of the → *vertex distance degrees* σ_i , which are the row sums of the → *distance matrix* $\mathbf{D}$ [Balaban, 1982, 1983a]:

$$J = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2} = \frac{1}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\bar{\sigma}_i \cdot \bar{\sigma}_j)^{-1/2}$$

where σ_i and σ_j are the vertex distance degrees of the vertices v_i and v_j , a_{ij} the elements of the → *adjacency matrix* equal to one for pairs of adjacent vertices and zero otherwise, A the number of graph vertices, B the number of graph edges, and C the → *cyclomatic number*, that is, the number of rings. The denominator $C + 1$ is a normalization factor against the number of rings in the molecule. $\bar{\sigma}_i = \sigma_i/B$ is the **average vertex distance degree**; it was observed that within an isomeric series the average distance degrees are low in the more branched isomers.

To better discriminate among graph size, cyclicity, and branching, two modifications of the original Balaban distance connectivity index were later proposed [Balaban, Mills *et al.*, 2006]. The resulting indices, denoted as F and G , are defined as

$$F = B \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2} = (C+1) \cdot J$$

$$G = \frac{A^2 \cdot F}{A + C + 1} = \frac{A^2 \cdot (C+1)}{A + C + 1} \cdot J = \frac{A^2 \cdot B}{A + C + 1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2}$$

where the summation goes over all pairs of graph vertices but only pairs of adjacent vertices are accounted for by means of the elements a_{ij} of the adjacency matrix. A , B , and C are the number of vertices, edges, and rings, respectively. The index G is defined in terms of the index F and seems to be able to account for → *molecular complexity*.

Balaban-like indices are molecular descriptors calculated applying the same mathematical formula as the distance connectivity index J , but substituting the vertex distance degrees σ_i by row sums VS_i of $\rightarrow$ *graph-theoretical matrices* other than the distance matrix $\mathbf{D}$ or other $\rightarrow$ *local vertex invariants* $\mathcal{L}_i$. They are usually derived from $\rightarrow$ *weighted matrices* computed from vertex- and edge-weighted graphs, which properly represent molecules containing heteroatoms and/or multiple bonds:

$$J(\mathbf{M}; w) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (VS_i(\mathbf{M}; w) \cdot VS_j(\mathbf{M}; w))^{-1/2}$$

where $\mathbf{M}$ is a graph-theoretical matrix, a_{ij} the elements of the adjacency matrix $\mathbf{A}$ equal to one for pairs of adjacent vertices and zero otherwise, A the number of graph vertices, w the $\rightarrow$ *weighting scheme*, and VS the $\rightarrow$ *vertex sum operator* applied to the matrix $\mathbf{M}$.

This formula for the calculation of the Balaban-like indices was called **Ivanciuc–Balaban operator** by Ivanciuc and denoted as IB [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc, 2001c; Nikolić, Plavšić *et al.*, 2001].

The most general formula for computing Balaban-like indices in terms of any local vertex invariant is

$$J(\mathcal{L}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\mathcal{L}_i \cdot \mathcal{L}_j)^{-1/2}$$

where $\mathcal{L}_i$ and $\mathcal{L}_j$ are the local invariants of vertices v_i and v_j .

The **extended Ivanciuc–Balaban operator** was also defined as [Ivanciuc, Ivanciuc *et al.*, 2002e]

$$IB(\mathbf{M}; w, \lambda) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (VS_i(\mathbf{M}; w) \cdot VS_j(\mathbf{M}; w))^\lambda$$

where λ is a variable exponent.

The **J_t index** is a Balaban-like index defined as [Balaban, 1994a]

$$J_t = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (t_i \cdot t_j)^{-1/2}$$

where t_i and t_j are local invariants for vertices v_i and v_j defined as a combination of vertex distance degree σ and $\rightarrow$ *vertex degree* δ to obtain a greater discriminant power among isomers:

$$t_i = \frac{\sigma_i}{\delta_i}$$

where δ_i is the i th vertex degree of the vertex v_i . The idea behind these LOVIs is that usually the vertices with the highest distance sums have the lowest vertex degrees, thus enhancing the intramolecular differences.

The J index for multigraphs is calculated by the distance sums of the $\rightarrow$ *multigraph distance matrix* $\ast \mathbf{D}$ where the distances are obtained by weighting each edge with the reciprocal of its $\rightarrow$ *conventional bond order* ($\rightarrow$ *relative topological distance*):

$$J(*D) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (*\sigma_i \cdot *\sigma_j)^{-1/2}$$

where $*\sigma_i$ and $*\sigma_j$ are the $\rightarrow$ multigraph distance degrees of vertices v_i and v_j .

Example B1

Calculation of the Balaban distance connectivity index J and J_t index for 2-methylpentane. D is the topological distance matrix; σ_i and δ_i are the vertex distance sums and the vertex degrees. B equals 5 and C is zero.

Atom	1	2	3	4	5	6	σ_i	δ_i	t_i
1	0	1	2	3	4	2	12	1	12
2	1	0	1	2	3	1	8	3	2.667
3	2	1	0	1	2	2	8	2	4
4	3	2	1	0	1	3	10	2	5
5	4	3	2	1	0	4	14	1	14
6	2	1	2	3	4	0	12	1	12

$$J = \frac{B}{C+1} \times \left[(\sigma_1 \times \sigma_2)^{-1/2} + (\sigma_6 \times \sigma_2)^{-1/2} + (\sigma_2 \times \sigma_3)^{-1/2} + (\sigma_3 \times \sigma_4)^{-1/2} + (\sigma_4 \times \sigma_5)^{-1/2} \right] =$$

$$= 5 \times \left[(12 \times 8)^{-1/2} + (12 \times 8)^{-1/2} + (8 \times 8)^{-1/2} + (8 \times 10)^{-1/2} + (10 \times 14)^{-1/2} \right] = 2.6272$$

$$J_t = \frac{B}{C+1} \times \left[(t_1 \times t_2)^{-1/2} + (t_6 \times t_2)^{-1/2} + (t_2 \times t_3)^{-1/2} + (t_3 \times t_4)^{-1/2} + (t_4 \times t_5)^{-1/2} \right] =$$

$$= 5 \times \left[(12 \times 2.667)^{-1/2} + (12 \times 2.667)^{-1/2} + (2.667 \times 4)^{-1/2} + (4 \times 5)^{-1/2} + (5 \times 14)^{-1/2} \right] = 5.0141$$

To account for both bond multiplicity and heteroatoms, **Balaban modified distance connectivity indices** J^X and J^Y were proposed [Balaban, 1986a; Balaban, Catana *et al.*, 1990]. They are derived from the $\rightarrow$ multigraph distance matrix $*D$ as

$$J^X = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (*\sigma_i^X \cdot *\sigma_j^X)^{-1/2}$$

$$J^Y = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (*\sigma_i^Y \cdot *\sigma_j^Y)^{-1/2}$$

where B is the number of graph edges, C the number of graph rings, a_{ij} the elements of the adjacency matrix equal to one for pairs of adjacent vertices and zero otherwise, and A the number of graph vertices. Each edge is weighted by the inverse square root of the product of modified $\rightarrow$ multigraph distance degrees of the incident vertices according to the $\rightarrow$ X weighting scheme and $\rightarrow$ Y weighting scheme, respectively, as

$$*\sigma_i^X = X_i \cdot *\sigma_i = X_i \cdot \sum_{j=1}^A [*D]_{ij} \quad \text{and} \quad X_i = 0.4196 - 0.0078 \cdot Z_i + 0.1567 \cdot G_i$$

$${}^*\sigma_i^Y = Y_i \cdot {}^*\sigma_i = Y_i \cdot \sum_{j=1}^A [{}^*\mathbf{D}]_{ij} \quad \text{and} \quad Y_i = 1.1191 + 0.0160 \cdot Z_i - 0.0537 \cdot G_i$$

The quantities X and Y are recalculated atomic Sanderson electronegativities and covalent radii relative to carbon atom, respectively, obtained as a function of the atomic number Z_i and the group number of the Periodic System short form G_i of the atom; for atoms different from B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te, and I the X and Y values are set at one. X_i and Y_i are local indices that account for the presence of heteroatoms in the molecule.

The **3D Balaban index**, denoted as ${}^3\mathbf{D}$ J, is a Balaban-like index derived from the $\rightarrow$ geometry matrix $\mathbf{G}$ as [Mihalić, Nikolić *et al.*, 1992]

$${}^3\mathbf{D}(\mathbf{G}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot ({}^G\sigma_i \cdot {}^G\sigma_j)^{-1/2}$$

where ${}^G\sigma_i$ and ${}^G\sigma_j$ are the $\rightarrow$ geometric distance degrees of the vertices v_i and v_j , which are the row sums of the geometry matrix.

The **E-state topological parameter**, denoted as TI^E , is derived by applying the $\rightarrow$ Ivanciuc–Balaban operator to the $\rightarrow$ E-state index values used to characterize molecule atoms [Voelkel, 1994]:

$$\text{TI}^E = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (S_i \cdot S_j)^{-1/2}$$

where S_i and S_j are the E-state values for the vertices v_i and v_j .

It has to be pointed out that the proposed formula for the E-state topological parameter cannot be used for every molecule because it presents two drawbacks: (1) TI^E cannot be calculated when there exists one atom in the molecule with negative E-state value S ; (2) TI^E assumes very large values even when one S value tends to zero.

To overcome these drawbacks of the original formula, an alternative formula [Authors, This Book], adopted in the $\rightarrow$ DRAGON descriptors, is the following:

$$\text{TI}^E = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (1 + e^{S_i} \cdot e^{S_j})^{-1/2}$$

In this case, descriptor values can be obtained also for molecules with negative S values; moreover, they are in a suitable range for any molecule.

Other Balaban-like indices are $\rightarrow$ Balaban-like information indices, $\rightarrow$ Barysz index, $\rightarrow$ reversed Balaban index, $\rightarrow$ Harary–Balaban index, $\rightarrow$ Balaban-like resistance index, $\rightarrow$ variable Balaban index, $\rightarrow$ L_z index, $\rightarrow$ quotient Balaban index of the first kind, and $\rightarrow$ quotient Balaban index of the second kind.

The **Balaban DJ index** was still defined in terms of modified vertex distance degrees σ_i but using the formula of the $\rightarrow$ matrix sum indices as [Balaban and Diudea, 1993]

$$DJ = \sum_{i=1}^A d_{j_i} = \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot \left(\frac{\sigma_i}{w_i(1+f_i)} \cdot \frac{\sigma_j}{w_j(1+f_j)} \right)^{-1/2}$$

where A is the number of graph vertices, f the $\rightarrow$ *multigraph factor*, w a weighting factor accounting for heteroatoms, a_{ij} the elements of the adjacency matrix equal to one for pairs of adjacent vertices and zero otherwise, and d_j $\rightarrow$ *local vertex invariants* accounting for heteroatoms and bond multiplicity. When the factor w is equal to one and the multigraph factor is equal to zero then the index DJ is related to the Balaban index J by the following:

$$DJ = 2 \cdot \frac{C+1}{B} \cdot J = 2 \cdot \frac{C+1}{B} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2}$$

▢ [Balaban and Quintas, 1983; Barysz, Jashari *et al.*, 1983; Balaban and Filip, 1984; Balaban, Ionescu-Pallas *et al.*, 1985; Sabljic, 1985; Mekenyan, Bonchev *et al.*, 1987; Balaban and Ivanciuc, 1989; Balaban, Ciubotariu *et al.*, 1990; Balaban, Kier *et al.*, 1992a; Nikolic, Medicsaric *et al.*, 1993; Guo and Randic, 1999; Montanari, Cass *et al.*, 2000; Estrada and Gutierrez, 2001; Nikolic, Plavsic *et al.*, 2001; Balaban, Mills *et al.*, 2002; Ivanciuc, 2002a; Ivanciuc, Ivanciuc *et al.*, 2002e]

- **Balaban ID number** $\rightarrow$ ID numbers
- **Balaban-like information indices** $\rightarrow$ topological information indices
- **Balaban-like indices** $\rightarrow$ Balaban distance connectivity index
- **Balaban-like resistance index** $\rightarrow$ resistance matrix
- **Balaban modified distance connectivity indices** $\rightarrow$ Balaban distance connectivity index
- **Barnard keys** $\equiv$ *BCI keys* $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **Baroni–Urbani similarity coefficient** $\rightarrow$ similarity/diversity (Table S9)
- **Bartell resonance energy** $\rightarrow$ delocalization degree indices
- **barycenter** $\equiv$ *center of mass* $\rightarrow$ center of a molecule
- **Barysz index** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **Barysz distance matrix** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **basic graph** $\equiv$ *Sachs graph* $\rightarrow$ graph
- **basis of descriptors** $\rightarrow$ vectorial descriptors
- **Bate–Smith–Westall retention index** $\rightarrow$ chromatographic descriptors
- **BC(DEF) coordinates** $\equiv$ *BC(DEF) parameters*

■ **BC(DEF) parameters** ($\equiv$ *BC(DEF) coordinates*)

Proposed by Cramer III in 1980, they are five $\rightarrow$ *principal properties* (i.e., significant components calculated by $\rightarrow$ *Principal Component Analysis*) of a data matrix comprised of the values of six physico-chemical properties collected for 114 diverse liquid-state compounds [Cramer III, 1980a, 1983b].

The $\rightarrow$ *physico-chemical properties* used to derive BC(DEF) descriptors are activity coefficient in water, $\rightarrow$ *octanol–water partition coefficient*, boiling point, $\rightarrow$ *molar refractivity*, liquid state $\rightarrow$ *molar volume*, and heat of vaporization. The eigenvalues and corresponding cumulative explained variances of the five principal properties (denoted by B, C, D, E, and F) are reported in Table B1. It can be noted that the first two principal properties *B* and *C* already explain 95.7% of the original variance of the six physico-chemical properties; further analysis using different compounds and properties showed *B* and *C* to be independent of the data set used in their derivation, identifying them as measures of molecular bulk and cohesiveness, respectively. The other three parameters, *D*, *E*, and *F*, are of minor importance, however they were

retained due to their significance in the correlations involving some physico-chemical properties.

In general BC(DEF) parameters describe molecular properties related to nonspecific inter-molecular interactions in the liquid state and could therefore be useful in predicting biological activity or physico-chemical properties depending on such nonspecific interactions; 29 linear models were calculated by multivariate regression analysis that correlate BC(DEF) parameters to 29 different physico-chemical properties.

Table B1 Eigenvalues and cumulative variances of BC(DEF) principal properties.

Principal property	Eigenvalue	Cumulative variance (%)
B	3.870	64.4
C	1.870	95.7
D	0.168	98.5
E	0.045	99.2
F	0.029	99.7

Calculation of BC(DEF) parameters for new compounds different from the original 114 compounds can be accomplished either by their physico-chemical properties or their structure.

The property-derived BC(DEF) values are calculated from a set of known property values and the corresponding property models previously derived from the original 114×6 data set. A property model has the general form:

$$\gamma = b_0 + b_1 \cdot B + b_2 \cdot C + b_3 \cdot D + b_4 \cdot E + b_5 \cdot F$$

where γ is the known experimental property value and b the known regression coefficients taken from the specific property model. Using a set of at least six property models, all the BC(DEF) values together with their confidence intervals can be obtained as solutions of the linear equation system [Cramer III, 1983a]. In this case, the physico-chemical properties should be considered as independent variables and the BC(DEF) values as dependent variables.

Alternatively, the BC(DEF) values can be obtained by $\rightarrow$ *additive-constitutive models* based on the contributions of individual fragments and some correction factors to each parameter [Cramer III, 1980b]. A hierarchical additive-constitutive model was derived by multivariate regression of the BC(DEF) values of 112 original compounds (water and methane were excluded from the model) and occurrence frequencies of 35 molecular fragments. Moreover, in the same way a linear additive-constitutive model was also proposed; the fragment contributions to BC(DEF) parameters are reported in Table B2.

Table B2 Fragment contributions to BC(DEF) parameters.

Fragment	B	C	D	E	F
Intercept	-0.506	-0.056	0.007	0.031	0.028
-H	0.066	0.018	-0.027	-0.019	-0.019
-CH ₃	0.142	-0.020	-0.016	-0.023	-0.015
-CH ₂ -	0.076	-0.038	0.011	-0.004	0.003
>CH-	0.003	-0.058	0.053	0.018	0.015
>C<	-0.075	-0.076	0.091	0.043	0.034

(Continued)

Table B2 (Continued)

Fragment	B	C	D	E	F
–CH=CH–	0.147	–0.043	0.028	0.010	0.003
–CH=CH ₂	0.212	–0.025	0.000	–0.009	–0.015
>CH=CH ₂	0.147	–0.043	0.028	0.010	0.003
–C≡CH	0.171	0.074	0.027	0.002	–0.012
–C ₆ H ₅	0.467	–0.007	0.012	0.007	–0.017
≈CH–(aromatic)	0.088	0.002	–0.007	0.001	–0.003
–Naphthyl	0.766	0.018	–0.026	0.024	–0.028
–Cyclohexyl	0.489	–0.148	0.004	–0.029	–0.009
–F ^a	0.078	0.088	0.009	–0.019	–0.020
–Cl	0.165	0.087	–0.024	–0.012	–0.021
–Br ^a	0.213	0.095	–0.033	–0.008	–0.020
–I ^a	0.302	0.103	–0.056	–0.010	–0.031
–CF ₃	0.150	0.017	0.035	–0.037	–0.013
–CCl ₃	0.410	0.015	–0.009	–0.017	–0.017
–OH ^a	0.202	0.324	–0.012	–0.015	0.003
–O– ^a	0.044	0.155	0.061	0.019	–0.022
–C=O– ^a	0.135	0.246	0.061	0.023	–0.021
–CH=O ^a	0.219	0.244	0.010	–0.014	–0.027
–COO– ^a	0.167	0.170	0.062	0.015	–0.027
–COOH ^a	0.323	0.342	–0.011	–0.017	0.008
–NH ₂ ^a	0.167	0.269	0.037	0.027	–0.014
–NH– ^a	0.082	0.251	0.095	0.056	–0.010
–N– ^a	–0.006	0.189	0.125	0.069	0.014
–CN	0.241	0.269	–0.007	–0.023	–0.041
–N= ^a (pyridine)	0.102	0.183	0.031	–0.011	–0.020
–NO ₂ ^a	0.238	0.241	–0.012	–0.027	–0.037
–CONH ₂ ^a	0.444	0.499	–0.019	–0.039	–0.012
–S– ^a	0.136	0.130	0.028	0.032	–0.020
–SH ^a	0.231	0.155	–0.026	–0.011	–0.013

^aValue when attached to an aliphatic system.

- **BCF** $\equiv$ *bioconcentration factor* $\rightarrow$ environmental descriptors
- **BCI keys** $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **BCUT descriptors** $\rightarrow$ spectral indices ($\odot$ Burden eigenvalues)
- **benzene-likeness index** $\rightarrow$ delocalization degree indices
- **Bertz branching index** $\rightarrow$ molecular complexity ($\odot$ molecular branching)
- **Bertz complexity index** $\rightarrow$ molecular complexity
- **Bertz–Herndon relative complexity index** $\rightarrow$ molecular complexity
- **best hydrophilic volumes** $\rightarrow$ grid-based QSAR techniques ($\odot$ VolSurf descriptors)
- **best hydrophobic volumes** $\rightarrow$ grid-based QSAR techniques ($\odot$ VolSurf descriptors)
- **Beteringhe–Filip–Tarko descriptor** $\rightarrow$ MPR approach
- **Betti numbers** $\rightarrow$ Mezey 3D shape analysis
- **betweenness centrality** $\rightarrow$ center of a graph
- **Bhattacharyya distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)
- **bilinear indices** $\rightarrow$ TOMOCOMD descriptors

- **binary descriptors** → indicator variables
- **binary distance measures** → similarity/diversity
- **binary QSAR analysis** → scoring functions
- **binary sparse matrix** → algebraic operators (⊙ sparse matrices)
- **binding affinity** → drug design
- **binding property pairs** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **binding property torsions** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **binding site cavity** → drug design
- **binormalized centric index** → centric indices (⊙ Balaban centric index)
- **binormalized quadratic index** → Zagreb indices
- **bioaccumulation** → environmental indices
- **bioconcentration factor** → environmental indices

■ **biodescriptors**

These are numerical quantities encoding information about biochemical systems, addressing the problem of numerical characterization of macromolecules like proteins and nucleic acids and complex systems like proteomics maps.

The term *biodescriptors* was introduced by analogy with the common molecular descriptors, which are *chemodescriptors* since they are derived from the molecular structure of chemicals.

Biodescriptors cover a large field of mathematical strategies, including graph-theoretical approaches [Randić and Basak, 2002] and general theoretical links between topological structure of molecules and molecular biology networks were also investigated [Bonchev and Buck, 2007].

Most of biodescriptors were proposed to characterize sequences of peptides and nucleic acids trying to account for the sequential disposition of the constitutive elements of the considered biological system.

If each element of the sequence (e.g., an amino acid) is considered as an “atomic” unit, the physico-chemical properties of each atomic unit can be evaluated and used as “local bioinvariant” for the calculation of several descriptors defined for classical organic molecules.

The term **Quantitative Sequence-Activity Models** (QSAMs) is used instead of quantitative structure-activity relationships when referred to the research on relationships between structure and activities of molecules of biological interest [Jonsson, Norberg *et al.*, 1993].

Below, some common strategies for description of peptide and DNA sequences are briefly reviewed starting from amino acid descriptors, which are of fundamental importance for deriving most of the protein descriptors. The last section deals with some approaches for proteomics map characterization.

• **amino acid descriptors**

Due to the relevance and the complexity of proteins, some descriptors were defined to represent amino acid side chains, these being responsible for the packing of the regular elements of secondary structure and then for the tertiary structure of a protein. As a consequence, the structure of a protein can be expressed quantitatively by means of side chain amino acid properties. Starting from the pioneering work of Sneath, who described peptide sequences by semiquantitative experimental parameters of the 20 coded amino acids [Sneath, 1966], several amino acid descriptors have been proposed that contain information about properties of side chains of amino acids.

Ten principal properties were calculated by $\rightarrow$ *Principal Component Analysis* on 188 physico-chemical properties for the 20 coded amino acids [Kidera, Konisci *et al.*, 1985a, 1985b]. These 10 properties were called **KOKOS descriptors** by Pogliani on the basis of the Authors' names [Pogliani, 1994a]; they describe most of the conformational, bulk, hydrophobicity, α -helix, and β -structure properties of amino acids.

To calculate $\rightarrow$ *ACC transforms* of peptide sequences, each amino acid in the peptide sequence was described by three orthogonal $\rightarrow$ *z-scores* (Table B3), derived from a $\rightarrow$ *Principal Component Analysis* on 29 $\rightarrow$ *physico-chemical properties* of the 20 coded amino acids [Hellberg, Sjöström *et al.*, 1986, 1987a, 1987b; Wold, Eriksson *et al.*, 1987; Jonsson, Eriksson *et al.*, 1989]. These $\rightarrow$ *principal properties* were later extended to 87 amino acids including natural amino acids [Sandberg, Eriksson *et al.*, 1998]. Moreover, amino acid 3D principal properties (Table B3) were also derived from $\rightarrow$ *molecular interaction fields* [Norinder, 1991; Cocchi and Johansson, 1993; Cruciani, Baroni *et al.*, 2004].

Table B3 Z-scores (columns 4–6) and principal properties (PP) from molecular interaction fields, in the original form (columns 7–9) and reoriented and scaled between -1 and $+1$ (columns 10–12) [Cruciani, Baroni *et al.*, 2004].

ID	Code	Code	z_1	z_2	z_3	PP ₁ Polarity	PP ₂ Hydroph.	PP ₃ H-bond	PP ^S ₁ Polarity	PP ^S ₂ Hydroph.	PP ^S ₃ H-bond
1	Ala	A	10.07	-1.73	0.09	3.19	-2.21	-0.82	-0.96	-0.76	0.31
2	Arg	R	2.88	2.52	-3.44	-2.94	3.44	-2.56	0.80	0.63	0.99
3	Asn	N	3.22	1.45	0.84	-3.03	-1.45	-0.04	0.82	-0.57	0.02
4	Asp	D	3.64	1.13	2.36	-3.66	-2.74	2.60	1.00	-0.89	-1.00
5	Cys	C	0.71	-0.97	4.13	1.77	-1.02	-0.49	-0.55	-0.47	0.19
6	Glu	E	3.08	0.39	-0.07	-3.45	-1.34	2.58	0.94	-0.54	-0.99
7	Gln	Q	2.18	0.53	-1.14	-2.89	-0.34	0.99	0.78	-0.30	-0.38
8	Gly	G	2.23	-5.36	0.30	2.91	-3.20	-1.26	-0.88	-1.00	0.49
9	His	H	2.41	1.74	1.11	-2.51	0.43	-0.95	0.67	-0.11	0.37
10	Ile	I	-4.44	-1.68	-1.03	3.11	0.65	0.47	-0.94	-0.05	-0.18
11	Leu	L	-4.19	-1.03	-0.98	2.99	0.99	0.61	-0.90	0.03	-0.24
12	Lys	K	2.84	1.41	-3.14	-2.25	1.27	-2.60	0.60	0.10	1.00
13	Met	M	-2.49	-0.27	-0.41	2.69	1.01	0.22	-0.82	0.03	-0.08
14	Phe	F	-4.92	1.30	0.45	2.80	2.81	1.52	-0.85	0.48	-0.58
15	Pro	P	-1.22	0.88	2.23	2.65	-0.76	0.17	-0.81	-0.40	-0.07
16	Ser	S	1.96	-1.63	0.57	-1.58	-2.46	-1.48	0.41	-0.82	0.57
17	Thr	T	0.92	-2.09	-1.40	-1.55	-1.72	-0.95	0.40	-0.64	0.37
18	Trp	W	-4.75	3.65	0.85	-0.38	4.94	1.11	0.06	1.00	-0.47
19	Tyr	Y	-1.39	2.32	0.01	-1.23	2.59	0.51	0.31	0.42	-0.20
20	Val	V	-2.69	-2.53	-1.29	3.33	-0.87	0.36	-1.00	-0.43	-0.14

VHSE descriptor (*principal component score Vector of Hydrophobic, Steric, and Electronic properties*) is a $\rightarrow$ *vectorial descriptor*, containing eight principal properties, derived from Principal Component Analysis on 50 physico-chemical properties of the 20 coded amino acids [Mei, Liao *et al.*, 2005]. VHSE₁ and VHSE₂ are related to hydrophobic properties of amino acids, VHSE₃ and VHSE₄ to steric properties, and VHSE₅–VHSE₈ to electronic properties (Table B4).

Table B4 VHSE descriptor for the 20 coded amino acids [Mei, Liao *et al.*, 2005].

ID	Code	Code	VHSE ₁	VHSE ₂	VHSE ₃	VHSE ₄	VHSE ₅	VHSE ₆	VHSE ₇	VHSE ₈
1	Ala	A	0.15	-1.11	-1.35	-0.92	0.02	-0.91	0.36	-0.48
2	Arg	R	-1.47	1.45	1.24	1.27	1.55	1.47	1.30	0.83
3	Asn	N	-0.99	0.00	-0.37	0.69	-0.55	0.85	0.73	-0.80
4	Asp	D	-1.15	0.67	-0.41	-0.01	-2.68	1.31	0.03	0.56
5	Cys	C	0.18	-1.67	-0.46	-0.21	0.00	1.20	-1.61	-0.19
6	Gln	Q	-0.96	0.12	0.18	0.16	0.09	0.42	-0.20	-0.41
7	Glu	E	-1.18	0.40	0.10	0.36	-2.16	-0.17	0.91	0.02
8	Gly	G	-0.20	-1.53	-2.63	2.28	-0.53	-1.18	2.01	-1.34
9	His	H	-0.43	-0.25	0.37	0.19	0.51	1.28	0.93	0.65
10	Ile	I	1.27	-0.14	0.30	-1.80	0.30	-1.61	-0.16	-0.13
11	Leu	L	1.36	0.07	0.26	-0.80	0.22	-1.37	0.08	-0.62
12	Lys	K	-1.17	0.70	0.70	0.80	1.64	0.67	1.63	0.13
13	Met	M	1.01	-0.53	0.43	0.00	0.23	0.10	-0.86	-0.68
14	Phe	F	1.52	0.61	0.96	-0.16	0.25	0.28	-1.33	-0.20
15	Pro	P	0.22	-0.17	-0.50	0.05	-0.01	-1.34	-0.19	3.56
16	Ser	S	-0.67	-0.86	-1.07	-0.41	-0.32	0.27	-0.64	0.11
17	Thr	T	-0.34	-0.51	-0.55	-1.06	-0.06	-0.01	-0.79	0.39
18	Trp	W	1.50	2.06	1.79	0.75	0.75	-0.13	-1.01	-0.85
19	Tyr	Y	0.61	1.60	1.17	0.73	0.53	0.25	-0.96	-0.52
20	Val	V	0.76	-0.92	-0.17	-1.91	0.22	-1.40	-0.24	-0.03

SSIA descriptors (*Scores of Structural Information for Amino acids*) are z-scores derived from Principal Component Analysis on $\rightarrow 3D$ VAIF descriptors for the 20 coded amino acids [Zhou, Zhou *et al.*, 2006].

T-scale is a five-dimensional vectorial descriptor (Table B5) derived from Principal Component Analysis on 67 $\rightarrow$ topological indices of 135 amino acids [Tian, Zhou *et al.*, 2007].

Table B5 T-scale for the 20 coded amino acids [Tian, Zhou *et al.*, 2007].

ID	Code	T ₁	T ₂	T ₃	T ₄	T ₅	ID	Code	T ₁	T ₂	T ₃	T ₄	T ₅
1	Ala	-9.11	-1.63	0.63	1.04	2.26	11	Leu	-4.38	0.28	-0.49	1.45	0.02
2	Arg	0.23	3.89	-1.16	-0.39	-0.06	12	Lys	-2.59	2.34	-1.69	0.41	-0.21
3	Asn	-4.62	0.66	1.16	-0.22	0.93	13	Met	-4.08	0.98	-2.34	1.64	-0.79
4	Asp	-4.65	0.75	1.39	-0.40	1.05	14	Phe	0.49	-0.94	-0.63	-1.27	-0.44
5	Cys	-7.35	-0.86	-0.33	0.80	0.98	15	Pro	-5.11	-3.54	-0.53	-0.36	-0.29
6	Gln	-3.00	1.72	0.28	-0.39	0.33	16	Ser	-7.44	-0.65	0.68	-0.17	1.58
7	Glu	-3.03	1.82	0.51	-0.58	0.43	17	Thr	-5.97	-0.62	1.11	0.31	0.95
8	Gly	-10.61	-1.21	-0.12	0.75	3.25	18	Trp	5.73	-2.67	-0.07	-1.96	-0.54
9	His	-1.01	-1.31	0.01	-1.81	-0.21	19	Tyr	2.08	-0.47	0.07	-1.67	-0.35
10	Ile	-4.25	-0.28	-0.15	1.40	-0.21	20	Val	-5.87	-0.94	0.28	1.10	0.48

VSW descriptor (*Vector of principal component Scores for WHIMs*) is a vectorial descriptor derived from Principal Component Analysis on the 99 $\rightarrow$ WHIM descriptors calculated for the

20 coded amino acids [Tong, Liu *et al.*, 2008]. The VSW descriptor contains nine principal properties for each amino acid (Table B6).

Table B6 VSW descriptor for the 20 coded amino acids [Tong, Liu *et al.*, 2008].

ID	Code	Code	VSW ₁	VSW ₂	VSW ₃	VSW ₄	VSW ₅	VSW ₆	VSW ₇	VSW ₈	VSW ₉
1	Ala	A	-11.634	-1.897	1.978	-2.606	-1.715	-2.031	-0.818	0.640	1.080
2	Arg	R	11.871	-2.870	2.748	1.257	1.143	-0.477	-2.722	1.769	1.440
3	Asn	N	-5.350	7.683	4.117	4.174	4.249	-0.189	-1.065	-0.128	-0.839
4	Asp	D	-4.027	2.993	-3.359	-3.770	1.923	0.672	1.557	1.210	-0.301
5	Cys	C	-5.650	-2.879	-2.990	2.344	0.878	-1.945	1.069	-1.562	2.619
6	Gln	Q	2.176	-2.400	0.845	3.572	-1.201	-1.092	-0.114	0.052	-0.882
7	Glu	E	2.367	0.152	-4.048	0.804	2.037	0.990	1.087	2.345	0.166
8	Gly	G	-11.782	-13.698	3.470	0.201	0.965	3.074	0.440	-0.282	-0.702
9	His	H	2.339	0.361	-1.565	-1.076	2.002	-1.041	-1.300	-2.067	-1.570
10	Ile	I	0.412	6.404	-1.244	-1.622	-1.234	1.424	0.041	-0.179	-0.419
11	Leu	L	0.269	8.116	2.897	0.982	-1.934	3.156	0.058	-1.196	1.530
12	Lys	K	9.006	-2.097	-3.355	2.392	0.378	1.327	-0.462	-0.186	0.332
13	Met	M	4.363	-1.665	-3.977	-1.023	0.130	0.817	-0.540	-1.703	0.084
14	Phe	F	7.264	-4.366	-1.091	1.621	-3.196	-0.093	0.408	-0.110	-0.667
15	Pro	P	-5.307	3.184	0.595	4.277	-1.525	-1.512	3.067	0.320	-0.683
16	Ser	S	-9.155	2.320	-0.499	-2.269	-0.129	0.372	-0.853	0.997	0.546
17	Thr	T	-4.220	-0.272	-1.391	-2.538	1.070	-1.557	-1.338	-0.608	-0.558
18	Trp	W	11.702	0.162	5.620	-4.919	0.564	-0.732	2.216	-0.973	0.320
19	Tyr	Y	8.540	-1.526	1.741	-1.285	0.109	-0.953	1.142	1.100	-0.575
20	Val	V	-3.184	2.294	-0.492	-0.516	-4.515	-0.210	-1.874	0.561	-0.920

→ *Isotropic surface area* and → *electronic charge index* were proposed as the descriptors of steric character and local dipole of amino acid side chains [Collantes and Dunn III, 1995]. Moreover, amino acids were described, for example, by → *substituent descriptors* [Charton and Charton, 1982; Charton, 1990], → *connectivity indices* [Gardner, 1980; Pogliani, 1992a, 1992b, 1993a, 1993b, 1994a, 1994c, 1995a, 1996a, 1997a, 1997c, 1999a; Lučić, Nikolić *et al.*, 1995b], → *G-WHIM descriptors* [Zaliani and Gancia, 1999], → *side chain topological index* [Raychaudhury, Banerjee *et al.*, 1999], → *Molecular Holographic Distance Vector* [Liu, Yin *et al.*, 2001a], and → *WHIM descriptors* (Table B7) [Mauri, Ballabio *et al.*, 2008].

Table B7 Global WHIM descriptors for the 20 coded amino acids [Mauri, Ballabio *et al.*, 2008].

ID	Code	Code	Am	Km	Dm	ID	Code	Code	Am	Km	Dm
1	Ala	A	0.3634	0.4430	0.2330	11	Leu	L	1.1486	0.5210	0.3370
2	Arg	R	1.9266	0.7980	0.3130	12	Lys	K	1.5369	0.8120	0.3340
3	Asn	N	0.9274	0.4970	0.2960	13	Met	M	1.0385	0.4610	0.2940
4	Asp	D	0.8575	0.4290	0.3700	14	Phe	F	1.3731	0.5790	0.2710
5	Cys	C	0.6683	0.4990	0.2530	15	Pro	P	0.5536	0.4870	0.2910
6	Gln	Q	0.9970	0.5840	0.3810	16	Ser	S	0.4656	0.3910	0.2810
7	Glu	E	1.1128	0.4040	0.3260	17	Thr	T	0.6918	0.4850	0.3070
8	Gly	G	0.2343	0.5420	0.3220	18	Trp	W	2.3415	0.6410	0.2970

(Continued)

Table B7 (Continued)

ID	Code	Code	Am	Km	Dm	ID	Code	Code	Am	Km	Dm
9	His	H	1.0631	0.7590	0.2740	19	Tyr	Y	1.6385	0.6620	0.2840
10	Ile	I	0.9845	0.5820	0.2660	20	Val	V	0.7066	0.5100	0.2980

Am, Km, and Dm are the size, shape, and atom density global WHIM descriptors, respectively, weighted by the atomic masses.

One of the most comprehensive resources of amino acid properties freely available on line is the amino acid index database (*AAindex*), which includes numerical indices representing various physico-chemical, biochemical, and statistical properties of amino acids and pairs of amino acids. *AAindex* database has been made publicly available by the Japanese GenomeNet database service (<http://www.genome.jp/aaindex/>).

[Sneath, 1966; Wolfenden, Andersson *et al.*, 1981; Fauchère and Pliška, 1983; Sjöström and Wold, 1985; Abraham and Leo, 1987; Skagerberg, Sjöström *et al.*, 1987; Nakayama, Shigezumi *et al.*, 1988; Tsai, Testa *et al.*, 1991; El Tayar, Tsai *et al.*, 1992; El Tayar and Testa, 1993; Naray-Szabo and Balogh, 1993; Eriksson, Hermens *et al.*, 1995; Šoškić, Klaić *et al.*, 1995; Vallat, Gaillard *et al.*, 1995; Chapman, 1996; Pogliani, 1997a, 2000b; Randić and Krilov, 1997a; Sotomatsu-Niwa and Ogino, 1997; Pérez and Contreras, 1998; Grgas, Nikolić *et al.*, 1999; Raychaudhury and Nandy, 1999; Stein, Gordon *et al.*, 1999; Tao, Wang *et al.*, 1999; Testa, Raynaud *et al.*, 1999; Alifrangis, Christensen *et al.*, 2000; Nyström, Andersson *et al.*, 2000; Randić, Mills *et al.*, 2000; Nikolić and Raos, 2001; Oprea and Gottfries, 2001b; Pacios, 2001; Wold, Sjöström *et al.*, 2001; Shen, LeTiran *et al.*, 2002; Estrada, 2004b; Marrero-Ponce, Marrero *et al.*, 2004; Boon, Van Alsenoy *et al.*, 2005; Restrepo and Villaveces, 2005; Liang, Zhou *et al.*, 2006; Zhang, Ding *et al.*, 2008]

• **peptide sequences** ($\equiv$ *amino acid sequences*)

A peptide sequence is the ordered sequence of amino acid residues, connected by peptide bonds, which compose a peptide or protein. The sequence is generally reported from the N-terminal end containing free amino group to the C-terminal end containing free carboxyl group. Peptide sequences are often called **protein sequences** if they represent the protein primary structure. The primary structure of a peptide (or protein) is just the sequence of amino acids along its backbone. The secondary structure of proteins is defined by patterns of hydrogen bonds between backbone amide and carboxyl groups, while the tertiary structure is the three-dimensional structure, as defined by the atomic coordinates.

Side chains of amino acids are responsible for the packing of the regular elements of secondary structure and then for the tertiary structure of a protein. As a consequence, the structure of a protein can be expressed quantitatively by means of side chain amino acid properties. Several $\rightarrow$ *amino acid descriptors* have been proposed, which contain information about properties of side chains of amino acids.

A general index based on the $\rightarrow$ *total information content* of biological compounds, such as peptide sequences, is the **information index on amino acid composition**, defined as [Bonchev, 1983]

$$I_{\text{AAC}} = k \cdot \left(\ln N! - \sum_{g=1}^G \ln n_g! \right)$$

where k is the Boltzmann constant, N the total number of amino acid residues, G the number of $\rightarrow$ *equivalence classes*, that is, the number of different amino acid residues, and n_g the number of

amino acid residues of type g , that is, belonging to the g th equivalence class. Unlike other information indices, factorials are used in the expression to take into account combinations of amino acid residues.

A simple approach to protein description consists of representing a protein by a sequence of properties of its constituent amino acids. Each amino acid is described by one or more properties and therefore the total number of protein descriptors is given by the product of the number of amino acids in the protein and the number of selected amino acid properties. As this number of descriptors increases very fast with the size of proteins, this approach is usually applied to small- and medium-size peptides. Moreover, in QSAR studies that require $\rightarrow$ *uniform-length descriptors*, it can be used only to describe a series of peptide analogues, which are peptide sequences with the same length. To enable QSAR studies of peptide sequences with different length, some method is required that is able to translate the peptide sequences into $\rightarrow$ *vectorial descriptors* with the same number of variables. For example, $\rightarrow$ *ACC transforms* were applied to compress information about $\rightarrow$ *principal properties* of amino acids into peptide sequences with different length.

To characterize size and shape of side chains in amino acids, a topological descriptor was proposed [Raychaudhury, Banerjee *et al.*, 1999] based on a graph-theoretical approach applied to rooted weighted molecular graphs (hydrogen included) representing the side chains. Each vertex of the chain other than the link vertex (C_α carbon atom) is weighted and all shortest weighted paths between the link vertex C_α (assumed at zero position) and terminal vertices are taken into account; the weight of each path is given by the sum of the atomic weights of all involved atoms. Moreover, if there are more than one shortest path between two vertices, then the selected path is that with the minimum sum of the weights of its vertices.

A probability value p_i is assigned to the directed path connecting the link vertex to each i th terminal vertex (Figure B1), calculated as the following:

$$p_i = (\delta'_1 \cdot \dots \cdot \delta'_{i-1})^{-1}$$

where δ' is the number of incident bonds of each atom involved in the path without considering those already counted at the previous step. Bonds in the rings not involved in any path should be deleted to get probability values.

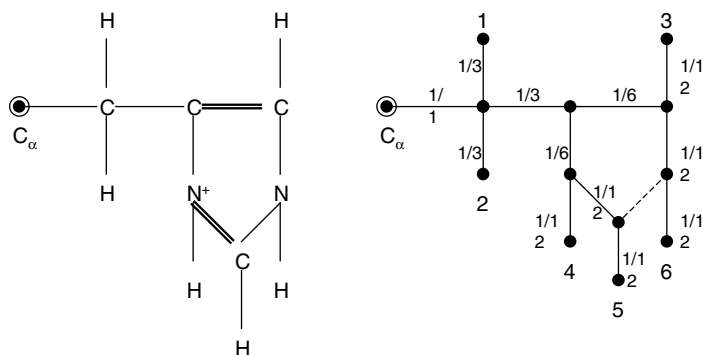


Figure B1 Histidine molecule and corresponding molecular graph. The number associated to each bond is the probability corresponding to the incident vertex, calculated starting from C_α .

By using the calculated probability values, the path value $P_{i,p}$ is calculated as

$$P_{i,p} = p_i \cdot \sum_k w_k$$

where the sum runs over all the vertices between the link atom and the i th vertex; w represents the weights of the atoms involved in the path. A molecular shape and size related index M_S^S (Figure B2), here called **side chain topological index**, is calculated for the link vertex C_α as the sum of all the path values:

$$M_S^S = \sum_{i=1}^{N_T} P_{i,p}$$

where N_T is the number of terminal vertices in the side chain.

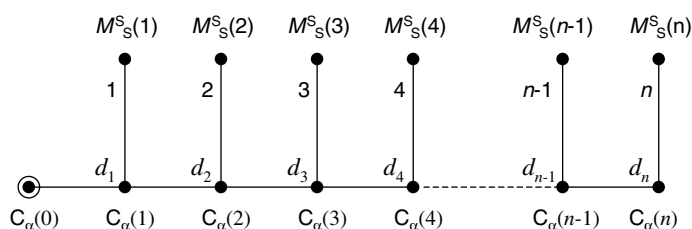


Figure B2 Amino acid sequence starting from $C_\alpha(0)$.

From this index, a descriptor for a sequence of amino acids, called **distance exponent index** D^x , was also proposed, defined as

$$D^x = \sum_k (M_S^S)_k \cdot d_k^x$$

where the sum runs over the considered sequence, being each term at topological distance d_k from the C_α representing the origin of the sequence. Each side chain topological index M_S^S is calculated independently of its link atom C_α . The exponent x may take any real values; values $x = -3$ and -4 were usefully proved in modeling side chain properties [Raychaudhury and Klopman, 1990].

A general approach to derive protein descriptors is based on representing a protein by a **macromolecular graph** in which vertices represent the α -carbon of the amino acid residues and edges represent the covalent peptidic bonds. Loops on vertices can be added to account for noncovalent interactions within a chain or between chains.

Then, $\rightarrow$ *amino acid descriptors* are used to weight vertices in the macromolecular graph in the same way as the atomic properties are used to weight vertices in a common molecular graph. At this stage, all the classic $\rightarrow$ *graph invariants* can be calculated from the weighted macromolecular graph and used as the protein descriptors in QSAR studies. Examples of these descriptors are linear, bilinear, and quadratic $\rightarrow$ *TOMOCOMD descriptors*.

By means of the macromolecular graph, the peptide description is simplified, considering that (a) the physico-chemical properties of the amino acids are responsible for the 3D structure and the functionality of the peptide and (b) all amino acids share common structural features, including an α -carbon to which an amino group, a carboxyl group, and a variable side chain are

bonded. The macromolecular graph allows reducing the complexity of the structures, since the number of amino acids in a peptide is significantly lower than the number of atoms.

To be able to calculate 3D descriptors, amino acids have to be characterized by (x, y, z) Cartesian coordinates. Being the α -carbon present in all coded amino acids, the Cartesian coordinates of that atom are selected as the coordinates of the whole amino acid. From the peptide topological representation and/or the corresponding geometrical representation (using only α -carbon spatial coordinates), several constitutional, topological and geometrical descriptors can be calculated.

For instance, several protein descriptors both topological and geometric were calculated by weighting amino acids with the $\rightarrow$ WHIM descriptors related to size (Am), shape (Km), and atom distribution density (Dm) of the single amino acids [Mauri, Ballabio *et al.*, 2008]. These amino acid properties were calculated on the isolated 3D structure of amino acids and are collected in Table B7.

An important characteristic of the 3D structure of proteins is the degree of folding of the protein chain. The degree is a quantitative measure of how folded a protein backbone is [Estrada and Uriarte, 2005]. Protein fold to optimize the conformational preferences of amino acids subject to local and global constraints. The $\rightarrow$ folding degree index obtained by diagonalization of the $\rightarrow$ distance/distance matrix [Randić, Kleiner *et al.*, 1994; Randić and Krilov, 1999] is an example of quantitative measure of folding degree, together with $\rightarrow$ molecular profiles [Randić and Krilov, 1997a]. Other size and/or shape descriptors of proteins and, in general, macromolecules, are the $\rightarrow$ characteristic ratio, $\rightarrow$ span, $\rightarrow$ Kuhn length, $\rightarrow$ end-to-end distance, $\rightarrow$ persistence length, and $\rightarrow$ radius of gyration.

Moreover, the **protein folding degree index**, denoted as I_3 , is based on the torsion angles of the protein backbone chain (ϕ , ψ , and ω) [Estrada, 2000, 2002a, 2004a; Estrada, Uriarte *et al.*, 2006]. The torsion angle ϕ_i describes the rotation about $N_i-C_{\alpha i}$ peptide bond, ψ_i the rotation about the $C_{\alpha i}-C_i$ peptidic bond, and ω_i describes the rotation around the C_i-N_{i+1} bond. A graph is defined whose vertices represent ϕ , ψ , and ω torsion angles and two vertices are connected if, and only if, the corresponding angles are contiguous in the backbone chain of the protein. Then, a matrix **B** is defined to represent the protein backbone as

$$\mathbf{B} = \mathbf{A} + \mathbf{T}$$

where **A** is the $\rightarrow$ adjacency matrix of the graph torsion angles and **T** is a diagonal matrix of the cosine of ϕ , ψ , and ω angles. Finally, the protein folding degree index I_3 is defined as

$$I_3 = \frac{1}{N-3} \cdot \sum_{j=1}^{N-3} e^{\lambda_j}$$

where N is the number of atoms in the protein backbone and λ_j are the eigenvalues of the **B** matrix. Note that this index is strictly related to the $\rightarrow$ Estrada index derived from the adjacency matrix of a molecular graph [Estrada and Hatano, 2007], then it can be expressed as the infinite sum of $\rightarrow$ spectral moments of **B** divided by $k!$ [Estrada, 2004b]. Consequently, the protein folding degree index can be interpreted as the sum of contributions from the sequences of torsion angles of different lengths, in such a way that large sequences of contiguous angles receive lower weights than shorter ones. It was shown that this index well describes the degree of folding of protein chains and that it takes larger values when the folded regions are close to the center of the chain. Moreover, it was demonstrated that local contribution of the i th amino acid

to the global protein folding is expressed as follows [Estrada, 2004b; Estrada and Uriarte, 2005; Estrada and Rodríguez-Velásquez, 2005b]:

$$I_3(i) = \sum_{j=1}^N e^{\lambda_j} \cdot \{[\ell_j(\psi_i)]^2 + [\ell_j(\phi_i)]^2\}$$

where $\ell_1, \ell_2, \dots, \ell_N$ are the eigenvectors of $\mathbf{B}$ associated to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_N$; $\ell_j(\psi_i)$ and $\ell_j(\phi_i)$ are the components of the eigenvector ℓ_j corresponding to the torsion angles ψ_i and ϕ_i of the i th amino acid (the angle ω_i is not considered because it corresponds to the peptidic bond shared by two contiguous amino acids).

📖 [Lee and Richards, 1971; Richards, 1977; Connolly, 1983b; Wagner, Colvin *et al.*, 1985; Eisenberg and McLachlan, 1986; Åqvist and Tapia, 1987; Arteca and Mezey, 1990; Wang, Shi *et al.*, 1990; Wold, Jonsson *et al.*, 1993; Leicester, Finney *et al.*, 1994b; Liang and Mislow, 1994; Kuz'min, Trigub *et al.*, 1995; Arteca, 1996; Poirrette, Artymiuk *et al.*, 1997; Andersson, Sjöström *et al.*, 1998; Štambuk, 1999; Lin and Lin, 2001; Liu, Yin *et al.*, 2001a, 2001b; Tusnády and Simon, 2001; Gironés, Amat *et al.*, 2002; Ivanciuc, Schein *et al.*, 2002; Torrens, 2002; Allen, Grant *et al.*, 2003; Ivanciuc, 2003d; Rost, Liu *et al.*, 2003; Ivanciuc, Oezguen *et al.*, 2004; Randić, Zupan *et al.*, 2004; Bai and Wang, 2005; Estrada, 2006b; Pissurlenkar, Malde *et al.*, 2007; Župerl, Pristovšek *et al.*, 2007]

• DNA sequences

The genome sequencing research projects are among the most challenging enterprises of these last decades. Elucidation of complete DNA sequences or protein sequences constitutes only a first step, while the further step lies in the interpretation of this huge number of data by automatic procedures.

A DNA sequence is a sequence of four letters A, T, G, and C that, respectively, denote four nucleic acid bases: adenine, thymine, guanine, and cytosine. RNA sequences contain the base uracil U in place of thymine T.

Graphical representations of DNA sequences were proposed by Hamori [Hamori, 1983, 1985, 1989], Gates [Gates, 1985], Nandy [Nandy, 1994, 1996a, 1996b; Nandy and Nandy 1995; Ray, Raychaudhury *et al.*, 1998; Nandy and Basak, 2000], and Leong and Mogenthaler [Leong and Mogenthaler, 1995].

The methods proposed by Gates and Nandy are based on choosing the four cardinal directions in (x, y) coordinate two-dimensional Cartesian system to represent the four bases in DNA sequences (Figure B3). The method essentially consists of plotting a point corresponding to a base by moving one unit in the positive or negative direction x - or y -axis depending on the defined association of a base with a cardinal direction. The cumulative plot of such points produces a graph that corresponds to the sequence.

In the Gates axis system (**TCAG-axis system**), one would move one unit in the positive x -direction for a cytosine (C), along the positive y -direction for a thymine (T), the negative x -direction for a guanine (G), the negative y -direction for an adenosine (A), implying a cumulative plot of the count of instantaneous C–G against T–A. The Nandy axis system (**CGTA-axis system**) associates G with positive x -direction, C with positive y -direction, A with negative x -direction, and T with negative y -direction (Figure B4). In the Leong and Mogenthaler axis system (**TAGC-axis system**), A is associated with positive x -direction, T with positive y -direction, C with negative x -direction, and G with negative y -direction.

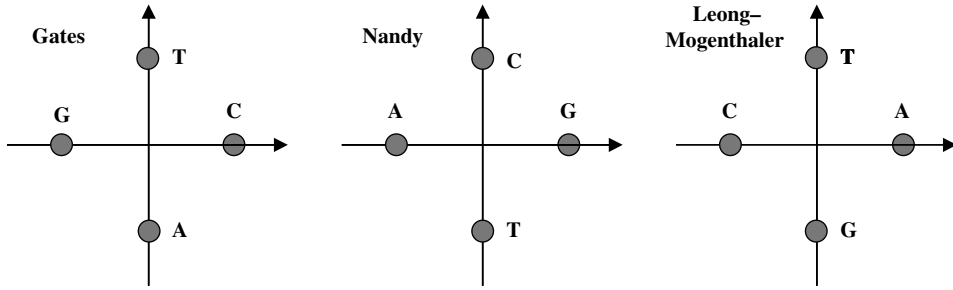


Figure B3 Graphical representation of a DNA sequence by two-dimensional Cartesian systems, as proposed by Gates, Nandy, and Leong-Mogenthaler.

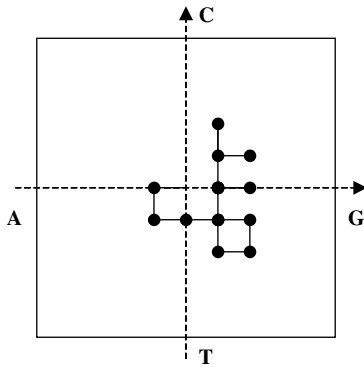


Figure B4 Graphical representations of a DNA sequence by the Nandy coordinate system.

A set of moments relative to the distribution of the graph points around the origin was proposed as the descriptor of the DNA sequence depicted by the Nandy graphical representation. These moments are derived as the weighted average of both x - and y -coordinates as [Raychaudhury and Nandy, 1999]

$$\mu_x = \frac{\sum_{i=1}^N x_i}{N} \quad \mu_y = \frac{\sum_{i=1}^N y_i}{N}$$

where N is the total length of the DNA sequence.

Then, the **graph radius**, denoted as g_R , was proposed as a further sequence descriptor, defined as

$$g_R = \sqrt{\mu_x^2 + \mu_y^2}$$

and the corresponding $\rightarrow$ *Euclidean distance* between two DNA sequences s and t was proposed as the measure of sequence dissimilarity:

$$d_{st} = \sqrt{(\mu_x^2(s) - \mu_x^2(t))^2 + (\mu_y^2(s) - \mu_y^2(t))^2}$$

To reduce the degeneracy of Nandy's graphical representation, an approach based on the idea to deviate from the original cardinal axes directions more than two of the four unit vectors that represent the corresponding bases [Liu, Guo *et al.*, 2002].

Starting from the Nandy's representation of DNA sequences, the eigenvalues obtained from the $\rightarrow$ *geometric distance/topological distance quotient matrix* and its increasing powers were used to perform similarity/diversity analysis of DNA sequences [Randić, 2000a].

Following the same philosophy of the previous graphical approaches, a representation into the 3D spaces was proposed assigning the four nucleic acid bases, the four directions associated with the regular tetrahedron [Randić, Vračko *et al.*, 2000]. To specify directions, the origin of the Cartesian (x , y , z) coordinate system was assigned in the center of a cube so that the four corners of the cube, which define the tetrahedral directions, constitute the main axes. Then, each basis is moved along the directions as arbitrarily defined in Table B8.

Table B8 The directions of the four nucleic bases in the tetrahedron space as proposed by [Randić, Vračko *et al.*, 2000].

Base	x	y	z
A	+1	-1	-1
G	-1	+1	-1
C	-1	-1	+1
T	+1	+1	+1

The $\rightarrow$ *leading eigenvalues* obtained from the $\rightarrow$ *geometric distance/topological distance quotient matrix* and its higher order matrices were proposed to describe the sequence. In any case, the degeneracy of this approach still remains large.

Still trying to remove degeneracy of the DNA sequence representations, another 2D representation was proposed moving the nucleic bases into the 2D plane, with coordinates that depend on an integer parameter d [Guo, Randić *et al.*, 2001] (Table B9). Values $d = 4$ and $d = 8$ were found to generate DNA graphical representations with lower degeneracy.

Table B9 The directions of the four nucleic bases in the tetrahedron space, as proposed by [Guo, Randić *et al.*, 2001].

Base	x	y	Base	x	y
A	-1	$+\frac{1}{d}$	C	$+\frac{1}{d}$	+1
G	+1	$+\frac{1}{d}$	T	$+\frac{1}{d}$	-1

Also in this approach, the $\rightarrow$ *leading eigenvalues* obtained from the $\rightarrow$ *geometric distance/topological distance quotient matrix* and its higher order matrices were proposed as the descriptors of DNA sequences.

In another approach, all the possible 64 combinations of three out of four nucleic bases, called triplets, were considered. A cubic matrix $4 \times 4 \times 4$ was constructed whose entries denote the frequencies of occurrence of all the 64 triplets in a DNA sequence [Randić, Guo *et al.*, 2001]. However, in practice, the cubic matrix is not used directly, but three groups of four bidimen-

sional matrices of size 4×4 are derived, each group of which contains all entries of the cubic matrix; thus a total of 12 matrices of size 4×4 are obtained.

From these 12 matrices, the $\rightarrow$ *leading eigenvalues* were calculated and used for similarity/diversity analysis.

Another 2D representation of the four nucleic bases of DNA sequences was proposed through a scatter plot where the x -axis is defined by the actual sequence of nucleic bases and the y -axis by the four ordered labels C, G, T, A for nucleic bases [Randić, Vračko *et al.*, 2003] (Figure B5).

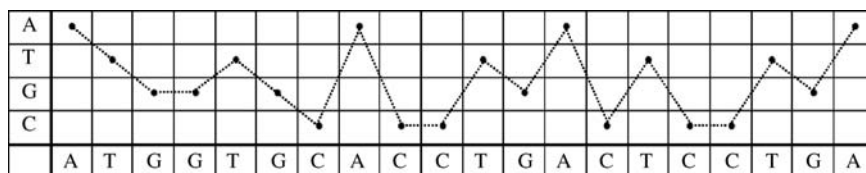


Figure B5 Scatter plot of a DNA sequence as proposed by [Randić, Vračko *et al.*, 2003].

Each point of the scatter plot indicates which basis is present in each position of the sequence. By joining two consecutive points by a line, a zigzag curve is obtained, which is a graphical representation of the sequence.

The numerical characterization of this zigzag curve is performed by the $\rightarrow$ *Euclidean-distance matrix*, where geometrical distances between every pair of vertices of the zigzag curve are collected, and two other $\rightarrow$ *graph-theoretical matrices*, called *M/M quotient matrix* and *L/L quotient matrix*. The **M/M quotient matrix** is a symmetric matrix whose off-diagonal elements are given as the ratio of the Euclidean distance between two vertices of the curve over the number of edges between the two vertices, that is, their $\rightarrow$ *topological distance*; diagonal elements are equal to zero. The M/M quotient matrix is the analogue of the $\rightarrow$ *geometric distance/topological distance quotient matrix* defined for molecular structures.

The **L/L quotient matrix** is a symmetric matrix whose off-diagonal elements are defined as the ratio of the Euclidean distance between two vertices of the curve over the sum of the geometrical lengths of the edges along the path connecting the two vertices. Note that this matrix is called $\rightarrow$ *quotient map matrix*, denoted as **Q**, in the framework of proteomics maps [Golbraikh, Bonchev *et al.*, 2001b; Randić, 2001e].

From these Euclidean-distance matrix, *M/M quotient matrix* and *L/L quotient matrix*, $\rightarrow$ *leading eigenvalues* were calculated to perform similarity/diversity analysis.

The **average distance between pairs of bases** (X, Y), being X, Y = A, C, G, T, is another descriptor of DNA sequences [Randić and Basak, 2001b]. To calculate this descriptor, the 16 pairs of DNA bases are arranged into a square matrix as

AA	AC	AG	AT
CA	CC	CG	CT
GA	GC	GG	GT
TA	TC	TG	TT

For each element of type X in the sequence, the $\rightarrow$ *topological distance* in the sequence between this element and the next one of type Y is recorded into a matrix, which has size $(N_X \times N_Y)$, where N_X and N_Y are the number of occurrences of the basis type X and Y.

Note that calculating the distance between pairs, their order is not considered, that is, the pair XY is considered the same as the pair YX. For example, given a DNA sequence as

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	G	G	T	G	C	A	C	C	T	G	A	C	T	C	C	T	G	A

the pair AA is characterized by the following distance matrix:

A/A	1	8	13	20
1	0	7	12	19
8	7	0	5	12
13	12	5	0	7
20	19	12	7	0

and the pair CA by the following unsymmetrical distance matrix:

A/C	7	9	10	14	16	17
1	6	8	9	13	15	16
8	1	1	2	6	8	9
13	6	4	3	1	3	4
20	13	11	10	6	4	3

From each matrix **XY**, the average distance between bases X and Y is calculated as

$$\overline{XY} = \frac{\sum_i \sum_j [XY]_{ij}}{N_X \cdot N_Y}$$

where $[XY]_{ij}$ are the elements of the matrix **XY** and N_X and N_Y are the number of occurrences of the bases of type X and Y, respectively.

Then, for the example given above, the average sum of distances between pairs AA in the 4×4 matrix is $124/(4 \times 4) = 7.75$, while for the AC pair is $162/(6 \times 4) = 6.75$.

Characteristic sequences of DNA were defined in terms of three different classification criteria [He and Wang, 2002]. All the possible combinations of two out four nucleic acid bases were assigned to six different classes:

$R = \{A, G\}$ $Y = \{C, T\}$ classification based on chemical structure
 $M = \{A, C\}$ $K = \{G, T\}$ classification based on distinguishing amino/keto groups
 $W = \{A, T\}$ $S = \{G, C\}$ classification based on hydrogen bond strength

The first classification criterion consists in assigning each basis in a DNA sequence a value 1 if the basis belongs to the class *R*, that is, is A or G, and zero if the basis belongs to the class *Y*, that is, the basis is C or T. Similar operations are performed considering the other class partitions, thus obtaining three binary vectors of characteristic sequences (*R*, *Y*), (*M*, *K*), and (*W*, *S*), each having length equal to the length of the DNA sequence.

Exploiting this binary representation of sequences, three $2 \times 2 \times 2$ cubic matrices are generated, accounting for the eight possible triplets in each characteristic sequence: 000, 001, 010, 011, 100, 101, 110, and 111. The entries of these matrices are defined as

$$f_{ijk}^X = \frac{100 \cdot m_{ijk}^X}{N-2}$$

where m_{ijk} is the number of occurrences of the triplet $i-j-k$ in X and N is the length of the string (i.e., the length of the DNA sequence). X stands for one of the three classes, (R , Y), (M , K), or (W , S).

The $2 \times 2 \times 2$ cubic matrices are splitted into six 2×2 matrices, considering separately the four entries with triplets beginning with zero (F_0^X) and the four entries with triplets beginning with one (F_1^X), as shown in Figure B6.

F_0^R	0	1	F_1^R	0	1
0	10	10	0	10	13
1	9	14	1	13	20

Figure B6 2×2 matrices for class (R , Y), according to the He–Wang approach.

The $\rightarrow$ *leading eigenvalues* of the six matrices were proposed to describe the whole DNA sequence.

Another approach to the description of DNA sequences is based on the partial-ordering given by the $\rightarrow$ *Hasse diagrams*. The order relationships between the C, T, A, G bases of a DNA sequence are recorded into the $\rightarrow$ *Hasse matrix* [Todeschini, Consonni *et al.*, 2006]. The variables used in building the Hasse matrix are the position of the bases in the sequence and a physico-chemical property of the bases, such as the mass. For example, the same small DNA sequence given above:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	T	G	G	T	G	C	A	C	C	T	G	A	C	T	C	C	T	G	A

can be described as shown in Table B10. From these data, the Hasse matrix (20×20) is calculated simply comparing, for each pair of bases in the sequence, the values of the two

Table B10 Data relative to the sequence in the text.

Basis	ID	MW
A	1	135.13
T	2	126.00
G	3	151.13
G	4	151.13
T	5	126.00
...	...	...
...	...	...
T	18	111.10
G	19	135.13
A	20	151.13

ID is the basis position in the sequence and MW the corresponding molecular weight.

variables, that is, position in the sequence (ID) and molecular weight (MW). The obtained corresponding Hasse diagram is shown in Figure B7.

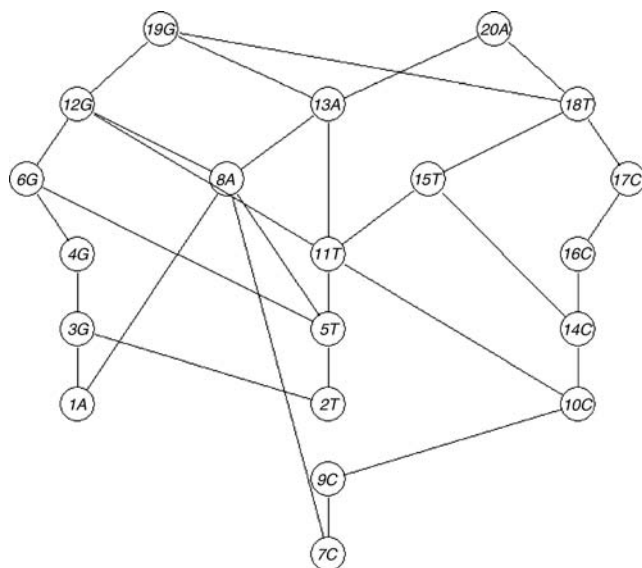


Figure B7 The Hasse diagram obtained by the sequence in the text. For each element, the number corresponds to its absolute position in the sequence.

From each property that ranks the four bases in a different way, a different Hasse diagram is obtained. Descriptors of the DNA sequence are finally obtained from the absolute values of the Hasse matrix elements and the largest eigenvalue was proposed to analyze similarity/diversity of DNA sequences.

📖 [Le, Nussinov *et al.*, 1989; Shapiro and Zhang, 1990; Wold, Jonsson *et al.*, 1993; Norinder, 1994; Bucher, Karplus *et al.*, 1996; Ray, Raychaudhury *et al.*, 1998; Randić, 2000b; Randić and Vračko, 2000; Štambuk, 2000; Nandy, Nandy *et al.*, 2002; Gan, Pasquali *et al.*, 2003; Guo and Nandy, 2003; Nandy and Nandy, 2003; Randić, Vračko *et al.*, 2003; Randić and Balaban, 2003; Yan, Wang *et al.*, 2003; Yuan, Liao *et al.*, 2003; Dobeš, Kmunicek *et al.*, 2004; Liao and Wang, 2004a, 2004b, 2004c, 2004d, 2005; Liao and Ding, 2005; Liao, Zhang *et al.*, 2005; Liao, Tan *et al.*, 2005a, 2005b; Liao, 2005; Liao, Ding *et al.*, 2005; Liao, Wang *et al.*, 2005; Nandy and Basak, 2005; Randić, Lerš *et al.*, 2005a; Zhang, Liao *et al.*, 2005; Dai, Liu *et al.*, 2006; Gao and Zhang, 2006a, 2006b; Luo, Liao *et al.*, 2006; Nandy, Harle *et al.*, 2006; Randić, Novič *et al.*, 2006; Wang and Wang, 2006; Zhang, Liao *et al.*, 2006; Zhang and Chen, 2006; Liao, Zhu *et al.*, 2007; Nandy, Basak *et al.*, 2007; Zhang, Luo *et al.*, 2007; Zhang, 2007]

• proteomics maps

Proteomics maps, together with NMR spectral maps, graphical representation of DNA, and protein sequences, belong to the general class of graphical and visual data represented by a 2D

map [Jeffrey, 1990; Blackstock and Weir, 1999; Bradfield, 2004]. A map is intended as a region of a plane where N discrete points are described by their Cartesian coordinates and relative intensities. **Map invariants** are numerical quantities that characterize the map and are independent of the orientation of coordinate axes and labeling of points [Randić, Lerš *et al.*, 2004b]. Advantages of numerical characterization of 2D maps are the possibility of visual data storage in digital format, significant data compression, quantitative evaluation of $\rightarrow$ *similarity/diversity* between maps and modeling of relationships between the structure of foreign agents and, for instance, the effects they have on a proteome.

A proteomics map is the result of horizontal separation of proteins by electrophoresis and vertical separation by chromatography, so that proteins on the left of the map have greater charge and proteins at the top have greater mass [Bajzer, Randić *et al.*, 2003]. Proteomics data can be reported as tables of x, y Cartesian coordinates (i.e., charge and mass) of protein spots and their abundance, or formatted into the *bubble diagram* (Figure B8), in which a point, whose coordinates represent charge and mass of a protein, is the center of a circle with radius proportional to the abundance of that protein [Randić, Witzmann *et al.*, 2001].

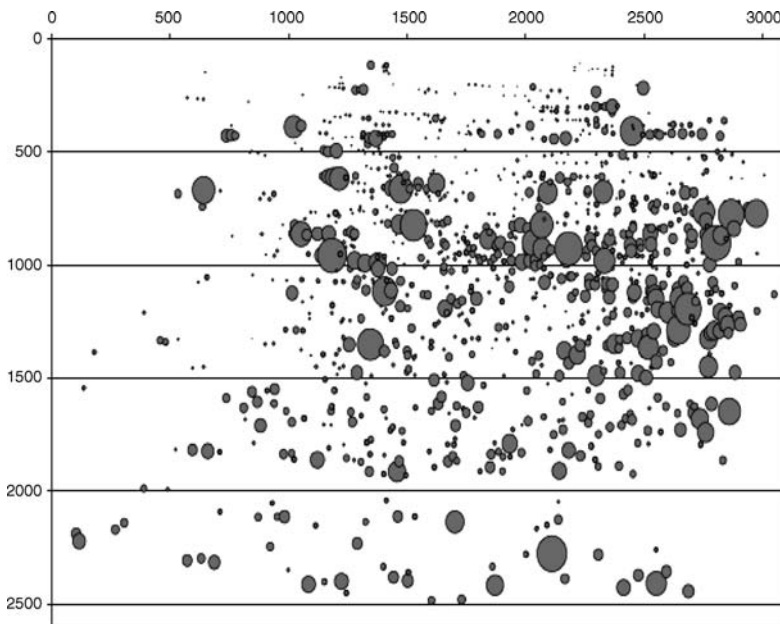


Figure B8 Example of a proteomics map, after a preliminary data pretreatment [Randić, Witzmann *et al.*, 2001].

Alternatively, proteins can be represented by points in three-dimensional space where x, y and z coordinates are proportional to charge, mass, and abundance, respectively. Since proteomics data represent different physical quantities, they are usually scaled, for instance, to the interval $(-1, 1)$ or in such a way as the average charge, the average mass, and the average abundance are all equal to one or to a selected reference value [Bajzer, Randić *et al.*, 2003].

To generate map invariants, the following procedure is used [Randić, 2001e, 2002a; Randić, Witzmann *et al.*, 2001; Randić, Zupan *et al.*, 2001; Randić and Basak, 2002; Randić, Novič *et al.*,

2002]. The first step consists of associating a suitable mathematical object of fixed geometry with a map; then, for the selected mathematical object a numerical representation is constructed in the form of a matrix; once a matrix representing the map has been derived, local invariants and matrix invariants can be calculated in a similar way to $\rightarrow$ *local vertex invariants* and $\rightarrow$ *graph invariants* which encode information about a molecular graph.

Examples of mathematical objects used to generate a matrix representation of a map are (1) an *embedded zigzag curve* (or *embedded path graph*), (2) an *embedded graph of partial ordering*, (3) an *embedded cluster graph*, and (4) an *embedded neighborhood graph*. These will be briefly explained below.

To construct an **embedded zigzag curve**, first points in the map are ordered by assigning them with labels that rank points relative to their abundance giving the most abundant protein point label 1 [Randić, 2001e; Randić, Zupan *et al.*, 2001; Randić, Witzmann *et al.*, 2001; Randić, Novič *et al.*, 2002]. Then, points with adjacent numerical labels are connected by an edge thus resulting into a complicated path, which overlaps itself many times; this path is called zigzag curve. The zigzag curve is then the result of a total ordering of protein spots relative to their abundance (Figure B9).

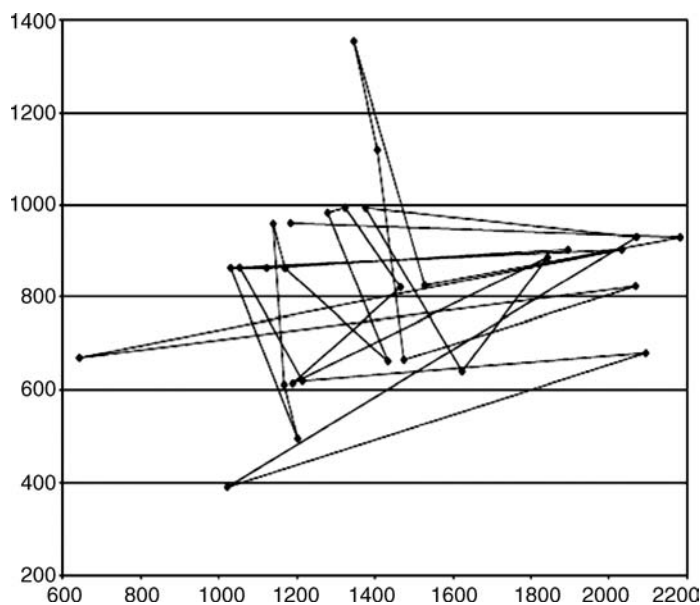


Figure B9 Zigzag curve of a proteomics map [Randić, Witzmann *et al.*, 2001].

The **embedded graph of partial order** is based on the partial order of proteins obtained by ordering them relatively to their charge and mass, respectively [Randić and Basak, 2002; Randić, 2002a; Randić, Zupan *et al.*, 2002]. In this graph, only those protein spots are connected that either dominate or are dominated in both the mass and the charge by the neighboring spots. If a direction from left to right is associated with each connection line, then a directed graph is obtained, which leads to an $\rightarrow$ *adjacency matrix* with positive and negative values depending on the direction of the edge connecting two vertices.

In the embedded graph of partial order the vertices, representing protein spots, are at fixed geometrical location and all the edges have positive slopes; this is a consequence of the partial order in which vertices at the top and right location dominate vertices which are at lower height and shifted towards the left.

The **embedded cluster graph** is obtained by making connections between the protein spots that are separated by Euclidean geometrical distances shorter than or equal to a selected critical distance [Randić and Basak, 2002; Bajzer, Randić *et al.*, 2003].

The **embedded neighborhood graph** is constructed by using the following procedure [Randić, Lerš *et al.*, 2004a, 2004b; Randić, Novič *et al.*, 2005]. First, x and y coordinates are scaled dividing them by the maximal Euclidean distance between two spots in the map. Relative abundances are calculated by dividing each protein abundance by the abundance of the protein corresponding to the spot with label 1, which is the maximally abundant protein. Then, the clustering method KNN is applied as follows: Euclidean distances between pairs of protein spots are calculated, for each spot a short list of the nearest neighbors is constructed, and, finally, the considered spot is connected by lines with its nearest neighbors. Different graphs are obtained by varying the number of nearest neighbors. The nearest neighbors can be 2D if Euclidean distance between spots are calculated in the space defined by the coordinates (x, y) or 3D if distances are calculated using (x, y, z) coordinates, where z refers to protein abundance. In any case, when dealing with 2D neighborhood graphs, information about relative abundances of spots can be accounted for by assigning each spot a two-component vector containing a local invariant derived from a map matrix $\mathbf{M}$ (e.g., the matrix row sum) and relative abundance z . Then, the length of this vector is

$$|v_i| = \sqrt{\left(\sum_{j=1}^N [\mathbf{M}]_{ij}\right)^2 + z_i^2}$$

where $\mathbf{M}$ is any matrix describing relationships among protein spots and N is the number of spots. The vector length can be used as the descriptor of each spot and the average length of the vectors of all the spots as a map descriptor.

From the selected graph representation of a proteomics map, different *map matrices* can be derived which encode information about distances and adjacency between protein spots. Examples of these matrices are reported below.

The **Euclidean-distance map matrix**, denoted as **ED**, is the analogue of the $\rightarrow$ *geometry matrix* $\mathbf{G}$ derived from a molecular graph. In this case, vertices of the map graph are assigned (x, y) or (x, y, z) coordinates, z being intended as the $\rightarrow$ *weighting scheme* for vertices; it is defined as [Bajzer, Randić *et al.*, 2003]

$$[\mathbf{ED}]_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

where x , y , and z are the spatial coordinates of a protein spot (x and y) in the map and its abundance (z), respectively. The Euclidean-distance map matrix can also be calculated by considering only x and y coordinates. To describe the embedded zigzag curve, Euclidean distances through space were measured directly (e.g., in millimeters) from the map for all the pairs of vertices [Randić, 2001e].

The **path-distance map matrix**, denoted as **PD**, resembling the $\rightarrow$ *bond length-weighted distance matrix* of a molecular graph, is defined as [Bajzer, Randić *et al.*, 2003]

$$[\mathbf{PD}]_{ij} = \min_{p_{ij}} \left(\sum_{kq} [\mathbf{ED}]_{kq} \right)_{ij}$$

where $[\mathbf{ED}]_{kq}$ denotes entries of the Euclidian-distance map matrix, p_{ij} a path connecting vertices i and j , and the summation goes over all the pairs of adjacent vertices along the considered path. Then, each entry of the path-distance map matrix is the shortest distance between two vertices measured along the path by summing the geometrical length of the edges connecting adjacent vertices along the path.

The **quotient map matrix**, denoted as $\mathbf{Q}$, is defined in a similar way to the $\rightarrow$ *geometric distance/topological distance quotient matrix*, whose entries are defined in terms of the ratio of distances between a pair of vertices measured through the space and along the bonds. The elements of the quotient matrix $\mathbf{Q}$ are formally defined as [Randić, 2001e; Bajzer, Randić *et al.*, 2003]

$$[\mathbf{Q}]_{ij} = \begin{cases} \frac{[\mathbf{ED}]_{ij}}{[\mathbf{PD}]_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $[\mathbf{ED}]_{ij}$ and $[\mathbf{PD}]_{ij}$ are the elements of the Euclidean-distance and path-distance map matrix, respectively. The elements of the quotient matrix are equal to 1 for all the pairs of adjacent protein spots and smaller than 1 for pairs of nonadjacent spots.

The **neighborhood-distance map matrix**, denoted as $\mathbf{ND}$, encodes information about vertex proximities; its elements are different from zero only for those pairs of protein spots that are within a certain neighborhood. This matrix is the analogue of the $\rightarrow$ *neighborhood geometry matrix* derived from a molecular graph; then it is defined as [Bajzer, Randić *et al.*, 2003]

$$[\mathbf{ND}]_{ij} = \begin{cases} [\mathbf{ED}]_{ij} & \text{if } [\mathbf{ED}]_{ij} \leq D_C \\ 0 & \text{if } [\mathbf{ED}]_{ij} > D_C \text{ or } i, j \text{ are not connected} \end{cases}$$

where $[\mathbf{ED}]_{ij}$ are the elements of the Euclidean-distance map matrix and D_C is a critical distance.

The **Euclidean-adjacency map matrix**, denoted as $\mathbf{EA}$, is the analogue of the $\rightarrow$ *bond length-weighted adjacency matrix* defined for molecular graphs. It is defined by replacing elements equal to one, corresponding to pairs of adjacent spots in the $\rightarrow$ *adjacency matrix* with the corresponding elements in the Euclidean-distance map matrix $\mathbf{ED}$ as [Randić and Basak, 2002; Randić, Lerš *et al.*, 2004b]

$$[\mathbf{EA}]_{ij} = \begin{cases} [\mathbf{ED}]_{ij} & \text{if } i, j \text{ are connected} \\ 0 & \text{if } i = j \text{ or } i, j \text{ are not connected} \end{cases}$$

The **map connectivity matrices** are another set of map matrices based on partitioning of the $\rightarrow$ *Randić connectivity index* and the higher order $\rightarrow$ *connectivity indices* into contributions arising from paths of length k [Randić and Basak, 2002]. They are defined as

$$\begin{aligned} [\mathbf{D}_{1\chi}]_{ij} &= (\delta_i \cdot \delta_j)^{-1/2} \cdot \delta(d_{ij}, 1) \\ [\mathbf{D}_{2\chi}]_{ij} &= (\delta_i \cdot \delta_l \cdot \delta_j)^{-1/2} \cdot \delta(d_{ij}, 2) \\ &\dots\dots\dots \\ [\mathbf{D}_{k\chi}]_{ij} &= (\delta_i \cdot \dots \cdot \delta_j)^{-1/2} \cdot \delta(d_{ij}, k) \end{aligned}$$

where δ_i indicates the $\rightarrow$ vertex degrees and $\delta(d_{ij}, k)$ indicates the Kronecker delta function that is equal to one for pairs of vertices v_i and v_j at a topological distance of k , and zero otherwise. The term $(\delta_i \cdot \dots \cdot \delta_j)$ indicates the product of the degrees of the vertices along the path connecting the vertices v_i and v_j .

Higher order map matrices are $\rightarrow$ power matrices derived by a map matrix $\mathbf{M}$ by either using the standard matrix multiplication of linear algebra or by using the $\rightarrow$ Hadamard matrix product, which leads to matrices whose elements are defined as $[\mathbf{M}]_{ij}^k$, where k is an integer exponent.

Map invariants usually calculated from map matrices are the $\rightarrow$ leading eigenvalue λ_1 of a map matrix and the normalized leading eigenvalues of its higher order matrices (e.g., $\lambda_1/k!$, k being the matrix order) [Bajzer, Randić *et al.*, 2003]. Other map invariants are the average row sum of a map matrix and the average of those row sums corresponding to protein spots lying in a selected region of the proteomics map [Randić, Lerš *et al.*, 2004b]. Moreover, $\rightarrow$ Wiener-type indices of map matrices were also investigated.

📖 [Marengo, Leardi *et al.*, 2003; Randić and Basak, 2004; Vračko and Basak, 2004; Randić, Lerš *et al.*, 2005b; Marengo, Robotti *et al.*, 2006; Randić, Witzmann *et al.*, 2006; Marengo, Robotti *et al.*, 2008]

➤ **bioisosterism** $\rightarrow$ drug design

■ biological activity indices

These are molecular properties related to the effect of a substance produces on an organism or any biological target. Biological activity depends on peculiarities of compounds (molecular structure and $\rightarrow$ physico-chemical properties), biological entity (species, gender, age, etc.), and mode of treatment (dose, route, exposure, etc.).

The *dose* is the amount of a substance that is administrated to an organism (animal, human) trough food or other administration routes (topic, gavage, injection, etc.).

The *dose–response curve* is a sigmoidal curve (Figure B10) that highlights the relation between the amount of a drug or chemical administered to an organism and the degree of response it produces. This response is measured by the percentage of the exposed population that shows the defined effect.

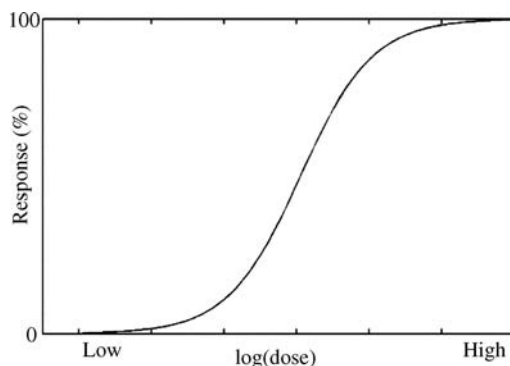


Figure B10 Dose–response curve.

Dose–response experiments typically use 10–20 doses, approximately spaced on a logarithmic scale. For example,

Dose (nM)	1	3	10	30	100	300	1000	3000	10000
Dose (log)	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4

• pharmacological indices

In pharmacology, the **Effective Dose** (ED) is the minimal dose that produces the desired effect of a drug. The effective dose is often determined based on analyzing the dose–response relationship specific to the drug. The dosage that produces a desired effect in half the test population is referred to as the **median effective dose** ED₅₀, that is, the amount of drug that produces a therapeutic response in 50% of the people taking it.

The **therapeutic index** (or **therapeutic ratio**) is a comparison of the amount of a therapeutic agent that causes the therapeutic effect to the amount that causes toxic effects. Quantitatively, it is the ratio of the dose required to produce the toxic effect over the therapeutic dose. A commonly used measure of therapeutic index is the lethal dose of a drug for 50% of the population (LD₅₀) divided by the effective dose for 50% of the population (ED₅₀):

$$\text{therapeutic index} = \frac{\text{LD}_{50}}{\text{ED}_{50}}$$

The therapeutic index of a drug indicates the selectivity of the drug and consequently its usability. It should be noted that a single drug can have many therapeutic indices; while some are for each of its undesirable effects relative to a desired drug action, the others for each of its desired effects if the drug has more than one action.

• toxicological indices

Toxicity is a relative property of a chemical that refers to its potential to have a harmful effect on a living organism. It is experimentally determined through toxicity tests in which organisms are exposed through food (*oral toxicity*) or are exposed at a concentration of the chemical in a given environmental compartment, such as water, air, or soil (*environmental toxicity*).

In acute oral tests, organisms are subject to a single dose of the chemical and the toxicity is not a function of exposure time. Several → *structural alerts* were proposed for identifying toxicological effects, particularly for carcinogenicity and mutagenicity of the chemicals.

The **lethal dose** (LD) is the dose of a chemical or biological preparation that is likely to cause death, giving an indication of the lethality of a given substance. Because resistance varies from one individual to another, the “lethal dose” represents a dose (usually recorded as weight of the dose per kilogram of subject body weight, e.g., mg/kg b.w.) at which a given percentage of subjects will die. The most commonly used lethality indicator is the **median lethal dose** LD₅₀, a dose at which 50% of subjects will die.

In long-term oral toxicity tests, the organism is fed for several days (in some cases months or years) with food contaminated by the tested chemical. In this case, results are expressed as concentration in food and are a function of exposure time. The total dose may also be calculated from the concentration and the total amount of food ingested.

In environmental toxicity tests, the **Lethal Concentration** (LC) is a measure (as weight/weight or weight/volume) of the concentration of the toxic chemical producing death in a given percentage of organisms. The concentration at which 50% of subjects will die is denoted as LC₅₀ and is called **median lethal concentration**. It is always a function of exposure time (for example, 96 h LC₅₀ in short-term toxicity tests on fish).

The **Lethal Time** (LT₅₀) is the time needed for 50% of the subjects to die after the exposure at a determined concentration of a substance.

Median Inhibitory Concentration (IC₅₀) is a measure of the concentration required for producing 50% inhibition of a biological activity (i.e., an enzyme reaction, cell growth, reproduction, etc.). In simpler terms, it measures how much of a particular substance/molecule is needed to inhibit some biological process by 50%. IC₅₀ is commonly used as a measure of drug-receptor binding affinity.

No-Observed-Effect Level (NOEL) is the greatest concentration or amount of a substance, found by experiment or observation, that causes no alterations of morphology, functional capacity, growth, development, or life span of target organisms distinguishable from those observed in normal (control) organisms of the same species under the same defined conditions of exposure.

The **No-Observed-Adverse-Effect Level** (NOAEL) is used for those chemicals that at low levels may be beneficial or necessary (e.g., natural micronutrients, such as some heavy metals).

The **Lowest-Observed-Effect-Level** (LOEL) is the lowest level to which a studied effect is observed.

The **Acceptable Daily Intake** (ADI) is the daily intake of a chemical that, during an entire lifetime, appears to be without appreciable risk. It is expressed as in milligrams of the chemical per kilogram of body weight (mg/kg b.w.). It is usually estimated as

$$\text{ADI} = \frac{\text{NOAEL}}{\text{SF}}$$

where NOAEL is the No-Observed-Adverse-Effect Level and SF is a safety factor. Safety factors are a function of the level of uncertainty of the NOAEL and the toxicological mode of action and may range from 10 to 10000.

The **Iball index** is defined as the percentage of skin cancer or papilloma-developing mice (skin painting experiments) divided by the average latent period in days for the affected animals multiplied by 100 [Daudel and Daudel, 1966; Herndon and Szentpály, 1986; Barone, Camilo Jr. *et al.*, 1996; Braga, Barone *et al.*, 1999; Barone, Braga *et al.*, 2000].

📖 [Wang and Milne, 1993; Benigni, 2003; Öberg, 2004a]

- **biological activity profile score** → scoring functions
- **Bird aromaticity indices** → delocalization degree indices
- **BLOGP** → lipophilicity descriptors

■ Blurock spectral descriptors

They are atomic or bond descriptors derived from a spectral representation of molecules, like the following: a property is associated with each atom (or bond) in such a way as to also represent the atom and its environment, the range of the property values in the whole data set of molecules

is divided into equal-sized intervals and the number of times the values of the molecule atoms fall within each interval is counted. The spectrum of the molecule is the distribution of these property values [Blurock, 1998].

The atomic properties considered are partial charges, electron densities, and polarizabilities, calculated by $\rightarrow$ *computational chemistry* methods; moreover, bond properties have been proposed as the difference between the property values of the atoms forming the bond. The range of each property is determined by the maximum and minimum values for all the atoms in all the molecules, thus obtaining uniform spectrum length for all the molecules in the data set.

Inductive learning was suggested for the prediction of the molecular property values.

- **Bocek–Kopecky analysis** $\rightarrow$ Free–Wilson analysis
- **Bocek–Kopecky model** $\rightarrow$ Free–Wilson analysis
- **Bodor hydrophobic model** $\equiv$ *BLOGP* $\rightarrow$ lipophilicity descriptors
- **Bodor LOGP** $\equiv$ *BLOGP* $\rightarrow$ lipophilicity descriptors
- **boiling point** $\rightarrow$ physico-chemical properties
- **Bonchev centric information indices** $\rightarrow$ centric indices
- **Bonchev complexity information index** $\rightarrow$ molecular complexity
- **Bonchev topological complexity indices** $\rightarrow$ molecular complexity
- **bond alternation coefficient** $\rightarrow$ delocalization degree indices
- **bond angles** $\rightarrow$ molecular geometry
- **bond connectivity index** $\equiv$ *edge connectivity index* $\rightarrow$ edge adjacency matrix
- **bond connectivity indices** $\equiv$ *extended edge connectivity indices* $\rightarrow$ edge adjacency matrix
- **bond count** $\equiv$ *bond number*
- **bond dipole moment** $\rightarrow$ bond ionicity indices
- **bond distances** $\rightarrow$ molecular geometry
- **bond distance-weighted edge adjacency matrix** $\rightarrow$ edge adjacency matrix
- **bond eccentricity** $\rightarrow$ edge distance matrix
- **bonded pair descriptors** $\rightarrow$ substructure descriptors
- **bond E-state index** $\rightarrow$ electrotopological state indices
- **bond flexibility** $\rightarrow$ flexibility indices
- **bond flexibility index** $\rightarrow$ flexibility indices
- **bond index** $\rightarrow$ quantum-chemical descriptors
- **bonding information content** $\rightarrow$ indices of neighborhood symmetry
- **bonding orbital information index** $\rightarrow$ information theoretic topological index

■ bond ionicity indices

Such indices encode information about the bond character, being the importance of bond character to the physical and chemical behavior of compounds well known. Bond character is closely related to the capacity of bonded atoms to exchange electrons and such capacity is commonly well represented by the $\rightarrow$ *electronegativity* χ of the bonded atoms.

The difference in electronegativity between two bonded atoms was called **bond dipole moment** [Malone, 1933], defined as

$$\mu_{ij} = |\chi_i - \chi_j|$$

but there was poor correlation between this index and bond ionicity. Therefore, starting from bond dipole moment, several empirical relationships have been proposed to define bond ionicity indices f_b . The most popular are [Barbe, 1983]:

1. $f_b = 1 - \exp \left[-0.25 \cdot (\chi_i - \chi_j)^2 \right]$
2. $f_b = 1 - \exp \left[-0.21 \cdot (\chi_i - \chi_j)^2 \right]$
3. $f_b = 0.160 \cdot (\chi_i - \chi_j) + 0.035 \cdot (\chi_i - \chi_j)^2$
4. $f_b = \frac{\chi_i - \chi_j}{\chi_i + \chi_j}$
5. $f_b = \frac{\chi_i - \chi_j}{2}$
6. $f_b = \frac{\chi_i - \chi_j}{\chi_i}$ with $\chi_i > \chi_j$

- **Bondi volume** → volume descriptors (⊙ van der Waals volume)
- **bond length-corrected connectivity index** → connectivity indices (⊙ Kupchik modified connectivity indices)
- **bond length-weighted adjacency matrix** → molecular geometry
- **bond length-weighted distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **bond length-weighted Wiener index** → weighted matrices (⊙ weighted distance matrices)
- **bond matrix** ≡ *edge adjacency matrix*

■ **bond number** (B) (≡ *edge counting; bond count*)

This is the simplest graph invariant defined as the number of edges in the simple → *molecular graph* G where multiple bonds are considered as single edges. The bond number is calculated from the → *adjacency matrix* A as half the → *total adjacency index* A_V :

$$B = \frac{1}{2} \cdot A_V = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A a_{ij}$$

where a_{ij} are the elements of the adjacency matrix and A is the total number of graph vertices.

The bond number is related to molecular size and gives equal weight to chemically nonequivalent groups, such as $\text{CH}_2\text{--CH}_2$, $\text{CH}_2\text{=CH}$, $\text{CH}_2\text{--NH}_2$, and $\text{CH}_2\text{--Cl}$.

When bond multiplicity in the molecule must be considered several → *multiple bond descriptors* can be used instead of the bond number.

The number of bonds is considered in the → *cyclomatic number* and appears in several → *molecular descriptors* such as the → *Balaban distance connectivity index*, the → *mean Randić branching index*, the → *information bond index*, and several → *topological information indices*.

- **bond order** → quantum-chemical descriptors
- **bond order–bond length relationships** → bond order indices

■ **bond order indices**

These are descriptors for molecule bonds proposed with the aim of estimating the → *bond order* defined in quantum-chemical theory or of generally defining bond weights so as to distinguish the bonds in a → *molecular graph*.

The most common definitions of bond order indices are reported below. Moreover, the term **fractional bond order** was suggested to refer to the inverse of any bond order index. Fractional bond order permits individual treatment of σ and π molecular systems; σ bonds give simple graphs, while π bonds introduce a weighted molecular framework with weights smaller than one [Randić, Brissey *et al.*, 1980].

• **conventional bond order (π^*)**

Within the framework of the graph theory, the conventional bond order π^* is defined as being equal to 1, 2, 3, and 1.5, for single, double, triple, and aromatic bonds, respectively. The $\rightarrow$ *bond vertex degree* of an atom is an important local invariant defined as the sum of the conventional bond orders of the edges incident to a vertex.

To consider chemical information relative to multiple bonds in terms of topological bond lengths, the inverse powers of the conventional bond order were proposed [Balaban, 1993c; Balaban, Bonchev *et al.*, 1993]. The **relative topological distance** is defined as

$$RTD_{ij} = (\pi_{ij}^*)^{-1}$$

where i and j refer to adjacent vertices in the graph. To obtain values more related to the standardized experimental interatomic average distances (as reference is taken the distance of single bonds, Table B11), the **chemical distance** was defined as

$$CD_{ij} = (\pi_{ij}^*)^{-1/4}$$

Using relative topological distance or chemical distance as well as conventional bond order to weight each edge in the graph several $\rightarrow$ *weighted matrices* were proposed which account for information about bond multiplicity.

Table B11 Average experimental bond length r , carbon relative value r^* , conventional bond order π^* , relative topological distance RTD , and chemical distance CD .

Bond type	r (Å)	r^*	π^*	RTD	CD
C–C	1.54	1.00	1	1.00	1.00
C $\approx$ C	1.40	0.91	1.5	0.67	0.90
C=C	1.33	0.86	2	0.50	0.84
C $\equiv$ C	1.20	0.78	3	0.33	0.76
C–N	1.47	1.00	1	1.00	1.00
C=N	1.29	0.88	2	0.50	0.84
C $\equiv$ N	1.16	0.79	3	0.33	0.76
N–N	1.45	1.00	1	1.00	1.00
N=N	1.26	0.87	2	0.50	0.90
C–O	1.41	1.00	1	1.00	1.00
C=O	1.21	0.86	2	0.50	0.84
O–O	1.45	1.00	1	1.00	1.00
N–O	1.47	1.00	1	1.00	1.00
N=O	1.15	0.78	2	0.50	0.84
C–S	1.81	1.00	1	1.00	1.00
C=S	1.61	0.89	2	0.50	0.84

The symbol $\approx$ stands for aromatic bonds.

From the conventional bond order, the **atomic multigraph factor** (or **multigraph factor**) is a $\rightarrow$ local vertex invariant, denoted as f_i , and defined as [Balaban and Diudea, 1993]

$$f_i = \sum_{j=1}^A a_{ij} \cdot (\pi_{ij}^* - 1) \quad \pi_{ij}^* = 0 \quad \text{if } (i, j) \notin E(G)$$

where the summation goes over all graph vertices, but the only nonvanishing elements are those corresponding to pairs of adjacent vertices (a_{ij} are the elements of the adjacency matrix); π_{ij}^* is the conventional bond order associated to the edge connecting vertices v_i and v_j for pairs of adjacent vertices, and zero otherwise. The multigraph factor is zero for atoms without multiple bonds and it is used, for instance, to derive local invariants from $\rightarrow$ layer matrices and the $\rightarrow$ Balaban DJ index. Moreover, the atomic multigraph factor is closely related to the $\rightarrow$ bond vertex degree δ_i^b as

$$\delta_i^b = \sum_{j=1}^A a_{ij} \cdot \pi_{ij}^* = \delta_i + f_i \quad \pi_{ij}^* = 0 \quad \text{if } (i, j) \notin E(G)$$

where δ is the simple $\rightarrow$ vertex degree, that is, the number of adjacent vertices.

• graphical bond order

The graphical bond order of the b th bond, denoted as $(TI'/TI)_b$, is derived from the $\rightarrow$ *H-depleted molecular graph* of the molecule by calculating a $\rightarrow$ graph invariant TI' for the subgraph G' obtained by erasing an edge b from the graph and then dividing it by the corresponding graph invariant TI calculated on the whole molecular graph G [Randić, Mihalić *et al.*, 1994]. If more than one subgraph is obtained by the erasure of each edge, the single contributions are summed up to give the graphical bond order or, alternatively, they can be multiplied [Mekenyan, Bonchev *et al.*, 1988a]. The ratio $(TI'/TI)_b$ was interpreted as a measure of the relative importance of the edge in the graph. The first proposed graphical bond order was that calculated from the $\rightarrow$ Hosoya *Z index* and was originally called **topological bond order**; it was shown to represent the weight of a bond in distributing π -electrons over the molecular graph [Hosoya, Hosoi *et al.*, 1975; Hosoya and Murakami, 1975]. Moreover, graphical bond orders can be considered special cases of $\rightarrow$ normalized fragment topological indices.

Molecular descriptors are derived by the additive contributions of the graphical bond orders of all bonds in the molecule as

$$TI'/TI = \sum_{b=1}^B \left(\frac{TI'}{TI} \right)_b$$

They are usually called **graphical bond order descriptors**.

The graphical bond order calculated using the $\rightarrow$ total path count P as molecular invariant was called **path graphical bond order** and denoted by the ratio $(P'/P)_{ij}$, where i and j refer to the vertices incident to the b th edge erased from the graph [Randić, 1991b; Randić and Trinajstić, 1993a; Plavšić, Šoškić *et al.*, 1996b].

From the path graphical bond orders, a square symmetric $\rightarrow$ weighted adjacency matrix of dimension $A \times A$, called **P-matrix** (or **path matrix**) and denoted as **P**, was derived [Plavšić and Graovac, 2001]; its elements are defined by the following:

$$[P]_{ij} = \begin{cases} (P'/P)_{ij} & (i, j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

where $E(G)$ is the set of graph edges.

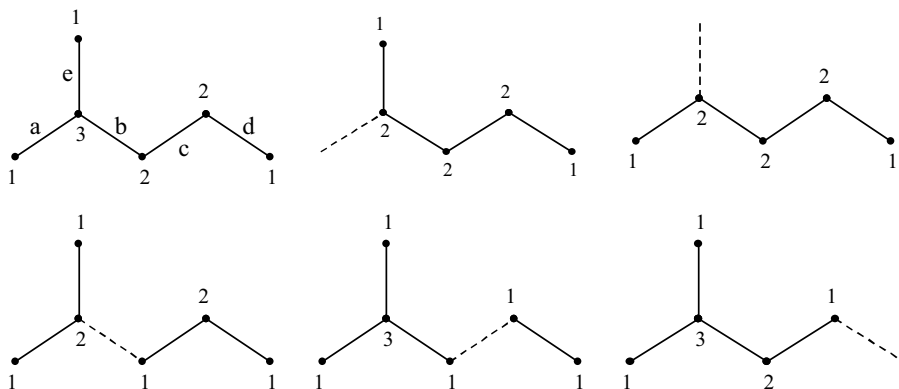
From the P-matrix, a $\rightarrow$ Wiener-type index, called **P'/P index**, is calculated applying the $\rightarrow$ Wiener operator Wi as

$$\frac{P'}{P} \equiv Wi(P) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [P]_{ij}$$

Other encountered graphical bond order descriptors are **χ'/χ index**, **W'/W index**, **WW'/WW index**, **J'/J index**, **CID'/CID index**, and $\rightarrow$ **Z'/Z index** derived, respectively, from $\rightarrow$ **Randić connectivity index**, $\rightarrow$ **Wiener index**, $\rightarrow$ **hyper-Wiener index**, $\rightarrow$ **Balaban distance connectivity index**, $\rightarrow$ **Randić connectivity ID number**, and $\rightarrow$ **Hosoya Z index**.

Example B2

χ'/χ graphical bond order index for 2-methylpentane.



$${}^1\chi(G) = 2(1 \cdot 3)^{-1/2} + (3 \cdot 2)^{-1/2} + (2 \cdot 2)^{-1/2} + (2 \cdot 1)^{-1/2} = 2.7701$$

$${}^1\chi(G-a) = {}^1\chi(G-e) = 2 \cdot (2 \cdot 1)^{-1/2} + 2 \cdot (2 \cdot 2)^{-1/2} = 2.4142$$

$$\left(\frac{\chi'}{\chi}\right)_a = \left(\frac{\chi'}{\chi}\right)_e = \frac{{}^1\chi(G-a)}{{}^1\chi(G)} = \frac{2.4142}{2.7701} = 0.8715$$

$${}^1\chi(G-b) = 4 \cdot (2 \cdot 1)^{-1/2} = 2.8284$$

$$\left(\frac{\chi'}{\chi}\right)_b = \frac{{}^1\chi(G-b)}{{}^1\chi(G)} = \frac{2.8284}{2.7701} = 1.0210$$

$${}^1\chi(G-c) = (1 \cdot 1)^{-1/2} + 3 \cdot (1 \cdot 3)^{-1/2} = 2.7321$$

$$\left(\frac{\chi'}{\chi}\right)_c = \frac{{}^1\chi(G-c)}{{}^1\chi(G)} = \frac{2.7321}{2.7701} = 0.9863$$

$${}^1\chi(G-d) = 2 \cdot (1 \cdot 3)^{-1/2} + (1 \cdot 2)^{-1/2} + (2 \cdot 3)^{-1/2} = 2.2701$$

$$\left(\frac{\chi'}{\chi}\right)_d = \frac{{}^1\chi(G-d)}{{}^1\chi(G)} = \frac{2.2701}{2.7701} = 0.8195$$

$$\frac{\chi'}{\chi} = \left(\frac{\chi'}{\chi}\right)_a + \left(\frac{\chi'}{\chi}\right)_e + \left(\frac{\chi'}{\chi}\right)_b + \left(\frac{\chi'}{\chi}\right)_c + \left(\frac{\chi'}{\chi}\right)_d = 2 \cdot 0.8715 + 1.0210 + 0.9863 + 0.8195 = 4.5698$$

An explicit formula for the direct calculation of the χ'/χ index from the molecular graph was derived as

$$\frac{\chi'}{\chi} = \frac{1}{\chi(G)} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot \frac{(A+C-\delta_i-\delta_j) \cdot (\delta_i \cdot \delta_j)^{1/2} + \delta_i \cdot [(\delta_i-1) \cdot \delta_j]^{1/2} + \delta_j \cdot [(\delta_j-1) \cdot \delta_i]^{1/2}}{\delta_i \cdot \delta_j}$$

where $\chi(G)$ is the $\rightarrow$ *Randić connectivity index* for the whole molecular graph, the summation goes over all pairs of vertices, but the only nonvanishing terms are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the adjacency matrix; δ_i and δ_j are the $\rightarrow$ *vertex degrees* of the vertices v_i and v_j , A the number of graph vertices, and C the $\rightarrow$ *cyclomatic number*. This formula holds for every connected graph with $A > 1$ vertices [Plavšić, Šoškić *et al.*, 1998].

In the same way, a general formula valid for any graph based on the number A of graph vertices and the distances between pairs of vertices was derived for the calculation of the WW'/WW index as [Plavšić, 1999]

$$\frac{WW'}{WW} = A + 1 - \frac{\sum_{i=1}^{A-1} \sum_{j=i+1}^A d_{ij} \cdot (d_{ij} + 1) \cdot (d_{ij} + 2)}{\sum_{i=1}^{A-1} \sum_{j=i+1}^A d_{ij} \cdot (d_{ij} + 1)},$$

where d_{ij} is the topological distance between vertices v_i and v_j .

Only for acyclic graphs, after special rearrangement, does the formula take the form

$$\frac{WW'}{WW} = A + 1 - \frac{\sum_{m=1}^D \frac{(m+2)!}{(m-1)!} \cdot {}^m P}{\sum_{m=1}^D \frac{(m+1)!}{(m-1)!} \cdot {}^m P}$$

where m is the length of the considered paths, ${}^m P$ the $\rightarrow$ *path count* of m th order, and D the $\rightarrow$ *topological diameter*, that is, the maximum topological distance in the graph.

• bond order–bond length relationships

Several bond order–bond length relationships were proposed in literature [Paolini, 1990; Alkorta, Rozas *et al.*, 1998]. The most known relationships are collected in Table B12.

Table B12 Relationships between bond length and bond order.

Equation	$r = f(\pi)$	$\pi = f(r)$	Reference
Pauling (1947)	$r_{ij} = \hat{r}_{ij} - 0.71 \cdot \log(\pi_{ij})$	$\pi_{ij} = \exp \left[-\frac{(r_{ij} - \hat{r}_{ij})}{0.71} \right]$	[Pauling, 1947]
Pauling (1986)	$r_{ij} = \hat{r}_{ij} - 0.700 \cdot \log \{ \pi_{ij} \cdot [1 + 0.064 \cdot (\nu - 1)] \}$	$\pi_{ij} = \frac{\exp \left[-\frac{(r_{ij} - \hat{r}_{ij})}{0.700} \right]}{1 + 0.064(\nu - 1)}$	[Pauling and Kamb, 1986]
Paolini (1990)	$r_{ij} = \hat{r}_{ij} - 0.78 \cdot (\pi_{ij}^{1/3} - 1)$	$\pi_{ij} = \left[1 - \frac{(r_{ij} - \hat{r}_{ij})}{0.78} \right]^3$	[Paolini, 1990]
Gordy (1947)	$r_{ij} = \sqrt{a/(b + \pi_{ij})}$	$\pi_{ij} = a \cdot r_{ij}^{-2} - b$	[Gordy, 1947]
Lendvay (2000)	$r_{ij} = \hat{r}_{ij} - 0.25 \cdot \ln \pi_{ij}$	$\pi_{ij} = \exp \left[-\frac{(r_{ij} - \hat{r}_{ij})}{0.25} \right]$	[Lendvay, 2000]

$\hat{r}$ is the equilibrium bond length of a single bond. The parameters a and b of the Gordy equation are given in Table B13. For Pauling (1947) formula, some equilibrium distances $\hat{r}$ are C–C = 1.542 Å, C=C = 1.330 Å, and C≡C = 1.204 Å; for Lendvay formula, C–C = 1.54 Å, C=O = 1.43 Å, and C–H = 1.08 Å.

Gordy's bond order is used in the calculation of the $\rightarrow$ *Bird aromaticity indices* and the empirical constants a and b are given in Table B13 [Gordy, 1947; Krygowski and Cyranski, 2001].

Table B13 Values of a and b constants used in the calculation of Gordy's bond order.

Bond	a	b	Bond	a	b
C≈C	6.80	1.71	N≈N	5.28	1.41
C≈N	6.48	2.00	N≈O	4.98	1.45
C≈O	5.75	1.85	N≈S	10.53	2.50
C≈S	11.9	2.59	O≈O	4.73	1.22
C≈P	13.54	3.02	O≈S	17.05	5.58
B≈B	9.12	1.94	S≈S	19.30	3.46
B≈C	8.05	2.11	C≈Se	15.24	3.09
B≈N	7.15	2.10	C≈Te	21.41	3.81
B≈O	6.75	2.14	N≈Se	13.31	2.86

The symbol $\approx$ stands for aromatic bonds.

📖 [Pauling, Brockway *et al.*, 1935; Bernstein, 1947; Gutman, Bosanac *et al.*, 1978; Randić, 1991g; Randić, 1993c; Hansen and Zheng, 1994; Randić, 1994b; Oláh, Blockhuys *et al.*, 2006; Sedlar, Andelic *et al.*, 2006]

- **bond order-weighted edge adjacency matrix** $\rightarrow$ edge adjacency matrix
- **bond order-weighted edge connectivity index** $\rightarrow$ edge adjacency matrix
- **bond order-weighted vertex connectivity indices** $\rightarrow$ connectivity indices
- **bond order-weighted Wiener index** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **bond profiles** $\rightarrow$ molecular profiles
- **bond rigidity** $\rightarrow$ flexibility indices
- **bond rigidity index** $\rightarrow$ flexibility indices ($\odot$ bond flexibility index)

- **bond spectral moments** → edge adjacency matrix
- **bond type *E*-state indices** → electrotopological state indices
- **bond vertex degree** → vertex degree
- **bootstrap** → validation techniques
- **Bowden–Wooldridge steric constant** → steric descriptors (⊙ number of atoms in substituent specific positions)
- **Bowden–Young steric constant** → steric descriptors (⊙ Charton steric constant)
- **branching ETA index** → ETA indices
- **branching index** $\equiv$ *Randić connectivity index* → connectivity indices
- **branching indices** → molecular complexity (⊙ molecular branching)
- **branching layer matrix** → layer matrices
- **Braun–Blanque similarity coefficient** → similarity/diversity (⊙ Table S9)
- **Bray–Curtis distance** $\equiv$ *Lance–Williams distance* → similarity/diversity (⊙ Table S7)
- **Brillouin redundancy index** → information content
- **Broto–Moreau–Vandicke hydrophobic atomic constants** → lipophilicity descriptors
- **Buckingham potential function** → molecular interaction fields (⊙ steric interaction fields)
- **bulk descriptors** → steric descriptors
- **bulkiness of an atom** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **bulk representation** → molecular descriptors
- **Burden eigenvalues** → spectral indices
- **Burden matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **Burden modified eigenvalues** → spectral indices (⊙ Burden eigenvalues)
- **Buser distance** $\equiv$ *Baroni–Urbani distance* → similarity/diversity (⊙ Table S7)

C

- **CACTVS screen vectors** → substructure descriptors (⊙ structural keys)
- **Calculated LOGP** $\equiv$ *CLOGP* → lipophilicity descriptors (⊙ Leo–Hansch hydrophobic fragmental constants)
- **Camilleri model based on surface area** → lipophilicity descriptors
- **Cammarata–Yau analysis** → Free–Wilson analysis
- **Cammarata–Yau model** → Free–Wilson analysis
- **Canberra distance** → similarity/diversity (⊙ Table S7)

■ **canonical numbering** ($\equiv$ *unique atomic ordering; unique atomic code*)

This is a procedure that assigns unique labels to the graph vertices so that the resulting matrix representations are in canonical form. The principal aim is to find a suitable numerical code for each given graph, which characterizes the graph up to isomorphism [Kvasnička and Pospichal, 1990; Faulon, 1998; Ivanciuc, 2003b].

The main canonical ordering procedures are listed below. Several different → *local vertex invariants* showing regular variation from central to terminal vertices were studied for canonical numbering of graph vertices [Filip, Balaban *et al.*, 1987; Bonchev and Kier, 1992]; examples are → *vertex distance degree*, → *local connectivity indices*, → *electrotopological state indices*, → *weighted atomic self-returning walk counts*, → *Randić atomic path code*, → *MPR descriptors*, → *centric operator*, → *centrocomplexity operator*, → *vertex complexity*, and → *vertex distance complexity*. The → *iterative vertex and edge centrality algorithm* (IVEC) also provides canonical ordering of vertices and edges in the graph and an algorithm based on the eigenvalues and eigenvectors of the adjacency matrix of the graph was also proposed [Liu and Klein, 1991].

• **Morgan's extended connectivity algorithm** ($\equiv$ *extended connectivity algorithm, ECA*)

Graph vertices are ordered on the basis of their extended connectivity values obtained after a number of iterations of the Morgan method until constant atom ordering is obtained in two consecutive steps [Morgan, 1965]. The **extended connectivity** (or **extended vertex degree**), denoted as EC_i , of a vertex is calculated as the iterative summation of connectivities of all first neighbors as the following:

$$EC_i^{k+1} = \sum_{j=1}^A a_{ij} \cdot EC_j^k$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix*, being equal to one only for pairs of adjacent vertices and zero otherwise; at the beginning ($k = 0$) the connectivity of each atom is simply the $\rightarrow$ *vertex degree* δ .

It must be pointed out that the extended connectivity EC^k of Morgan coincides with the $\rightarrow$ *atomic walk count* ($awc^{(k)}$) calculated as row sum of the k th power of the adjacency matrix **A** [Razinger, 1982; Rücker and Rücker, 1993; Figueras, 1993]. Then, the **extended connectivity indices**, denoted by EC^k and defined as [Rücker and Rücker, 1993]

$$EC^k = \sum_{i=1}^A EC_i^k$$

where A is the number of graph vertices, coincide with the $\rightarrow$ *molecular walk counts*.

The **normalized extended connectivity** (NEC_i) is derived as

$$NEC_i = \lim_{k \gg 1} \left(\frac{EC_i^k}{EC^k} \right) \cdot \sum_{i=1}^A EC_i^1$$

where the last summation coincides with twice the number of bonds in the molecular graph [Bonchev, Kier *et al.*, 1993].

The Morgan algorithm was later improved by a better formalization and considering stereochemical aspects [Wipke and Dyott, 1974a, 1974b]. The Stereochemically Extended Morgan Algorithm (SEMA) resulted in a higher discriminating ability of graph vertices than ECA [Wipke, Krishnan *et al.*, 1978]; it is based on the iterative summation of the properties of neighboring atoms.

📖 [Ouyang, Yuan *et al.*, 1999]

- **first eigenvector algorithm** (FEVA)

Vertices in a graph are ordered according to the relative magnitudes of the coefficients of the first eigenvector (corresponding to the largest eigenvalue) of the $\rightarrow$ *adjacency matrix* **A** [Randić, 1975d]. Generally nonequivalent vertices have different coefficient magnitudes, while equivalent vertices, that is, vertices constituting same orbits, have to be distinguished according to some alternative rules.

To obtain a vertex numbering similar to the Morgan algorithm, the convention to associate label 1 to the vertex with the largest coefficient and label A (the total number of vertices in the graph) to that with the smallest coefficient was established.

The largest coefficients correspond to $\rightarrow$ *central vertices*, the smallest to $\rightarrow$ *terminal vertices* and their neighbors.

- **smallest binary label** (SBL)

This is a binary label assigned to each graph vertex that consists of the corresponding row of the $\rightarrow$ *adjacency matrix*; this binary label can also be expressed as decimal number, as in the $\rightarrow$ *decimal adjacency vector*. The unique numbering given by the smallest binary label of each vertex can be achieved by iteratively renumbering the vertices of the graph, that is, iterative reordering of the adjacency matrix rows [Randić, 1974, 1975c; Mackay, 1975].

• Jochum–Gasteiger canonical numbering

A canonical numbering algorithm where nonterminal atoms are treated first, monovalent atoms (hydrogen and nonhydrogen atoms) then being numbered correspondingly to the nonterminal atoms [Jochum and Gasteiger, 1977]. The algorithm is based on the following steps:

1. the nonterminal atoms are put into the same equivalence class on the basis of their $\rightarrow$ *vertex eccentricity*; the classes are ordered according to increasing eccentricity values and the atoms within each equivalent class are then ordered separately by the following sequential rules, beginning with the first equivalence class;
2. for each equivalent class, the atom with the highest atomic number has priority;
3. the atom with the most free electrons has priority;
4. the atom with the highest number of first neighbors (i.e., highest vertex degree) has priority;
5. the atom that has a first-neighbor atom with an atomic number higher than the others has priority;
6. the atom that has a first-neighbor atom with more free electrons than the others has priority;
7. the atom that has more bonds to first-neighbor atoms than the others has priority;
8. the atom with the highest bond order to the heavier first-neighbor atom has priority.
9. the atom that lies closer to an atom already numbered has priority;
10. the atom that has a higher bond order to an atom already numbered has priority.

Finally, terminal atoms are then numbered according to rules 2 and 9.

• hierarchically ordered extended connectivities algorithms ($\equiv$ HOC algorithms)

HOC algorithm is an iterative procedure that finds topological equivalence classes (i.e., graph orbits) and provides canonical numbering of vertices in molecular graphs. It is based on the $\rightarrow$ *extended connectivity* like Morgan's algorithm but also on the hierarchical ordering at each stage provided by the rank of the previous iteration [Balaban, Mekenyan *et al.*, 1985a].

The whole procedure consists of some algorithms that allow handling using graphs of different levels of complexity.

The main algorithm for ordering the vertices in graph orbits is called HOC-1 and the steps are

Step 1. Vertices of the $\rightarrow$ *H-depleted molecular graph* are partitioned into equivalence classes according to their $\rightarrow$ *vertex degree* δ_i , that is, a first rank ${}^1K_i = \delta_i$ is assigned to each vertex.

Step 2. For each vertex, the first ranks of its adjacent vertices are listed in increasing order as

$${}^1K_i^1, {}^1K_i^2, \dots, {}^1K_i^{\delta_i}$$

Step 3. An additional discrimination within each class is performed by means of the extended connectivities EC, which are the sums of the vertex degrees (ranks) of the nearest

neighbors, as

$${}^1\text{EC}_i = \sum_{r=1}^{\delta_i} {}^1K_i^r$$

where the sum runs over the neighbor ranks of the i th vertex considered in increasing order. A second rank ${}^1K_i$ is assigned to each vertex according to the increasing order of the extended connectivities. When two or more vertices are of the same rank, that is, ${}^1K_i = {}^1K_j$, but the individual values of the terms contributing to the extended connectivity are different, that is, ${}^1K_i^r \neq {}^1K_j^r$, then the ordering is made according to the rank of the first different addendum.

Step 4. Steps 2 and 3 are iteratively repeated, replacing first ranks by second ranks until all the ${}^{k+1}K_i$ ranks become equal to ${}^kK_i$ ranks of the preceding stage for all vertices.

The HOC-3 algorithm is used for the canonical numbering of graph vertices and is equal to the HOC-1 algorithm from step 1 to step 4 with an additional step based on an artificial discrimination into the largest orbits including two or more vertices. To make such a discrimination, one of the vertices inside the largest orbit, with the highest cardinality ${}^kK_p$, is arbitrarily assigned a higher rank ${}^kK_i + 1$, and steps 2–4 are iteratively repeated.

The final vertex numbering is the inverse of the final HOC ranks so as to assign the lowest numbers to the most central vertices.

The HOC-2 and HOC-2A algorithms were proposed for the vertex ordering of special molecular graphs with pericondensed rings.

Based on HOC ranks, a *Unique Topological Representation* (UTR) was proposed in which topological equivalent vertices of the same rank in the graph are placed at the same level.

Moreover, two **HOC rank descriptors** based on the extended connectivity of graph vertices and the vertex rank obtained by HOC algorithms were proposed [Mekenyan, Bonchev *et al.*, 1984b]:

$$M = \sum_{i=1}^A M_i \quad \text{and} \quad N = \sum_{i=1}^A M_i^2$$

where the summation runs over all atoms and the term M_i is a local invariant defined as

$$M_i = \sum_{j=1}^i S_j \quad \text{and} \quad S_j = \sum_m K_m$$

where S_j is the sum of the ranks K of the adjacent vertices to the j th vertex restricted to those adjacent vertices of rank greater than j . The S_j values calculated for each vertex are ordered according to the increasing j index and are then summed up to i index to give the M_i local invariant. By this procedure, M_i gives a nondecreasing sequence.

Both descriptors increase with the size and cyclicity of the graph.

📖 [Mekenyan, Bonchev *et al.*, 1984a, 1985; Balaban, Mekenyan *et al.*, 1985b; Bonchev, Mekenyan *et al.*, 1985; Mekenyan, Balaban *et al.*, 1985; Ralev, Karabunarliev *et al.*, 1985]

• matrix method for canonical ordering

Graph vertices are partitioned and ordered into topological equivalence classes, that is, orbits, according to some special matrices developed for each atom [Bersohn, 1987]. These matrices give a representation of the whole molecule as seen from the considered atom.

The procedure consists first in assigning to each atom an atomic property P_i defined as

$$P_i = 1024 \cdot Z_i + 64 \cdot N_i^{\text{uns}} + 16 \cdot (4 - h_i)$$

where Z_i , N_i^{uns} , and h_i are the atomic number, the number of unsaturation, and the number of attached hydrogen atoms of the i th atom, respectively. The unsaturation number is 8 for an atom involved in a triple bond, 4 for either of the atoms in a double bond involving a heteroatom, 6 for an allene or ketene central atom with two double bonds, 2 for vinyl carbon atoms, 1 for aromatic atoms in a six-membered ring, and finally 0 for saturated atoms. The coefficients 1024, 64, and 16 have been chosen so that two atoms with the same atomic number cannot have the same property values.

The matrix representing the environment of the i th atom contains in the m th row the property values P_{mj} of the atoms located at a distance equal to m from the i th atom. The first row collects the property values of the first neighbors of the considered atom. The P_{mj} values are listed in descending order in the first entries of the matrix; the other entries are set to zero but are of no significance. The matrix dimension can be chosen for convenience.

The first partition of the vertices depends on their property values P_i ; two atoms X and Y are considered topologically equivalent if: (i) they have identical matrices and (ii) if X has a neighbor I belonging to a different equivalence class, then there must exist a neighbor J of the atom Y in the same equivalence class as the atom I .

Finally, the canonical ordering is performed on the basis of P_i values and in the case of equivalence according to the values of the matrix. Atoms with the greatest values are assigned the smallest numerical labels in the canonical ordering, such atoms being closest to the graph center.

Property values can also be modified to take into account geometric isomerism and chirality.

• self-returning walk ordering

Graph vertices are ordered on the basis of their $\rightarrow$ *self-returning walk counts*, that is, the number of walks of a given length starting and ending at the same vertex. In particular, local invariants representing the relative occurrence of the self-returning walks of the considered atom to all self-returning walks (SRWs) in the molecule, that is, $\rightarrow$ *topological atomic charge* TAC_i , provide a canonical numbering of graph vertices [Bonchev, Kier *et al.*, 1993]. The most important factors influencing the ordering, that is, the number of SRWs, are vertex branching, centrality, and cyclicity.

📖 [Balaban, 1976c; Randić, 1977c, 1978a, 1980b, 1995b; Hall and Kier, 1977a; Randić, Brissey *et al.*, 1979, 1981; Wilkins and Randić, 1980; Bonchev and Balaban, 1981; Bonchev, Mekenyan *et al.*, 1986; Polanski and Bonchev, 1986a; Klopman and Raychaudhury, 1988; Herndon, 1988; Diudea, Horvath *et al.*, 1992; Balaban, Filip *et al.*, 1992; Balasubramanian, 1995a; Babic, Balaban *et al.*, 1995; Laidboeur, Cabrol-Bass *et al.*, 1997; Agarwal, 1998; Lukovits, 1999]

■ Cao–Yuan indices

A set of three topological indices defined for a $\rightarrow$ *H-depleted molecular graph* in terms of the $\rightarrow$ *distance matrix* $\mathbf{D}$, the squared $\rightarrow$ *Harary matrix* $\mathbf{D}^{-2}$, the vector $\boldsymbol{\delta}$ of $\rightarrow$ *vertex degrees*, and the vector $\mathbf{p}$ of ring degrees [Cao and Yuan, 2001; Yuan and Cao, 2003].

The **ring degree** of a vertex v_i belonging to a cycle, denoted as ρ_i , is a local vertex invariant defined as the minimum number of nonhydrogen vertices bonded to vertex v_i , which must be removed to transform the i th vertex into an acyclic one [Cao and Yuan, 2001].

The definitions of the three Cao-Yuan descriptors are the following:

- **odd-even index (OEI)**

The odd-even index is derived from the **odd-even matrix OEM**, which is a binary matrix, defined as

$$[\mathbf{OEM}]_{ij} = \begin{cases} (-1)^{d_{ij}-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where the off-diagonal elements of the **OEM** matrix are ± 1 , depending on whether the topological distance d_{ij} is odd or even. The **OEI** matrix is then calculated by the Hadamard product of the **OEM** and $\mathbf{D}^{-2}$ matrices as

$$\mathbf{OEI} = \mathbf{OEM} \otimes \mathbf{D}^{-2}$$

where $\mathbf{D}^{-2}$ is the matrix whose elements are the reciprocal of the square topological distances. Finally, the odd-even index is calculated as

$$\text{OEI} = \sum_{i=1}^A \sum_{j \neq i}^A [\mathbf{OEI}]_{ij} = \sum_{i=1}^A \sum_{j \neq i}^A [[\mathbf{OEM}]_{ij} \cdot [\mathbf{D}^{-2}]_{ij}]$$

This index encodes information on interactions between vertices v_i and v_j , which are proportional to the inverse of their square distance.

- **vertex degree-distance index (VDI)**

It is defined as

$$\text{VDI} = \left(\prod_{i=1}^A f_i \right)^{1/A}$$

where f_i are the elements of the A -dimensional vector $\mathbf{f}$ defined as the product between the $\mathbf{D}^{-2}$ matrix and the vertex degree vector $\mathbf{v}$:

$$\mathbf{f} = \mathbf{D}^{-2} \cdot \mathbf{v}$$

In this way, the interactions between vertices v_i and v_j are determined not only by their distance, but also by their vertex degrees.

- **ring degree-distance index (RDI)**

To distinguish the different freedom of atoms belonging and not belonging to cycles, the RDI index is defined as

$$\text{RDI} = \left(\prod_{i=1}^A g_i \right)^{1/A}$$

where g_i are the elements of the A -dimensional vector $\mathbf{g}$ defined as the product between the matrix $\mathbf{D}^{-2}$ and the ring degree vector $\mathbf{p}$:

$$\mathbf{g} = \mathbf{D}^{-2} \cdot \mathbf{p}$$

As the ring degree is zero for vertices not in cycles, the RDI index is obviously zero for acyclic molecules.

• **edge degree-distance index (EDI)**

It is defined as [Yuan and Cao, 2003]

$$\text{EDI} = \left(\prod_{i=1}^A \text{ES}_i \right)^{1/A}$$

where ES_i are the elements of the A -dimensional vector $\mathbf{ES}$ defined as the product between the matrix $\mathbf{D}^{-2}$ and the vector of the $\rightarrow$ bond vertex degree δ^b that, unlike the vertex degree, accounts for bond multiplicity:

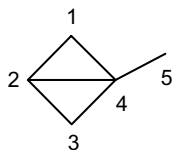
$$\mathbf{ES} = \mathbf{D}^{-2} \cdot \delta^b$$

This descriptor differs from VDI only for molecules containing multiple and/or aromatic bonds.

Note. The Authors use the name “edge degree” to refer to the bond vertex degree, but this is not correct because the edge degree was defined some years before [Bonchev, 1983] as the number of edges incident to an edge and not to a vertex.

Example C1

Cao-Yuan indices for 1-methylbicyclo[1.1.0]butane. $\mathbf{v}$ is the vector of vertex degrees, $\mathbf{p}$ the vector of ring degrees, $\mathbf{f}$ the vector obtained by the product of matrix $\mathbf{D}^{-2}$ and vector $\mathbf{v}$, $\mathbf{g}$ the vector obtained by the product of matrix $\mathbf{D}^{-2}$ and vector $\mathbf{p}$.


 $\mathbf{D} =$

Atom	1	2	3	4	5
1	0	1	2	1	2
2	1	0	1	1	2
3	2	1	0	1	2
4	1	1	1	0	1
5	2	2	2	1	0

 $\mathbf{D}^{-2} =$

Atom	1	2	3	4	5
1	0	1	0.25	1	0.25
2	1	0	1	1	0.25
3	0.25	1	0	1	0.25
4	1	1	1	0	1
5	0.25	0.25	0.25	1	0

 $\mathbf{OEI} =$

Atom	1	2	3	4	5
1	0	1	-0.25	1	-0.25
2	1	0	1	1	-0.25
3	-0.25	1	0	1	-0.25
4	1	1	1	0	1
5	-0.25	-0.25	-0.25	1	0

$\mathbf{v} = \{2, 3, 2, 4, 1\}$, $\mathbf{p} = \{1, 2, 1, 2, 0\}$, $\mathbf{f} = \{7.75, 8.25, 7.75, 8.00, 5.75\}$, $\mathbf{g} = \{4.25, 4.00, 4.25, 4.00, 3.00\}$

$\text{OEI} = 1 \times 12 - 0.25 \times 8 = 10.00$, $\text{VDI} = \text{EDI} = (7.75 \times 8.25 \times 7.75 \times 8.00 \times 5.75)^{1/5} = 7.44$
 $\text{RDI} = (4.25 \times 4.00 \times 4.25 \times 4.00 \times 3.00)^{1/5} = 3.87$

Boiling points of hydrocarbons were modeled by using Cao–Yuan indices, together with the fundamental contribution of the number of carbon atoms, as $N_C^{2/3}$.

- **capacity factor** → chromatographic descriptors
- **capacity factors** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **Carbó similarity index** → quantum similarity
- **cardinality layer matrix** → layer matrices
- **cardinality of a set** → algebraic operators
- **Carter resonance energy** → delocalization degree indices
- **Cartesian coordinates** → molecular geometry
- **CASE approach** → lipophilicity descriptors (⊙ Klopman hydrophobic models)
- **CAST** → molecular descriptors
- **CATS descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **CATS2D descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **CATS3D descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **CATS-charge descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **cavity term** → Linear Solvation Energy Relationships
- **cell-based density** → cell-based methods
- **cell-based entropy** → cell-based methods

■ **cell-based methods** ($\equiv$ *partition-based methods*)

Cell-based methods, as well as clustering or distance-based methods, aim at extracting representative structurally diverse subsets of compounds from large chemical databases [Cummins, Andrews *et al.*, 1996; Mason and Pickett, 1997; Pearlman and Smith, 1999; Farnum, DesJarlais *et al.*, 2003]. They are mainly used in design and optimization of combinatorial libraries; the most important aspect being here to ensure maximum diversity within and between libraries before they are produced. Moreover, cell-based methods are used for lead discovery purposes allowing the selection of the compounds most similar to the active reference target.

Cell-based methods represent compounds in a p -dimensional space where each dimension represents either a molecular descriptor or a linear combination of molecular descriptors. Moreover, these methods partition the chemical space into hyper-rectangular regions, that is, the cells, in which the compounds are placed according to their property values, and measure the occupancy of the resulting cells by means of several cell-based diversity measures. The chemical space is defined by a number of selected molecular descriptors representing properties of compounds and ranges of the values of these descriptors are used to define the cells. Molecular descriptors selected to span the chemical space are commonly molecular properties that would be expected to affect ligand–receptor binding. The range of values of each molecular descriptor is divided into a set of subranges, that is, the bins. This can be accomplished by the use of different binning schemes. The binning scheme is the algorithm that is used to partition the descriptor value range into appropriate bins.

If n_j is the number of bins of the j th molecular descriptor, then the total number of cells in which the chemical space is partitioned is obtained from

$$N_{\text{TOT}} = \prod_{j=1}^p n_j$$

where p is the number of descriptors defining the chemical space. This number increases fast with the number of descriptors; for example, the number of cells for 3 and 5 descriptors, each divided into 10 bins, is 10^3 and 10^5 , respectively.

There are different binning schemes; they usually satisfy two simple criteria [Bayley and Willett, 1999]: (1) the maximum and minimum values for each of the descriptors that specify the partition must be set so as to encompass all of the compounds that may need to be processed by the partitioning scheme and (2) it seems appropriate that each molecule be assigned to just a single cell, thus requiring that the cell ranges do not overlap at all.

There are two basic ways in which partition can be generated. The simpler, *descriptor-independent partitioning scheme*, assumes that the bin boundaries used to subdivide the j th descriptor are completely independent of the bin boundaries that have been used to subdivide the preceding $j - 1$ descriptors. Alternatively, a *descriptor-dependent partitioning scheme* generates bin boundaries to subdivide the j th descriptor by taking account of the bin boundaries that have been used to subdivide the preceding $j - 1$ descriptors. Moreover, two simple criteria can be used for controlling the bin width (and hence the occupancy of each bin): each descriptor can be divided into equally sized bins or into equally occupied bins. It is hence possible to identify four types of binning schemes (Figure C1), depending on whether the bin boundaries are, or are not, independent of the preceding bin boundaries and on the occupancy criterion that is used to define each of the bins.

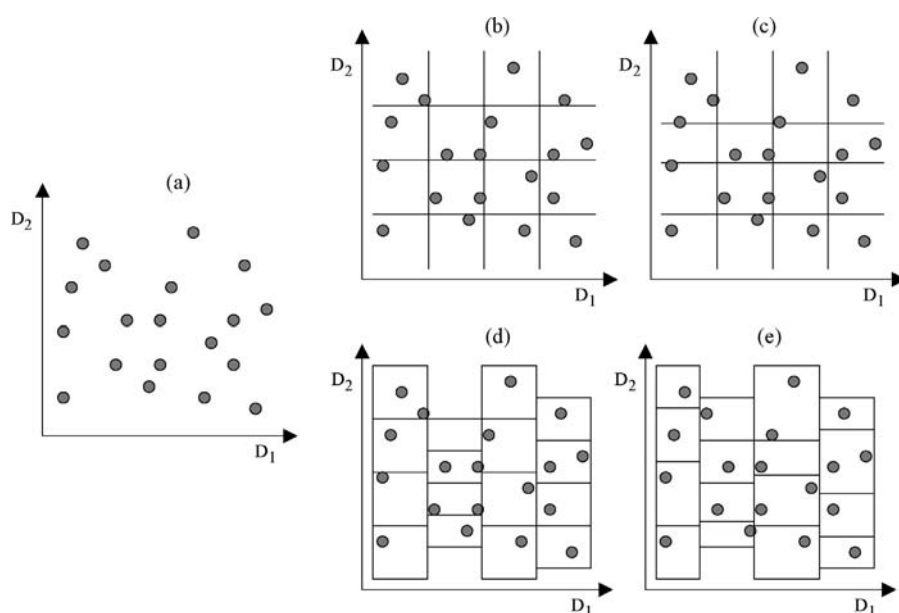


Figure C1 Example of the four binning schemes from [Bayley and Willett, 1999]. (a) Original data represented by $D_1 - D_2$ descriptors; (b) equisized independent binning scheme; (c) equifrequent independent binning scheme; (d) equisized dependent binning scheme; (e) equifrequent dependent binning scheme.

To select a subset of diverse compounds each molecule of the data set is assigned to the cell that matches the set of binned descriptors of the molecule; a structurally diverse subset is then obtained by selecting, for instance, one molecule from each of the cells to obtain the maximal coverage of the chemical space. On the contrary, in lead discovery, all the compounds falling in the same cell as a reference active compound are selected for further evaluation as candidates to be potential drugs.

Cell-based methods are significantly faster than distance-based methods, but are applicable to chemical space with low dimensionality (typically not more than five to six descriptors). In effect, when dimensionality is high, only a small fraction of the cells will be occupied, even with a low bin resolution. Moreover, the results are very sensitive to the grid resolution; indeed, if the bins are too large, the method loses its discriminating power; on the contrary, if the bins are too small, the data are very sparse, local behaviors of the data are highlighted and the general trend is lost. An algorithm based on a fractal approach was proposed to identify the optimal grid resolution by generating random subsets of k molecules from the whole data set, measuring their diversity at several grid resolutions, and identifying the resolution at which the relative variance of the diversity measure over all the random subsets assumes its maximum value [Agrafiotis and Rassokhin, 2002].

An advantage of cell-based methods is that they allow the explicit identification of those regions of the chemical space that are underrepresented, or, even unrepresented (i.e., diversity voids), in a database thus suggesting alternative potential structures to those of the existing chemicals [Pearlman and Smith, 1998].

Several cell-based diversity measures have been proposed in the literature [Pascual, Borrell *et al.*, 2003; Pascual, Mateu *et al.*, 2003]. These are $\rightarrow$ *concentration indices*, such as χ^2 statistics, Gini concentration ratio, and Pratt measure, used to evaluate the distribution of compounds throughout the grid specified by a binning scheme and, thus, sometimes referred to as **occupancy numbers**.

The most natural index is the **cell occupancy ratio**, defined as the ratio of the number of occupied cells N_{OCC} over the total number of cells N_{TOT} :

$$COR = \frac{N_{\text{OCC}}}{N_{\text{TOT}}}$$

Another occupancy measure is the χ^2 statistics defined as

$$\chi^2 = \sum_{k=1}^{N_{\text{TOT}}} \frac{(n_k - n_k^*)^2}{n_k^*}$$

where n_k is the number of compounds in the k th cell and n_k^* is the expected theoretical number of compounds for the k th cell, that is, n/N_{TOT} , n being the total number of compounds in the data set.

Cell-based entropy (I_{cell}) and **cell-based density** (H_{cell}), both based on $\rightarrow$ *information content*, are defined as

$$I_{\text{cell}} = - \sum_{k=1}^{N_{\text{TOT}}} (n_k \cdot \log_2(n_k)) \quad \text{and} \quad H_{\text{cell}} = - \sum_{k=1}^{N_{\text{TOT}}} \left(n_k \cdot \log_2 \left(\frac{n_k}{n_k^{\text{REF}}} \right) \right)$$

where n_k^{REF} is the number of compounds in each k th cell from a reference data set.

To quantify the performance of the partition when used for lead discovery purposes, each compound has to be associated with some activity data and the active cells are defined as those containing at least one active compound. Then, common $\rightarrow$ *classification parameters for two-class problems* can be used. For instance, a partition performance measure is the deviation from the ideal situation evaluated by summing the number of inactive compounds found within the active cells and then dividing the sum by the total number of active molecules, that is, $\rightarrow$ *error rate*.

$\rightarrow$ *Property filters* are a particular implementation of partitioning methods; they are used to select drug-like or lead-like compounds from large chemical libraries. Like the cell-based methods, these filters are based on a partition of the chemical space but each selected molecular descriptor is divided into only two or three subranges of values. While property filters mainly aim at optimizing drug-likeness, cell-based methods at optimizing diversity of chemical libraries.

Three applications of cell-based methods are reported below.

The **Diverse Property-Derived method (DPD method)** is based on the partitioning of six noncorrelated molecular descriptors and physico-chemical properties [Ashton, Jaye *et al.*, 1996]. These are a lipophilicity descriptor (CLOGP), an electrotopological index calculated as normalized sum of the squares of the atomic $\rightarrow$ *electrotopological state indices*, the number of hydrogen-bond acceptors (*HBA*), the number of hydrogen-bond donors (*HBD*), a flexibility index defined as the ratio of the $\rightarrow$ *Kier shape descriptors* $^1\kappa$ over $^2\kappa$, and the aromatic density defined as the number of aromatic rings over the molecular volume (Table C1).

Table C1 Descriptors and ranges used in DPD method.

Descriptor	Bins	Ranges
CLOGP	4	<1.5; 1.5–4.0; >4.0; N^+
Electrotopological index	3	<10; 10–20; >20
HBA	2	0–2; ≥ 3
HBD	2	0–1; ≥ 2
Flexibility index	3	<3.5; 3.5–6.5; >6.5
Aromatic density	4	0; 0.01–3.5; 3.5–6.5; >6.5

Analogously to $\rightarrow$ *Property and Pharmacophoric Features fingerprints (PPF fingerprints)*, **PDR-FP fingerprints** (or *Property Descriptor value Range-derived FingerPrints*) encode value ranges of 93 molecular descriptors that were selected on the basis of their potential to adopt class-selective value ranges for different activity classes [Eckert and Bajorath, 2006a]. Molecular descriptors were selected among those implemented in the program MOE, on the basis of a comparison of their value distribution in 26 different classes of activity. To select the descriptors that systematically respond to molecular features related to the different activities, the DynaMAD scoring function [Eckert, Vogt *et al.*, 2006] was used, which relates descriptor scores to the probability p of compounds to map a given activity range:

$$\text{score} = [1 - p(\text{classMin} \leq x \leq \text{classMax})] \cdot 100$$

where *classMin* and *classMax* are the minimum and maximum values within a class, respectively. This scoring function produces scores between zero, corresponding to no selectivity, and 100, corresponding to optimal selectivity of the descriptor.

Value ranges of molecular descriptors are encoded using from two to seven equiprobable nonoverlapping bins, leading to a final vector of 500 bits.

To generate the binary PDR-FP of a compound, its values for the 93 molecular descriptors are calculated, and for each descriptor (represented by n bins), it is determined into which of the predefined n bins the compound descriptor value falls. The associated bit is then set to 1; all other $n - 1$ bits are set to 0. When this scheme is followed, the bit string representation of any compound has exactly 93 bits that are set on.

Applications of PDR-FP fingerprints reported in literature are [Eckert and Bajorath, 2007a; Wang, Eckert *et al.*, 2007; Tovar, Eckert *et al.*, 2008].

The **Joint Entropy-based Diversity Analysis** (JEDA) is a method to select representative subsets of compounds from combinatorial libraries by using a scoring function based on the $\rightarrow$ *Shannon's entropy* and implemented in a probabilistic search algorithm [Landon and Schaus, 2006].

Unlike other cell-based diversity methods, which select one compound from each cell of the chemical space, JEDA allows selection of more than one compound belonging to the same cell.

Compounds are described by a number of molecular descriptors; these are first normalized and then subjected to the $\rightarrow$ *Principal Component Analysis* to reduce the dimensionality of the chemical space. The M most significant principal components are successively transformed into binary vectors where each bit corresponds to a single principal component (PC): the bit can be either 0 or 1 depending on whether the PC value is smaller or greater than the median of that component calculated on the whole library [Xue, Godden *et al.*, 2003b].

The median is here calculated as the value at which the entropy H of a molecular descriptor is maximal for the considered library:

$$H(t) = -\left(\frac{n_b}{n} \cdot \log_2 \frac{n_b}{n} + \frac{n_a}{n} \cdot \log_2 \frac{n_a}{n}\right)$$

where n , n_a , and n_b are the total number of compounds in the library, the number of descriptor values that fall above (a) and below (b) a threshold value t , respectively.

Because the chemical space is defined by M principal components and each component is partitioned into two regions, the chemical space is divided into 2^M cells.

To select the optimal subset of compounds, that is, a set of compounds having the maximal chemical diversity, a probabilistic search algorithm is applied, which consists in selecting a subset of compounds based on a probability assigned to each compound. This algorithm optimizes the $\rightarrow$ *joint entropy* (JH) of the subset of selected compounds. The task is performed iteratively, assigning each i th compound an initial uniform probability $p_i = 1/n$, then calculating the score s_i that is added to the previous compound probability as

$$p'_i = p_i + s_i, \quad s_i = 1 - p_i^{\left(\frac{JH}{n^T}\right)}$$

and renormalizing, at each step, the probability of the remaining compounds in the library so that the total probability of the library is equal to 1.

The exponent of the score function includes the joint entropy JH of the selected subset, the number of library compounds, and an adjustable parameter T used to control the speed of the search.

After the probability of compounds has been updated, the process is repeated until the probabilities of the compounds in the selected subset sum to 1.

📖 [Godden, Xue *et al.*, 2002b]

- **cell occupancy ratio** → cell-based methods
- **center distance-based criteria** → center of a graph

■ center of a graph

This is the set of central vertices and edges, whose definitions depend on the approach used to determine them [Bonchev, 1989]. The graph center can be a single vertex, a single edge, or a single group of equivalent vertices. Several graph center definitions are derived from approaches aimed at → *canonical numbering* of vertices. Other ways to identify central vertices are the → *pruning of the graph* and the application of → *centric operator* and → *centrocomplexity operator* to → *layer matrices* of the graph [Diudea, Horvath *et al.*, 1992].

According to the most popular definition, the central vertices in a graph are those vertices having the smallest → *atom eccentricity*. In acyclic graphs the center coincides with a single vertex (i.e., a central vertex) or two adjacent vertices (i.e., a single central edge), while in cyclic-containing graphs, it usually coincides with a group of vertices. Other local vertex invariants give information useful to distinguish between terminal and central vertices. → *Centric indices* are molecular descriptors that quantify the degree of compactness of molecules based on the recognition of the graph center.

Specifically applied to study general networks but valid also for molecular graphs, some simple descriptors of vertex centrality are the so-called **centrality measures**. The concept of centrality is related to the ability of a vertex to communicate with other vertices or to its closeness to many other vertices or to the number of pairs of vertices that need a specific vertex as intermediary in their communications [Freeman, 1977, 1979; Estrada, 2006b]. The two simplest centrality measures are the → *vertex degree* and the **degree centrality** DC_i , that is, the number of paths starting/ending at a vertex i [Albert, Jeong *et al.*, 1999].

The **betweenness centrality** BC_i characterizes the degree of influence a vertex has in communicating between vertex pairs and is defined as the fraction of shortest paths going through a given vertex i as [Freeman, 1977; Newman, 2005]

$$BC_i = \sum_{k=1}^{A-1} \sum_{j=k+1}^A \frac{\min P_{kj}(i)}{\min P_{kj}} \quad k, j \neq i$$

where $\min P_{kj}$ is the number of shortest paths connecting vertices k and j , and $\min P_{kj}(i)$ is the number of these shortest paths that pass through the vertex i . Moreover, a relative measure of betweenness centrality BC'_i is obtained by dividing the betweenness centrality BC_i by the maximal value relative to the central vertex of the corresponding → *star graph* as

$$BC'_i = \frac{2 \cdot BC_i}{A^2 - 3 \cdot A + 2}$$

where A is the number of vertices in the graph. From the relative betweenness centrality, a measure of dominance of the most central vertex is defined as

$$BC' = \frac{\sum_{i=1}^A [BC^* - BC'_i]}{A - 1}$$

where BC^* is the maximal centrality value for any vertex in the graph, that is, $\max(BC'_i)$.

The **closeness centrality** CC_i of the i th vertex is defined as [Freeman, 1979; Albert, Jeong *et al.*, 1999]

$$CC_i = \frac{A-1}{\sum_{j=1}^A d_{ij}} = \frac{A-1}{\sigma_i}$$

where A is the number of vertices in the graph and σ_i the $\rightarrow$ *vertex distance degree*, that is, the sum of all distances from the i th vertex; the quantity σ_i , in the network context, is called **farness**.

The **eigenvector centrality** EC_i of a vertex i is derived from the leading eigenvector of the $\rightarrow$ *adjacency matrix* $\mathbf{A}$ representing a connected subgraph or component of the network [Bonacich, 1972, 2007]. It is defined as the i th component of the eigenvector associated to the largest eigenvalue of $\mathbf{A}$:

$$EC_i = \ell_{i1}$$

A vertex has high value of EC either if it is connected to many other vertices or if it is connected to others that themselves have high EC ; in effect, unlike degree centrality, which weights every neighbor equally, the eigenvector weights connections with others according to their centralities.

The **information centrality** IC_i is based on the information that can be transmitted between any two vertices in a connected network [Stephenson and Zelen, 1989]. It is defined as follows:

$$IC_i = \left[\frac{1}{A} \cdot \sum_{j=1}^A \frac{1}{[\mathbf{I}]_{ij}} \right]^{-1} \quad [\mathbf{I}]_{ij} = \begin{cases} (c_{ii} + c_{jj} - c_{ij})^{-1} & \text{if } i \neq j \\ \infty & \text{if } i = j \end{cases}$$

where c_{ij} are the elements of the matrix $\mathbf{C}$ obtained by inverting the matrix $\mathbf{B}$, that is strictly related to the $\rightarrow$ *Laplacian matrix* and defined as

$$\mathbf{B} = \mathbf{V} - \mathbf{A} + \mathbf{U}$$

$\mathbf{A}$ is the adjacency matrix, $\mathbf{V}$ the $\rightarrow$ *vertex degree matrix*, that is, the diagonal matrix of the vertex degrees, and $\mathbf{U}$ is the $\rightarrow$ *unit matrix* with all its elements equal to one.

The $\rightarrow$ *subgraph centrality* $C_S(i)$ accounts for the weighted participation of vertices in all subgraphs of the network and is defined as [Estrada and Rodríguez-Velázquez, 2005b]

$$C_S(i) = \sum_{j=1}^A (\ell_{ij})^2 \cdot e^{\lambda_j}$$

where ℓ_{ij} is the i th component of the eigenvector associated to the j th eigenvalue λ_j of the adjacency matrix. This index counts the times that a vertex takes part in the different connected subgraphs of the network, with smaller subgraphs having higher importance.

A **generalized graph center** concept is obtained by a hierarchy of criteria applied recursively so as to reduce the number of vertices qualifying as central vertices [Bonchev, Balaban *et al.*, 1980, 1981].

The graph **center distance-based criteria** 1D–4D for a vertex v_i to belong to the graph center are: *Criterion 1D* $\equiv$ minimum $\rightarrow$ *atom eccentricity* η_i (i.e., the largest distance from the i th vertex):

$$\min_i(\eta_i)$$

Criterion 2D $\equiv$ for the vertices satisfying the first criterion, minimum $\rightarrow$ vertex distance degree σ_i :

$$\min_i(\sigma_i)$$

Criterion 3D $\equiv$ for the vertices satisfying the previous criteria, minimum number of occurrences of the largest distance in the $\rightarrow$ vertex distance code:

$$\min_i({}^{\eta_i}f_i)$$

where ${}^{\eta_i}f_i$ is the frequency of the maximum distance η_i from the vertex v_i to any other vertex, that is, atom eccentricity. If the largest distance occurs the same number of times for two or more vertices, the frequency of the next largest distance $\eta_i - 1$ is considered and so on.

The graph vertices qualified as central according to the first three criteria constitute a smaller graph called **pseudocenter** (or **graph kernel**).

Criterion 4D $\equiv$ iterative process of the first three criteria 1D–3D applied to the pseudocenter instead of the whole graph.

If more than one central vertex results from distance-based criteria, the graph center is called **polycenter**.

To discriminate even further among the vertices of the polycenter, the graph **center path-based criteria** 1P–4P can be applied to the polycenter:

Criterion 1P $\equiv$ minimum $\rightarrow$ vertex path eccentricity ${}^{\Delta}\eta_i$ (i.e., the largest distance from the i th vertex in the detour matrix):

$$\min_i({}^{\Delta}\eta_i)$$

Criterion 2P $\equiv$ for vertices satisfying the first criterion, minimum $\rightarrow$ vertex path sum π_i (i.e., the sum of the lengths m of all paths starting from the considered vertex v_i):

$$\min_i(\pi_i)$$

Criterion 3P $\equiv$ for the vertices satisfying the previous criteria, minimum number of occurrences of the largest order path in the $\rightarrow$ vertex path code:

$$\min_i({}^{\Delta}\eta_i P_i)$$

where ${}^{\Delta}\eta_i P_i$ is the number of paths of maximal length starting from vertex v_i to any other vertex. If the longest path occurs the same number of times for two or more vertices, the path count of the next largest order is considered and so on.

Criterion 4P $\equiv$ iterative process of the first three criteria 1P–3P applied to the small set of vertices selected according to the above described procedure.

The central vertices resulting from the criteria 1P–4P are called **oligocenter**. To further discriminate among the vertices of the oligocenter, analogous graph **center self-returning walk-based criteria** 1W–4W can be applied.

All the criteria defined above can also be applied to search for central edges in the graph using information provided from the $\rightarrow$ edge distance matrix. For example, the center distance-based criteria 1D–4D are defined as

Criterion 1D $\equiv$ minimum $\rightarrow$ bond eccentricity ${}^b\eta_i$:

$$\min_i({}^b\eta_i)$$

Criterion 2D $\equiv$ for the edges satisfying the first criterion, minimum $\rightarrow$ edge distance degree ${}^E\sigma_i$:

$$\min_i({}^E\sigma_i)$$

Criterion 3D $\equiv$ for the edges satisfying the previous criteria, minimum number of occurrences of the longest distance in the $\rightarrow$ edge distance code:

$$\min_i({}^b\eta_i f_i)$$

where ${}^b\eta_i f_i$ is the frequency of the maximum distance ${}^b\eta_i$ from the edge e_i to any other edge, that is, the bond eccentricity. If the longest distance occurs the same number of times for two or more edges, the frequency of the next longest distance ${}^b\eta_i - 1$ is considered and so on.

Criterion 4D $\equiv$ iterative process of the first three criteria 1D–3D applied to the pseudocenter instead of the whole graph.

Moreover, the application of the center distance-based criteria on simultaneously both the vertex distance matrix **D** and the edge distance matrix ${}^E\mathbf{D}$ resulted in a new algorithm, called **Iterative Vertex and Edge Centricity algorithm** (IVEC), for graph center definition and vertex $\rightarrow$ canonical numbering [Bonchev, Mekenyan *et al.*, 1989]. The graph center is selected through the sequential centric ordering of the graph vertices and edges, on the basis of their metric properties and incidence.

In the initial step, the distance-based criteria 1D–4D are applied to graph vertices to order them into equivalence classes identified by ranks 1, 2, 3, ... Rank 1 is assigned to the polycenter and the maximum rank to the most external vertices. Then the same procedure is applied to the edges on the basis of the edge distance matrix.

Additional discrimination within the vertex equivalence classes is obtained by summing the ranks of the edges incident to each vertex of the considered class. New ranks are assigned to the vertices on this basis: lower ranks are assigned to vertices with smaller sum. If the same rank sum is obtained for two or more vertices, the vertex for which the addendum in the sum is smaller is assigned the lower rank. The same operation is performed for the graph edges by summing the new ranks of their incident vertices.

The algorithm continues iteratively until the same vertex and edge equivalence classes are obtained in two consecutive iterations. The center of the graph then includes the vertices and edges of lowest rank.

A modified IVEC algorithm was also proposed to search for the center of graphs where multiple bonds are present [Balaban, Bonchev *et al.*, 1993].

📖 [Bonchev and Balaban, 1981, 1993; Bonchev, 1983; Barysz, Bonchev *et al.*, 1986]

■ center of a molecule ($\equiv$ molecule center)

Molecule centers are reference points used to calculate distributional properties of the molecule and, mathematically speaking, are the first-order moments of property distributions. Arithmetic mean and weighted arithmetic mean are the common way to calculate centers.

For example, the **geometric center** of a molecule is defined as the average value of atom coordinates calculated separately for each axis:

$$\bar{x} = \frac{1}{A} \cdot \sum_{i=1}^A x_i \quad \bar{y} = \frac{1}{A} \cdot \sum_{i=1}^A y_i \quad \bar{z} = \frac{1}{A} \cdot \sum_{i=1}^A z_i$$

where A is the number of atoms in the molecule.

The weighted center is analogously defined, but each i th atom coordinate is weighted by w_i , which represents an atomic property:

$$\bar{x} = \frac{1}{W} \cdot \sum_{i=1}^A w_i \cdot x_i \quad \bar{y} = \frac{1}{W} \cdot \sum_{i=1}^A w_i \cdot y_i \quad \bar{z} = \frac{1}{W} \cdot \sum_{i=1}^A w_i \cdot z_i$$

where W is the sum of the weights over all atoms in the molecule. For example, if the $\rightarrow$ *weighting scheme* w is based on the atomic masses m , the **center of mass** (or **barycenter**) of the molecule is obtained.

Geometric and mass centers of a molecule are not molecular descriptors, but they are commonly used as the reference origin in the calculation of several geometric descriptors to obtain invariance to translation and rotation of molecules (i.e., $\rightarrow$ *TRI descriptors*).

By weighting atoms by atomic charges, the first-order moment of charges is the $\rightarrow$ *dipole moment* in neutral molecules. Moreover, for molecules with zero net charge and non-vanishing dipole moment, the **center-of-dipole** was defined as the appropriate molecule center for multipolar expansions to obtain rotational invariance [Silverman and Platt, 1996].

Another definition of molecule center is obtained by applying the concept of *leverage* to 3D $\rightarrow$ *molecular geometry*, the **atom leverage-based center** is defined as the set of atoms with the minimum value of the diagonal elements h_i of the $\rightarrow$ *molecular influence matrix* $\mathbf{H}$ derived from the centered spatial coordinates of the atoms in a molecule [Todeschini and Consonni, 2000], that is,

$$\{a_i : h_i = \min_i(h_i)\}$$

The molecular influence matrix is calculated as

$$\mathbf{H} = \mathbf{M} \times (\mathbf{M}^T \times \mathbf{M})^{-1} \times \mathbf{M}^T$$

where $\mathbf{M}$ is the rectangular matrix of dimension $A \times 3$ of the atom spatial coordinates (x, y, z) , that is, the $\rightarrow$ *molecular matrix*. The diagonal values h_i are always between 0 and 1.

- **center-of-dipole** $\rightarrow$ center of a molecule
- **center of mass** $\rightarrow$ center of a molecule
- **center path-based criteria** $\rightarrow$ center of a graph
- **center self-returning walk-based criteria** $\rightarrow$ center of a graph
- **central edges** $\rightarrow$ graph
- **centrality measures** $\rightarrow$ center of a graph
- **central moments** $\equiv$ *moments about the mean* $\rightarrow$ statistical indices ($\odot$ moment statistical functions)
- **centralization** $\rightarrow$ distance matrix
- **central vertices** $\rightarrow$ graph

■ centric indices

$\rightarrow$ *Molecular descriptors* proposed to quantify the degree of compactness of molecules by distinguishing between molecular structures organized differently with respect to their centers. Based on the recognition of the $\rightarrow$ *graph center*, these indices are mainly defined by the

information theory concepts applied to a partition of the graph vertices made according to their positions relative to the center. Moreover, $\rightarrow$ *centric operator* and $\rightarrow$ *centrocomplexity operator* have been proposed to calculate $\rightarrow$ *local vertex invariants* from $\rightarrow$ *layer matrices* and corresponding molecular descriptors, which account for molecular centrality.

The main centric indices are listed below; they have been divided into two main groups, one containing the indices proposed by Balaban and the other indices proposed by Bonchev.

- **Balaban centric index (B)**

A topological index defined for acyclic graphs based on the **pruning of the graph**, a stepwise procedure for removing all the $\rightarrow$ *terminal vertices*, that is, vertices with a $\rightarrow$ *vertex degree* of one ($\delta_i = 1$), and the corresponding incident edges from the $\rightarrow$ *H-depleted molecular graph*. The vertices and edges removed at the k th step are n_k and the total number of steps to remove all vertices is R [Balaban, 1979].

The **pruning partition** of the graph is the reversed sequence of numbers n_k provided by the pruning procedure:

$$\{n_R, n_{R-1}, \dots, n_1\}$$

The pruning partition is related to $\rightarrow$ *molecular branching* and the reversed order of numbers n_k is due to the fact that the number of branches cannot decrease when starting from the center of the tree. Moreover, the first entry is always equal to one (center) or two (bicenter) and the pruning partition is a partition of A , that is, the number of graph vertices; this means that

$$\sum_{k=1}^R n_k = A$$

The Balaban centric index is calculated from the pruning partition in analogy with the $\rightarrow$ *first Zagreb index* M_1 as

$$B = \sum_{k=1}^R n_k^2$$

This index provides a measure of molecular branching: the higher the value of B , the more branched is the tree. It is called centric index because it reflects the topology of the tree as viewed from the center.

The **normalized centric index** C is derived by normalization, that is, imposing the same lower bound equal to zero for the least branched (linear) trees, on all the graphs. It is defined as

$$C = \frac{B - 2 \cdot A + U}{2}$$

where A is the number of graph vertices. The term U is defined as

$$U = \frac{1 - (-1)^A}{2} = \begin{cases} 0 & \text{if } A \text{ is even} \\ 1 & \text{if } A \text{ is odd} \end{cases}$$

The **binormalized centric index** C' is derived by a binormalization, that is, imposing on all the graphs the same lower bound and an upper bound equal to one for star graphs. In practice, it is

obtained from the normalized centric index C dividing it by the corresponding value of the star graph:

$$C' = \frac{B - 2 \cdot A + U}{(A - 2)^2 - 2 + U}$$

The binormalized centric index provides information on the topological shape of trees in a similar way to the $\rightarrow$ *binormalized quadratic index* Q' .

• **lopping centric information index ($\bar{I}_B$)**

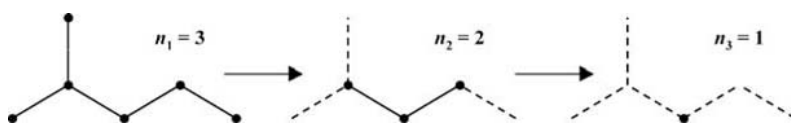
An index defined as the $\rightarrow$ *mean information content* derived from the pruning partition of a graph:

$$\bar{I}_B = - \sum_{k=1}^R \frac{n_k}{A} \cdot \log_2 \frac{n_k}{A}$$

where n_k is the number of terminal vertices removed at the k th step, A the number of graph vertices, and R the number of steps to remove all graph vertices [Balaban, 1979].

Example C2

Pruning partition of 2-methylpentane and some centric indices.



Pruning partition: {1, 2, 3}

$$B = n_1^2 + n_2^2 + n_3^2 = 3^2 + 2^2 + 1^2 = 14 \quad \bar{I}_B = -\frac{1}{6} \cdot \log_2 \frac{1}{6} - \frac{2}{6} \cdot \log_2 \frac{2}{6} - \frac{3}{6} \cdot \log_2 \frac{3}{6} = 1.459$$

$$C = \frac{B - 2 \cdot A + U}{2} = \frac{14 - 2 \cdot 6 + 0}{2} = 1 \quad C' = \frac{B - 2 \cdot A + U}{(A - 2)^2 - 2 + U} = \frac{14 - 2 \cdot 6 + 0}{(6 - 2)^2 - 2 + 0} = 0.143$$

• **information content based on center (IBC)**

Defined only for acyclic graphs and substituents, it is calculated as the $\rightarrow$ *total information content* based on the shells around the center of the graph:

$$IBC = 2W \cdot \log_2 2W - \sum_k q_k \cdot \log_2 q_k$$

where W is the $\rightarrow$ *Wiener index*, that is, the sum of all the distances in the graph and q_k is the sum of the $\rightarrow$ *vertex distance degree* (i.e., the sum of all distances from a vertex) of the vertices located at a $\rightarrow$ *topological distance* equal to k from the center [Balaban, Bertelsen *et al.*, 1994].

- **average information content based on center (AIBC)**

Defined only for acyclic graphs and substituents, it is the average of the information content based on center *IBC*, that is,

$$AIBC = \frac{IBC}{W}$$

IBC is divided by the $\rightarrow$ *Wiener index* *W* rather than $2W$ to have *AIBC* values higher than one [Balaban, Bertelsen *et al.*, 1994].

Bonchev centric information indices are centric indices derived from the vertex $\rightarrow$ *distance matrix* **D** and the $\rightarrow$ *edge distance matrix* **^ED**, based on the concept of graph center and calculated as $\rightarrow$ *mean information content* [Bonchev, Balaban *et al.*, 1980; Bonchev, 1983, 1989].

For **vertex centric indices**, the number of equivalent vertices in each equivalence class is calculated applying the $\rightarrow$ *center distance-based criteria* 1D–4D to the graph vertices, that is, the subsequent application of these criteria increases the discrimination of graph vertices. Analogously, for **edge centric indices**, the number of equivalent edges in each equivalence class is calculated applying the center distance-based criteria to the graph edges.

Once the $\rightarrow$ *polycenter* of the graph has been found, four other centric information indices, called **generalized centric information indices**, are calculated on the vertex (edge) partition based on the average topological distance between each vertex (edge) and the atoms of the polycenter. An increasing discrimination of the graph vertices (edges) is obtained by subsequently applying the remaining criteria 2D–4D.

Other centric information indices can be calculated by the same formulas on both vertex and edge graph partition based on graph $\rightarrow$ *center path-based criteria* 1P–4P and $\rightarrow$ *center self-returning walk-based criteria* 1W–4W. Moreover, **edge centric indices for multigraphs** have different values from those calculated on the parent graph, the edge distance matrix of the multigraph being different from the edge distance matrix of the parent graph.

The Bonchev centric information indices and the corresponding generalized centric information indices are listed below.

- **radial centric information index (${}^V\bar{I}_{C,R}$)**

It is defined as

$${}^V\bar{I}_{C,R} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having the same $\rightarrow$ *atom eccentricity*, that is, the maximum distance from a vertex to any other vertex in the graph, *G* the number of different vertex equivalence classes, and *A* the number of graph vertices.

- **distance degree centric index (${}^V\bar{I}_{C,deg}$)**

It is defined as

$${}^V\bar{I}_{C,deg} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having both the same atom eccentricity and the same $\rightarrow$ *vertex distance degree* (i.e., the sum of all distances from a vertex), G the number of different vertex equivalence classes, and A the number of graph vertices.

• **distance code centric index** (${}^V\bar{I}_{C,code}$)

It is defined as

$${}^V\bar{I}_{C,code} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices contemporarily having the same atom eccentricity, the same vertex distance degree, and the same $\rightarrow$ *vertex distance code* (i.e., occurrence number of distances of different length from a vertex), G the number of different vertex equivalence classes, and A the number of graph vertices.

• **complete centric index** (${}^V\bar{I}_{C,C}$)

It is defined as

$${}^V\bar{I}_{C,C} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices contemporarily having the same atom eccentricity, the same vertex $\rightarrow$ *distance degree*, the same $\rightarrow$ *vertex distance code*, but also distinguishing the vertices defining the $\rightarrow$ *pseudocenter*, that is, removing existing degeneracy of pseudocenter vertices. G is the number of different vertex equivalence classes and A is the number of graph vertices.

• **generalized radial centric information index** (${}^V\bar{I}_{C,R}^G$)

It is defined as

$${}^V\bar{I}_{C,R}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having the same average topological distance to the polycenter, G the number of different vertex equivalence classes, and A the number of graph vertices.

• **generalized distance degree centric index** (${}^V\bar{I}_{C,deg}^G$)

It is defined as

$${}^V\bar{I}_{C,deg}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having both the same average topological distance to the polycenter and the same vertex distance degree, G the number of different vertex equivalence classes, and A the number of graph vertices.

- **generalized distance code centric index** (${}^V\bar{I}_{C,code}^G$)

It is defined as

$${}^V\bar{I}_{C,code}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having the same average topological distance to the polycenter, the same vertex distance degree, and the same vertex distance code, G is the number of different vertex equivalence classes, and A the number of graph vertices.

- **generalized complete centric index** (${}^V\bar{I}_{C,C}^G$)

It is defined as

$${}^V\bar{I}_{C,C}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices contemporarily having the same average topological distance to the polycenter, the same vertex distance degree, the same $\rightarrow$ *vertex distance code*, but also distinguishing the atoms defining the pseudocenter, that is, removing existing degeneracy of pseudocenter atoms. G is the number of different vertex equivalence classes and A is the number of graph vertices.

- **edge radial centric information index** (${}^E\bar{I}_{C,R}$)

It is defined as

$${}^E\bar{I}_{C,R} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having the same $\rightarrow$ *bond eccentricity* (i.e., the maximum value in each i th row of the edge distance matrix), G the number of different edge equivalence classes, and B the number of edges.

- **edge distance degree centric index** (${}^E\bar{I}_{C,deg}$)

It is defined as

$${}^E\bar{I}_{C,deg} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having both the same $\rightarrow$ *bond eccentricity* and the same $\rightarrow$ *edge distance degree* (i.e., the sum of the i th row entries of the edge distance matrix), G the number of different edge equivalence classes, and B the number of graph edges.

- **edge distance code centric index** (${}^E\bar{I}_{C,code}$)

It is defined as

$${}^E\bar{I}_{C,code} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges contemporarily having the same $\rightarrow$ *bond eccentricity*, the same $\rightarrow$ *edge distance degree*, and the same $\rightarrow$ *edge distance code* (i.e., the occurrence of edge distance values for each i th edge), G is the number of different edge equivalence classes, and B the number of graph edges.

• **edge complete centric index** (${}^E\bar{I}_{C,C}$)

It is defined as

$${}^E\bar{I}_{C,C} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges contemporarily having the same $\rightarrow$ *bond eccentricity*, the same $\rightarrow$ *edge distance degree*, the same $\rightarrow$ *edge distance code*, but also distinguishing the edges defining the $\rightarrow$ *pseudocenter*, that is, removing existing degeneracy of pseudocenter edges. G is the number of different edge classes and B is the number of graph edges.

• **generalized edge radial centric information index** (${}^E\bar{I}_{C,R}^G$)

It is defined as

$${}^E\bar{I}_{C,R}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having the same average topological distance to the polycenter, G the number of different edge equivalence classes, and B the number of edges.

• **generalized edge distance degree centric index** (${}^E\bar{I}_{C,\text{deg}}^G$)

It is defined as

$${}^E\bar{I}_{C,\text{deg}}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having both the same average topological distance to the $\rightarrow$ *polycenter* and the same edge distance degree, G the number of different edge equivalence classes, and B the number of graph edges.

• **generalized edge distance code centric index** (${}^E\bar{I}_{C,\text{code}}^G$)

It is defined as

$${}^E\bar{I}_{C,\text{code}}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having the same average topological distance to the $\rightarrow$ *polycenter*, the same $\rightarrow$ *edge distance degree*, and the same $\rightarrow$ *edge distance code*; G is the number of different edge equivalence classes and B is the number of graph edges.

• **generalized edge complete centric index** (${}^E\bar{I}_{C,C}^G$)

It is defined as

$${}^E\bar{I}_{C,C}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges contemporarily having the same average topological distance to the $\rightarrow$ *polycenter*, the same $\rightarrow$ *edge distance degree*, the same $\rightarrow$ *edge distance code*, but also distinguishing the edges defining the pseudocenter, that is, removing existing degeneracy of pseudocenter edges. G is the number of different equivalence edge classes and B is the number of graph edges.

- **centricity** $\equiv$ *molecular centricity* $\rightarrow$ molecular complexity
- **centric operator** $\rightarrow$ layer matrices
- **centric topological index** $\rightarrow$ layer matrices
- **centrocomplexity operator** $\rightarrow$ layer matrices
- **centrocomplexity topological index** $\rightarrow$ layer matrices
- **CEP matrix** $\equiv$ *weighted electronic connectivity matrix* $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **CFM** $\equiv$ *Compressed Feature Matrix* $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **CGTA-axis system** $\rightarrow$ biodescriptors ($\odot$ DNA sequences)
- **CHAA₁ index** $\rightarrow$ charged partial surface area descriptors
- **CHAA₂ index** $\rightarrow$ charged partial surface area descriptors ($\odot$ CHAA₁ index)

■ **chainlength**

The chainlength is defined as the size of the longest heavy-atom chain in the molecule with none of the constituent atoms of the chain belonging to rings [Feher and Schmidt, 2003].

- **chain subgraph** $\rightarrow$ molecular graph
- **chance correlation** $\rightarrow$ validation techniques
- **characteristic graph** $\equiv$ *Sachs graph* $\rightarrow$ graph
- **characteristic polynomial** $\rightarrow$ algebraic operators

■ **characteristic polynomial-based descriptors**

These are various molecular descriptors derived from characteristic polynomials of the molecular graph G . They were originally used in the framework of the molecular orbital theory to study unsaturated compounds [Živković, Trinajstić *et al.*, 1975; Gutman, 1979, 1983; Graovac, Gutman *et al.*, 1977; Knop and Trinajstić, 1980; Rosenfeld and Gutman, 1989; Trinajstić, 1992; Gutman, Klavžar *et al.*, 2001; Noy, 2003]. Then, they were generalized to study any compound, finding a lot of applications in modeling physico-chemical properties of molecules [Hosoya, 1971, 1988; Trinajstić, 1988; Ivanciuc, Ivanciuc *et al.*, 1999b]. A comprehensive collection of characteristic polynomial-based descriptors with examples of calculation is presented in the reviews of Ivanciuc [Ivanciuc and Balaban, 1999c; Ivanciuc, Ivanciuc *et al.*, 1999a].

The $\rightarrow$ *characteristic polynomial* of the molecular graph is the characteristic polynomial of a $\rightarrow$ *graph-theoretical matrix* $\mathbf{M}$ derived from the graph [Graham and Lovasz, 1978; Diudea, Ivanciuc *et al.*, 1997; Diudea, Gutman *et al.*, 2001; Ivanciuc, 2001c]:

$$\begin{aligned} Ch(\mathbf{M}; w; x) &= \det(x\mathbf{I} - \mathbf{M}) = \sum_{i=0}^n (-1)^i c_i x^{n-i} \\ &= x^n - c_1 x^{n-1} + c_2 x^{n-2} + \dots + (-1)^{n-1} c_{n-1} x + (-1)^n c_n \end{aligned}$$

where “det” denotes the matrix determinant, $\mathbf{I}$ is the identity matrix of dimension $n \times n$, x is a scalar variable, and c_i are the $n + 1$ polynomial coefficients. $\mathbf{M}$ is any square $n \times n$ matrix computed on weighted or unweighted molecular graphs; w is the $\rightarrow$ *weighting scheme* applied to the molecular graph to encode chemical information. Note that $w = 1$ denotes unweighted graphs. If $\mathbf{M}$ is a vertex matrix then n is equal to A , the number of graph vertices, while, if $\mathbf{M}$ is an edge matrix, then n is equal to B , the number of graph edges. Polynomial coefficients are graph invariants and are thus related to the structure of a molecule graph.

A large number of graph polynomials were proposed in the literature, which differ from each other according to the molecular matrix $\mathbf{M}$ they are derived from, and the weighting scheme w used to characterize heteroatoms and bond multiplicity of molecules.

The most known polynomial is the characteristic polynomial of the $\rightarrow$ *adjacency matrix* ($\mathbf{M} = \mathbf{A}$), which is usually referred to as the **graph characteristic polynomial** [Harary, 1969a; Cvetković, Doob *et al.*, 1995; Bonchev and Rouvray, 1991; Trinajstić, 1992]:

$$Ch(\mathbf{A}; 1; x) = \det(x\mathbf{I} - \mathbf{A})$$

For any acyclic graph, two general rules are observed: (i) the power of x decreases by two and (ii) the absolute values of $Ch(\mathbf{A}; 1; x)$ coefficients are equal to the coefficients of the $\rightarrow$ *Z-counting polynomial* $Q(\mathcal{G}; x)$, which are the nonadjacent numbers $a(\mathcal{G}, k)$ of order k , that is, the numbers of k mutually nonincident edges [Nikolić, Plavšić *et al.*, 1992]. Moreover, if two graphs are $\rightarrow$ *isomorphic graphs*, their characteristic polynomials coincide, while the converse is not true, or, in other words, there exist nonisomorphic graphs with identical characteristic polynomials and spectra; these are called $\rightarrow$ *isospectral graphs* [Harary, King *et al.*, 1971; Herndon, 1974a; Randić, Trinajstić *et al.*, 1976].

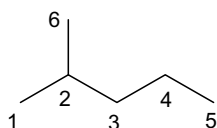
Depending on the elements of the matrix $\mathbf{M}$, characteristic polynomial can have very large coefficients and, spanning the x axis, often, asymptotic curves are obtained, whose characteristic points are not very representative as graph descriptors. To face with this problem, the characteristic polynomial can be transformed according to some **Hermite-like wave functions** for graphs, as [Gálvez, García-Domenech *et al.*, 2006]

$$\Psi \equiv Ch(\mathbf{M}; w; x) \exp\left(-\frac{x^2}{2}\right)$$

where $Ch(\mathbf{M}; w; x)$ is the characteristic polynomial of a graph. The most significant difference is that the area under the curve becomes finite in this approach, thus allowing the definition of more sound graph invariants, such as the *area under the curve* (AUC), the *maximum Ψ value* ($\Psi^{\max}$), and the *maximum amplitude* (MA) of the obtained sinusoidal curve.

Example C3

H-depleted molecular graph of 2-methylpentane and its adjacency matrix **A**.



$$\mathbf{A} =$$

Atom	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	1
3	0	1	0	1	0	0
4	0	0	1	0	1	0
5	0	0	0	1	0	0
6	0	1	0	0	0	0

The characteristic polynomial of the adjacency matrix of 2-methylpentane is

$$Ch(\mathbf{A}; x) = x^6 - 5 \cdot x^4 + 5 \cdot x^2$$

where coefficients c_1 , c_3 , c_5 , and c_6 are zero. Absolute values of nonzero coefficients are $|c_0| = 1$, which corresponds to the nonadjacent number of zero order, $a(\mathcal{G}, 0) = 1$ (by definition); $|c_2| = 5$, which corresponds to the nonadjacent number of first order, $a(\mathcal{G}, 1) = 5$ (the number of graph edges); $|c_4| = 5$, which corresponds to the nonadjacent number of second order, $a(\mathcal{G}, 2) = 5$ (the number of ways two edges may be selected so that they are nonadjacent).

The **Laplacian polynomial** is the characteristic polynomial of the $\rightarrow$ *Laplacian matrix* **L** of the molecular graph [Ivanciuc, 1993; Gutman, Lee *et al.*, 1994; Trinajstić, Babic *et al.*, 1994; Gutman, Vidović *et al.*, 2002d; Gutman, 2003b; Cash and Gutman, 2004]:

$$Ch(\mathbf{L}; 1; x) = \det(x\mathbf{I} - \mathbf{L})$$

The **distance polynomial** is the characteristic polynomial of the $\rightarrow$ *distance matrix* **D** of the molecular graph [Hosoya, Murakami *et al.*, 1973; Graham, Hoffman *et al.*, 1977; Graham and Lovasz, 1978]:

$$Ch(\mathbf{D}; 1; x) = \det(x\mathbf{I} - \mathbf{D}) = x^n - \sum_{i=1}^n c_i x^{n-1}$$

Note that the coefficients other than c_0 , which is always equal to one, are negative.

The **detour polynomial** is the characteristic polynomial of the $\rightarrow$ *detour matrix* **Δ** of the molecular graph [Nikolić, Trinajstić *et al.*, 1999b]:

$$Ch(\mathbf{\Delta}; 1; x) = \det(x\mathbf{I} - \mathbf{\Delta}) = x^n - \sum_{i=1}^n c_i x^{n-1}$$

As for the distance polynomial, the coefficients other than c_0 , which is always equal to one, are negative.

The **reciprocal distance polynomial** is the characteristic polynomial of the $\rightarrow$ *Harary matrix* **D⁻¹** of the molecular graph [Diudea, Ivanciuc *et al.*, 1997; Ivanciuc, Ivanciuc *et al.*, 1999b]:

$$Ch(\mathbf{D}^{-1}; 1; x) = \det(x\mathbf{I} - \mathbf{D}^{-1})$$

All these polynomials can also be calculated for the corresponding $\rightarrow$ *weighted matrices*, according to different $\rightarrow$ *weighting schemes* *w*.

Example C4

Laplacian, distance, reciprocal distance, and detour polynomials of 2-methylpentane are

$$Ch(\mathbf{L}; 1; x) = x^6 - 10 \cdot x^5 + 35 \cdot x^4 - 52 \cdot x^3 + 32 \cdot x^2 - 6 \cdot x$$

$$Ch(\mathbf{D}; 1; x) \equiv Ch(\mathbf{A}; 1; x) = x^6 - 84 \cdot x^4 - 368 \cdot x^3 - 580 \cdot x^2 - 368 \cdot x - 80$$

$$Ch(\mathbf{D}^{-1}; 1; x) = x^6 - 6.7083 \cdot x^4 - 8.3403 \cdot x^3 - 1.2843 \cdot x^2 + 1.9400 \cdot x + 0.6522$$

Note that, 2-methylpentane being an acyclic molecule, the detour polynomial coincides with the distance polynomial.

By analogy with the $\rightarrow$ Hosoya *Z* index that, for acyclic graphs, can be calculated as the sum of the absolute values of the coefficients of the characteristic polynomial of the adjacency matrix, the **stability index** (or **modified Z index**) is a molecular descriptor calculated for any graph as the sum of the absolute values of the coefficients c_{2i} , appearing alternatively in the characteristic polynomial of the adjacency matrix [Hosoya, Hosoi *et al.*, 1975]:

$$\tilde{Z} = \sum_{i=0}^{[A/2]} |c_{2i}|$$

where the square brackets indicate the greatest integer not exceeding $A/2$ and A is the number of graph vertices.

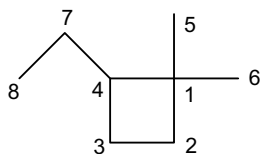
The same approach applied to the distance polynomial led to the definition of the **Hosoya Z' index** (or **Z' index**) [Hosoya, Murakami *et al.*, 1973]:

$$Z' = \sum_{i=0}^A |c_i|$$

where c_i are the coefficients of the distance polynomial of the molecular graph.

Example C5

H-depleted molecular graph and nonadjacent numbers of 4-ethyl-1,1-dimethylcyclobutane are



$$\begin{aligned} a(G, 0) &= 1 \\ a(G, 1) &= 8 \\ a(G, 2) &= 16 \\ a(G, 3) &= 8 \end{aligned}$$

The Hosoya *Z* index is $Z = 1 + 8 + 16 + 8 = 33$

The graph characteristic polynomial is $Ch(\mathbf{A}; 1; x) = x^8 - 8x^6 + 14x^4 - 6x^2$

The stability index is $\tilde{Z} = 1 + 8 + 14 + 6 = 29$

The distance polynomial is

$$Ch(\mathbf{D}; 1; x) = x^8 - 161x^6 - 1216x^5 - 3728x^4 - 5760x^3 - 4752x^2 - 2048x - 384$$

The *Z'* index is $Z' = 1 + 161 + 1216 + 3728 + 5760 + 4752 + 2048 + 384 = 18048$

An extension of the Z' index are the **Hosoya-type indices** that are defined as the sum of the absolute values of the coefficients of the characteristic polynomial of any square graph-theoretical matrix $\mathbf{M}$ [Ivanciuc, 1999c, 2001c]:

$$\text{Ho}(\mathbf{M}; w) = \sum_{i=0}^n |c_i|$$

where n is the matrix dimension and w the $\rightarrow$ *weighting scheme* applied to compute the matrix $\mathbf{M}$. The formula for the calculation of Hosoya-type indices was called by Ivanciuc **Hosoya operator**.

For any graph, when $\mathbf{M}$ is the distance matrix of a simple graph, $\text{Ho}(\mathbf{D}; 1) = Z'$, when $\mathbf{M}$ is the adjacency matrix of a simple graph, $\text{Ho}(\mathbf{A}; 1) = \tilde{Z}$; moreover, for acyclic graphs, when $\mathbf{M}$ is the adjacency matrix of a simple graph, $\text{Ho}(\mathbf{A}; 1) = \tilde{Z} = Z$ (Hosoya Z index).

Example C6

Hosoya-type indices derived from adjacency $\mathbf{A}$, Laplacian $\mathbf{L}$, distance $\mathbf{D}$, reciprocal distance $\mathbf{D}^{-1}$, and detour $\mathbf{\Delta}$ matrices of 2-methylpentane in the case of unweighted molecular graph ($w = 1$).

$$\text{Ho}(\mathbf{A}; 1) \equiv Z = 1 + 5 + 5 = 11$$

$$\text{Ho}(\mathbf{L}; 1) = 1 + 10 + 35 + 52 + 32 + 6 = 136$$

$$\text{Ho}(\mathbf{D}; 1) \equiv \text{Ho}(\mathbf{\Delta}; 1) = 1 + 84 + 368 + 580 + 368 + 80 = 1401$$

$$\text{Ho}(\mathbf{D}^{-1}; 1) = 1 + 6.7083 + 8.3403 + 1.2843 + 1.9400 + 0.6522 = 19.9251$$

Table C2 Some Hosoya-type indices for the data set of 18 octane isomers (Appendix C – Set 1) calculated on the unweighted molecular graph and the following graph-theoretical matrices: $\mathbf{A}$, adjacency matrix; $\mathbf{D}$, distance matrix; $\mathbf{L}$, Laplacian matrix; $\mathbf{D}^{-1}$, reciprocal distance matrix; χ , χ matrix; $\mathbf{G}^{-1}$, reciprocal geometry matrix.

C8	Ho(A;1)	Ho(D;1)	Ho(L;1)	Ho(D ⁻¹ ;1)	Ho(χ;1)	Ho(G ⁻¹ ;1)
n-Octane	34	34049	987	53.689	5.281	84663.6
2M	29	31028	932	55.879	4.625	68727.9
3M	31	30513	924	48.781	5.042	69915.5
4M	30	30424	923	44.932	4.958	71521.5
3E	32	29889	915	42.933	5.375	74399.4
22MM	23	26516	848	57.278	3.938	64305.2
23MM	27	27413	868	48.934	4.500	70235.7
24MM	26	27656	872	51.514	4.389	65051.7
25MM	25	28181	880	57.591	4.056	63337.2
33MM	25	25748	836	48.749	4.438	67147.2
34MM	29	26969	861	46.440	4.944	74731.5
2M3E	28	26864	860	44.505	4.833	72374.3
3M3E	28	25049	825	44.242	5.063	72784.0

(Continued)

Table C2 (Continued)

C8	Ho(A;1)	Ho(D;1)	Ho(L;1)	Ho(D ⁻¹ ;1)	Ho(x;1)	Ho(G ⁻¹ ;1)
223MMM	22	23168	788	51.570	3.917	70745.8
224MMM	19	23897	800	59.762	3.417	63623.3
233MMM	23	22925	784	48.839	4.083	71846.6
234MMM	24	24572	816	51.153	4.074	69364.3
2233MMMM	17	19685	720	55.535	3.125	74424.3

The characteristic polynomial encodes several important properties of the matrix, most notably its eigenvalues, spectral moments, determinant, and trace, which are largely used as molecular descriptors.

Information indices on polynomial coefficients are information indices defined as $\rightarrow$ *total information content* and $\rightarrow$ *mean information content* based on the partition of the coefficients of the characteristic polynomial of the graph. For acyclic molecules they coincide with the $\rightarrow$ *Hosoya total information index* and $\rightarrow$ *Hosoya mean information index*, respectively.

The **graph eigenvalues** λ_i are the roots of the characteristic polynomial of the matrix **M**, that is, the values of the x variable for which

$$Ch(\mathbf{M}; w; \lambda_i) = \det(\lambda_i \mathbf{I} - \mathbf{M}) = 0 \quad i = 1, n$$

where n is the size of the matrix **M**.

The complete set of the n eigenvalues of the matrix **M** is called **spectrum of the graph**:

$$\Lambda(\mathbf{M}; w) = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$$

Two fundamental properties of the characteristic polynomial of a matrix **M** are

$$tr(\mathbf{M}) = \sum_{i=1}^n \lambda_i = -c_1 \quad \det(\mathbf{M}) = \prod_{i=1}^n \lambda_i = (-1)^n \cdot c_n$$

where c_1 and c_n are two coefficients of the characteristic polynomial, “ tr ” denotes the trace of the matrix **M**, that is, the sum of its diagonal elements, and “ $\det$ ” the determinant of the matrix **M**. Therefore, it is noteworthy that the coefficient c_1 is always equal to zero for all the graph-theoretical matrices having zero on the main diagonal such as the adjacency, distance, and reciprocal distance matrices of simple graphs. For matrices derived from weighted molecular graphs, the coefficient c_1 , and, accordingly, the sum of the eigenvalues often is equal to the sum of the vertex weights, which are usually being located on the matrix main diagonal. For instance, in the Example C6, $-c_1$ in the characteristic polynomial of the Laplacian matrix **L** of 2-methylpentane is 10, which is the sum of the $\rightarrow$ *vertex degrees* δ_i , that is, the numbers of adjacent vertices: $\delta_1 = 1$, $\delta_2 = 3$, $\delta_3 = 2$, $\delta_4 = 2$, $\delta_5 = 1$, and $\delta_6 = 1$.

A large number of $\rightarrow$ *spectral indices*, which are molecular descriptors based on eigenvalues of molecular matrices, were defined both to study molecular graphs and model physico-chemical properties of molecules.

Spectral moments of the matrix **M**, denoted by $\mu^k(\mathbf{M}; w)$ and calculated with a weighting scheme w , are defined as

$$\mu^k(\mathbf{M}; w) = \sum_{i=1}^n \lambda_i^k = \sum_{i=1}^n [\mathbf{M}^k]_{ii}$$

where $k = 1, \dots, n$ is the order of the spectral moment, λ_i the eigenvalues of the matrix $\mathbf{M}$, and the last sum goes over the diagonal elements of the k th power of the matrix $\mathbf{M}$. Note that the spectral moment of order 1 simply is the sum of the matrix eigenvalues, coinciding with the sum of the diagonal elements of the matrix $\mathbf{M}$ and the coefficient c_1 of its characteristic polynomial:

$$\mu^1(\mathbf{M}; w) = \sum_{i=1}^n \lambda_i = \text{tr}(\mathbf{M}) = -c_1$$

If $\mathbf{M}$ is the adjacency matrix $\mathbf{A}$ of a simple graph, each diagonal element of the k th power of the adjacency matrix $\mathbf{A}^k$ is the $\rightarrow$ *atomic self-returning walk count* of order k , that is, the number of self-returning walks of length k of each atom, and, thus, the sum of the atomic self-returning walk counts of all the atoms is the $\rightarrow$ *molecular self-returning walk count* of order k , denoted as srw^k . Therefore, in the case of $\mathbf{M} = \mathbf{A}$, the following relationships hold:

$$\mu^k(\mathbf{A}; 1) = \text{srw}^k = \text{tr}(\mathbf{A}^k)$$

A number of theoretical studies and applications of spectral moments can be found in the literature [Živković, Trinajstić *et al.*, 1975; Jiang, Tang *et al.*, 1984; Hall, 1986; Kiang and Tang, 1986; Jiang and Zhang, 1989, 1990; Poshusta and McHughes, 1989; Marković and Gutman, 1991; Gutman, 1992a; Khadikar, Deshpande *et al.*, 1994; Bonchev and Seitz, 1995; Jiang, Qian *et al.*, 1995; Gutman and Rosenfeld, 1996; Helguera Morales, Cabrera Pérez *et al.*, 2006; Zhou, Gutman *et al.*, 2007].

$\rightarrow$ *Spectral moments of the edge adjacency matrix* and $\rightarrow$ *spectral moments of iterated line graph sequence* were largely investigated in QSAR/QSPR analysis by E. Estrada [Estrada, 1996, 1997, 1998a, 1998b, 1999c; Estrada, Peña *et al.*, 1998; Estrada and Gutierrez, 1999; Marković and Gutman, 1999; Marković, 1999, 2003; Marković, Marković *et al.*, 2001, 2002; Estrada, Paltewicz *et al.*, 2003].

Example C7

Eigenvalues and spectral moments of adjacency $\mathbf{A}$, Laplacian $\mathbf{L}$, distance $\mathbf{D}$, and reciprocal distance $\mathbf{D}^{-1}$ matrices for 2-methylpentane in the case of unweighted molecular graph.

$$\begin{aligned}\Lambda(\mathbf{A}; 1) &= \{1.9021; 1; 1756; 0; 0; -1.1756; -1.9021\} \\ \Lambda(\mathbf{L}; 1) &= \{4.2143; 3; 1.4608; 1; 0.3249; 0\} \\ \Lambda(\mathbf{D}; 1) &= \{11.0588; -0.5115; -0.6730; -1.1726; -2.0000; -6.1717\} \\ \Lambda(\mathbf{D}^{-1}; 1) &= \{3.0788; 0.4827; -0.5000; -0.5714; -1.1271; -1.3631\} \\ \mu(\mathbf{A}; 1) &= \{0; 10; 0; 30; 0; 100\} \\ \mu(\mathbf{L}; 1) &= \{10; 30; 106; 402; 1580; 6341.8\} \\ \mu(\mathbf{D}; 1) &= \{0.53; 166.5; 1107.3; 16425.6; 156413.2; 1884474\} \\ \mu(\mathbf{D}^{-1}; 1) &= \{0; 13.4; 25.0; 95.1; 270.0; 860.2\}\end{aligned}$$

Note that spectral moments of the distance matrix increase very quickly, thus requiring a proper scaling to be used in QSAR/QSPR modeling.

Graph eigenvectors are the eigenvectors associated with the eigenvalues λ_i of the characteristic polynomial of a matrix $\mathbf{M}$; for each $i = 1, \dots, n$, the following relationship holds:

$$Ch(\mathbf{M}; w; \lambda) = \det(\mathbf{M} - \lambda_i \mathbf{I}) = 0, \quad i = 1, \dots, n$$

where n being the number of eigenvalues and the size of the matrix $\mathbf{M}$.

Therefore, for each eigenvalue, $\mathbf{M} - \lambda_i \mathbf{I}$ is singular and, thus, it exists a nonzero n -dimensional vector $\mathbf{v}$ satisfying:

$$\mathbf{M} \cdot \mathbf{v}_i = \lambda_i \cdot \mathbf{v}_i$$

Any vector $\mathbf{v}$ satisfying this relationship is called eigenvector of $\mathbf{M}$ for the eigenvalue λ_i .

Based on the eigenvectors of the adjacency and distance matrices, $\rightarrow$ VEA indices, $\rightarrow$ VRA indices, $\rightarrow$ VED indices, and $\rightarrow$ VRD indices were proposed as molecular descriptors.

Characteristic polynomials belong to a more general class of graph polynomials, which are used to encode some information on molecular graphs. Among these, there are $\rightarrow$ Z-counting polynomial, $\rightarrow$ matching polynomial, and $\rightarrow$ Wiener polynomial.

Characteristic polynomials and Hosoya-type indices were also derived from $\rightarrow$ distance-valency matrices, $\rightarrow$ distance-path matrix, $\rightarrow$ reciprocal distance-path matrix, $\rightarrow$ distance-delta matrix, $\rightarrow$ Szeged matrices [Ivanciuc and Ivanciuc, 1999], $\rightarrow$ layer matrices, and $\rightarrow$ edge adjacency matrix.

Characteristic polynomial, spectrum, spectral moments, eigenvectors, and Hosoya-type indices were also computed on square molecular matrices encoding information about spatial interatomic distances such as the $\rightarrow$ geometry matrix $\mathbf{G}$ and the $\rightarrow$ reciprocal geometry matrix $\mathbf{G}^{-1}$ [Ivanciuc and Balaban, 1999c].

📖 [Balaban and Harary, 1971; Randić, Trinajstić *et al.*, 1976; Balasubramanian, 1982, 1984a 1984b; Balasubramanian and Randić, 1982; Randić, 1982, 1983; Krivka, Jericevic *et al.*, 1985; Barysz, Nikolić *et al.*, 1986; Dias, 1987a, 1987b; Ivanciuc, 1988a, 1992, 1998d, 1998e, 2001b; Trinajstić, 1988; Rosenfeld and Gutman, 1989; Balasubramanian, 1990; Živković, 1990; Shalabi, 1991; Randić, Müller *et al.*, 1997; Cash, 1999; John and Diudea, 2004]

- **characteristic ratio** $\rightarrow$ shape descriptors
- **characteristic root index** $\rightarrow$ spectral indices
- **characteristic sequences** $\rightarrow$ biodescriptors (☉ DNA sequences)
- **charge density matrix** $\rightarrow$ quantum-chemical descriptors

■ charge descriptors

These are $\rightarrow$ electronic descriptors defined in terms of $\rightarrow$ atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such as atoms, bonds, molecular fragments, and orbitals. The charges measure the extent of electronic density localization in a molecule: negative q_i values mean that excess electronic charge is at center i while positive values mean that center i is electron-deficient. Electrical charges in the molecule are the driving force of electrostatic interactions and it is well known that local electron densities or charges play a fundamental role in many chemical reactions, physico-chemical properties and receptor-ligand $\rightarrow$ binding affinity.

Charge descriptors are calculated by methods of $\rightarrow$ computational chemistry and are among $\rightarrow$ quantum-chemical descriptors [Lowe, 1978; Streitwieser, 1961]. In the framework of quantum chemistry, $\rightarrow$ population analysis is the basic tool used to calculate atomic charges and the most common approaches are the $\rightarrow$ Mulliken population analysis and $\rightarrow$ Löwdin population

analysis. Moreover, $\rightarrow$ *partial equalization of orbital electronegativity* is the most popular approach for the calculation of partial atomic charges.

Charge descriptors are derived from atomic charges in different ways and a list of the most known descriptors is presented below.

- **maximum positive charge ($Q_{\max}^+$)**

The maximum positive charge of the atoms in a molecule:

$$Q_{\max}^+ = \max_i(q_i^+)$$

where q^+ are the net atomic positive charges.

- **maximum negative charge ($Q_{\max}^-$)**

The maximum negative charge of the atoms in a molecule:

$$Q_{\max}^- = \max_i(q_i^-)$$

where q^- are the net atomic negative charges.

- **total positive charge (Q^+)**

The sum of all of the positive charges of the atoms in a molecule:

$$Q^+ = \sum_i q_i^+$$

where q^+ are the net atomic positive charges.

- **total negative charge (Q^-)**

The sum of all of the negative charges of the atoms in a molecule:

$$Q^- = \sum_i q_i^-$$

where q^- are the net atomic negative charges.

- **total absolute atomic charge (Q)**

The sum over all atoms in a molecule of the absolute values of the atomic charges q_i :

$$Q = \sum_i |q_i|$$

This is a measure of molecule polarity, also called **Electronic Charge Index (ECI)**; for example, it has been used to study amino acid side chains [Collantes and Dunn III, 1995].

Moreover, the summation of the partial atomic charges of all the carbon atoms of the molecule was proposed for modeling hydrocarbons and called – quite improperly – **electrotopological descriptor**, denoted as $-\sum q_C$. To model properties of alcohols, partial charges of oxygen atoms were also accounted for, thus resulting into a different charge descriptor, denoted as $-(\sum q_C + \sum q_O)$ [Arupiyoti and Iragavarapu, 1998].

The **charge polarization** is the mean absolute atomic charge in a molecule, defined as

$$P = \frac{\sum_i |q_i|}{A} = \frac{Q}{A}$$

where A is the number of atoms.

Another measure of molecular polarity is obtained from the **total square atomic charge**, defined as

$$Q^2 = \sum_i q_i^2$$

These total charge descriptors can also be calculated restricted to a molecular fragment as well as to a functional group.

- **potential of a charge distribution (ϕ)**

The theoretical potential function of a discrete charge distribution, that is, of atomic point charges q_i , is given at point $\mathbf{r}$ as the following:

$$\phi(\mathbf{r}) = \sum_{i=1}^A \frac{q_i}{r_i}$$

where r_i is the distance from each atom to the point $\mathbf{r}$.

- **Submolecular Polarity Parameter (SPP; ${}^1\Delta$)**

An electronic descriptor defined as the maximum excess charge difference for a pair of atoms in the molecule [Kaliszan, Osmialowski *et al.*, 1985; Osmialowski, Halkiewicz *et al.*, 1985], that is, calculated from the difference between the atomic maximum positive charge $Q_{\max}^+$ and the atomic maximum negative charge $Q_{\max}^-$ in a molecule:

$${}^1\Delta = |Q_{\max}^+ - Q_{\max}^-|$$

The **second-order submolecular polarity parameter** ${}^2\Delta$ is determined analogously, and is the second largest difference of excess charges [Luco, Yamin *et al.*, 1995].

The interatomic distance $r_{\pm}$ between the two atoms bearing the maximum positive and negative charges is used to derive the **DP descriptor** as follows:

$$DP = \frac{|Q_{\max}^+ - Q_{\max}^-|}{r_{\pm}^2} = \frac{{}^1\Delta}{r_{\pm}^2}$$

where the denominator accounts for the decreasing of atom interaction when interatomic distance increases.

- **topographic electronic descriptors (T^E)**

Topographic electronic descriptors are calculated from partial atomic charges q as the following [Osmialowski, Halkiewicz *et al.*, 1985, 1986; Katritzky and Gordeeva, 1993]:

$$T^E = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{|q_i - q_j|}{r_{ij}^2} \quad {}^cT^E = \sum_{b=1}^B \left(\frac{|q_i - q_j|}{r_{ij}^2} \right)_b$$

where the first index considers all pairs of atoms (both connected and disconnected) and the second is restricted to all pairs i - j of bonded atoms; r_{ij} are $\rightarrow$ interatomic distances; A and B are the number of atoms and bonds, respectively. These descriptors are calculated in such a way that they reflect, to some extent, differences in size, shape, and constitution, these quantities affecting the electronic charge distribution and interatomic distances of the molecules.

- **partial charge weighted topological electronic index (PCWT^E)**

A molecular electronic descriptor defined as [Osmialowski, Halkiewicz *et al.*, 1986]

$$\text{PCWT}^E = \frac{1}{Q_{\max}^-} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{|q_i - q_j|}{r_{ij}^2} = \frac{T^E}{Q_{\max}^-}$$

where q are the Zefirov partial atomic charges [Zefirov, Kirpichenok *et al.*, 1987] of the atoms i and j , r_{ij} the corresponding interatomic distance, and $Q_{\max}^-$ the maximum negative charge.

- **local dipole index (D)**

A molecular descriptor calculated as the average of the charge differences over all pairs i - j of bonded atoms:

$$D = \frac{\sum_b |q_i - q_j|_b}{B}$$

where B is the number of bonds [Clare and Supuran, 1994; Karelson, Lobanov *et al.*, 1996].

- **electronic-topological descriptors (E^T)**

Proposed by analogy with the $\rightarrow$ connectivity indices, they are calculated for a $\rightarrow$ hydrogen-included molecular graph using absolute values of partial charges q_i as vertex weights instead of $\rightarrow$ vertex degrees δ as [Katritzky and Gordeeva, 1993]

$$\begin{aligned} {}^0 E^T &= \sum_{i=1}^A (|q_i|)^{-1/2} & {}^1 E^T &= \sum_{b=1}^B (|q_i \cdot q_j|_b)^{-1/2} \\ {}^2 E^T &= \sum_{k=1}^{N_2} (|q_i \cdot q_l \cdot q_j|_k)^{-1/2} & {}^3 E^T &= \sum_{k=1}^{^3 P} (|q_i \cdot q_l \cdot q_h \cdot q_j|_k)^{-1/2} \end{aligned}$$

where A is the number of graph vertices, B the number of edges, N_2 the $\rightarrow$ connection number, and $^3 P$ the number of paths of length three. Each term in the summations is the inverse square root of the product of the absolute partial charges of the vertices contained in the considered path.

- **charge-related indices**

They are global molecular descriptors derived from a $\rightarrow$ H-depleted molecular graph where each vertex is weighted by a $\rightarrow$ local vertex invariant called **Atom-in-Structure Invariant Index (ASII)** defined as [Bangov, 1988]

$$\text{ASII}_i = \text{ASII}_i^0 - h_i + q_i$$

where ASII_i^0 is a standard value for the atom-type and hybridization state of each i th atom (Table C3), h_i the number of hydrogen atoms bonded to the i th atom, and q_i its net $\rightarrow$ atomic charge.

Table C3 Standard values of the Atom-in-Structure Invariant Index for different atom-types.

Atom	ASII ⁰	Atom	ASII ⁰
C sp ³	4	O sp ³	23
C sp ²	11	O sp ²	25
C sp ² (ar)	13	S	28
C sp	7	F	32

(Continued)

Table C3 (Continued)

Atom	ASII ⁰	Atom	ASII ⁰
N sp ³	15	Cl	33
N sp ²	18	Br	34
N sp	20	I	35

From *ASII*, the *ASIIg* index and **Charge Topological Index (CTI)** were derived as the following:

$$ASIIg = \frac{10}{\sqrt{\sum_{i=1}^A ASII_i}} \cdot \left[\sum_{b=1}^B (ASII_i \cdot ASII_j)_b \right]^{1/2}$$

$$CTI = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{ASII_i \cdot ASII_j}{d_{ij}}$$

where *A* and *B* are the number of atoms and bonds, respectively, d_{ij} the topological distance between atoms v_i and v_j , and the summation in *CTI* index runs over all the pairs of atoms. The *ASIIg* index was particularly useful in dealing with isomers [Bangov, 1990]; the *CTI* index was proposed as a highly discriminant index with a low degree of degeneracy [Demirev, Dyulgerov *et al.*, 1991].

📖 [Del Re, 1958; Buydens, Massart *et al.*, 1983; Gasteiger, Röse *et al.*, 1988; Abraham and Smith, 1988; Baumer, Sala *et al.*, 1989; Gombar and Enslein, 1990; Dixon and Jurs, 1992; Reynolds, Essex *et al.*, 1992; Palyulin, Baskin *et al.*, 1995; Hannongbua, Lawtrakul *et al.*, 1996a; Payares, Díaz *et al.*, 1997]

■ Charged Partial Surface Area descriptors ($\equiv$ CPSA descriptors)

These constitute a set of different descriptors [Stanton and Jurs, 1990] that combine shape and electronic information to characterize molecules and, therefore, encode features responsible for polar interactions between molecules. The molecule representation used for deriving CPSA descriptors views molecule atoms as hard spheres defined by the $\rightarrow$ *van der Waals radius*. The $\rightarrow$ *solvent-accessible surface area* SASA is used as molecular surface area; it is calculated using a sphere with a radius of 1.5 Å to approximate the contact surface formed when a water molecule interacts with the considered molecule. Moreover, the contact surface where polar interactions can take place is characterized by a specific electronic distribution obtained by mapping atomic partial charges on the solvent-accessible surface.

Let SA_a^+ and SA_a^- be the surface area contributions of the *a*th positive and negative atoms, respectively; q_a^+ and q_a^- the partial atomic charges for the *a*th positive and negative atoms; and Q^+ and Q^- the total sum of partial positive and negative charges in the molecule, respectively. The CPSA descriptors are defined as the following:

• partial negative surface area (PNSA₁)

It is the sum of the solvent-accessible surface areas of all negatively charged atoms, that is,

$$PNSA_1 = \sum_{a-} SA_a^-$$

where the sum is restricted to negatively charged atoms a^- .

- **partial positive surface area ($PPSA_1$)**

It is the sum of the solvent-accessible surface areas of all positively charged atoms, that is,

$$PPSA_1 = \sum_{a^+} SA_a^+$$

where the sum is restricted to positively charged atoms a^+ .

- **total charge weighted negative surface area ($PNSA_2$)**

It is the partial negative solvent-accessible surface area multiplied by the $\rightarrow$ total negative charge Q^- , that is,

$$PNSA_2 = Q^- \cdot \sum_{a^-} SA_a^-$$

- **total charge weighted positive surface area ($PPSA_2$)**

It is the partial positive solvent-accessible surface area multiplied by the $\rightarrow$ total positive charge Q^+ , that is,

$$PPSA_2 = Q^+ \cdot \sum_{a^+} SA_a^+$$

- **atomic charge weighted negative surface area ($PNSA_3$)**

It is the sum of the product of atomic solvent-accessible surface area by the partial charge q_a^- over all negatively charged atoms, that is,

$$PNSA_3 = \sum_{a^-} q_a^- SA_a^-$$

- **atomic charge weighted positive surface area ($PPSA_3$)**

It is the sum of the product of atomic solvent-accessible surface area by the partial charge q_a^+ over all positively charged atoms, that is,

$$PPSA_3 = \sum_{a^+} q_a^+ SA_a^+$$

- **difference in charged partial surface area ($DPSA_1$)**

It is the partial positive solvent-accessible surface area minus the partial negative solvent-accessible surface area, that is,

$$DPSA_1 = PPSA_1 - PNSA_1$$

- **difference in total charge weighted surface area ($DPSA_2$)**

It is the total charge weighted positive solvent-accessible surface area minus the total charge weighted negative solvent-accessible surface area, that is,

$$DPSA_2 = PPSA_2 - PNSA_2$$

- **difference in atomic charge weighted surface area ($DPSA_3$)**

It is the atomic charge weighted positive solvent-accessible surface area minus the atomic charge weighted negative solvent-accessible surface area, that is,

$$DPSA_3 = PPSA_3 - PNSA_3$$

- **fractional charged partial negative surface areas ($FNSA_1, FNSA_2, FNSA_3$)**

They are the partial negative surface area ($PNSA_1$), the total charge weighted negative surface area ($PNSA_2$), and the atomic charge weighted negative surface area ($PNSA_3$), divided by the total molecular solvent-accessible surface area ($SASA$), that is,

$$FNSA_1 = \frac{PNSA_1}{SASA} \quad FNSA_2 = \frac{PNSA_2}{SASA} \quad FNSA_3 = \frac{PNSA_3}{SASA}$$

- **fractional charged partial positive surface areas ($FPSA_1, FPSA_2, FPSA_3$)**

They are the partial positive surface area ($PPSA_1$), the total charge weighted positive surface area ($PPSA_2$), and the atomic charge weighted positive surface area ($PPSA_3$), divided by the total molecular solvent-accessible surface area ($SASA$), that is,

$$FPSA_1 = \frac{PPSA_1}{SASA} \quad FPSA_2 = \frac{PPSA_2}{SASA} \quad FPSA_3 = \frac{PPSA_3}{SASA}$$

- **surface weighted charged partial negative surface areas ($WNSA_1, WNSA_2, WNSA_3$)**

They are the partial negative surface area ($PNSA_1$), the total charge weighted negative surface area ($PNSA_2$), and the atomic charge weighted negative surface area ($PNSA_3$), multiplied by the total molecular solvent-accessible surface area ($SASA$) and divided by 1000, that is,

$$WNSA_1 = \frac{PNSA_1 \cdot SASA}{1000} \quad WNSA_2 = \frac{PNSA_2 \cdot SASA}{1000} \quad WNSA_3 = \frac{PNSA_3 \cdot SASA}{1000}$$

- **surface weighted charged partial positive surface areas ($WPSA_1, WPSA_2, WPSA_3$)**

They are the partial positive surface area ($PPSA_1$), the total charge weighted positive surface area ($PPSA_2$), and the atomic charge weighted positive surface area ($PPSA_3$), multiplied by the total molecular solvent-accessible surface area ($SASA$) and divided by 1000, that is,

$$WPSA_1 = \frac{PPSA_1 \cdot SASA}{1000} \quad WPSA_2 = \frac{PPSA_2 \cdot SASA}{1000} \quad WPSA_3 = \frac{PPSA_3 \cdot SASA}{1000}$$

- **relative negative charge ($RNCG$)**

It is the partial charge of the most negative atom divided by the $\rightarrow$ total negative charge, that is,

$$RNCG = \frac{Q_{\max}^-}{Q^-}$$

- **relative positive charge (RPCG)**

It is the partial charge of the most positive atom divided by the $\rightarrow$ total positive charge, that is,

$$RPCG = \frac{Q_{\max}^+}{Q^+}$$

- **relative negative charge surface area (RNCS)**

It is the solvent-accessible surface area of the most negative atom divided by the relative negative charge (RNCG), that is,

$$RNCS = \frac{SA_{\max}^-}{RNCG}$$

- **relative positive charge surface area (RPCS)**

It is the solvent-accessible surface area of the most positive atom divided by the relative positive charge (RPCG), that is,

$$RPCS = \frac{SA_{\max}^+}{RPCG}$$

- **total hydrophobic surface area (TASA)**

It is the sum of solvent-accessible surface areas of atoms with absolute value of partial charges less than 0.2, that is,

$$TASA = \sum_a SA_a \quad \forall a : |q_a| < 0.2$$

- **total polar surface area (TPSA)**

It is the sum of solvent-accessible surface areas of atoms with absolute value of partial charges greater than or equal to 0.2.

$$TPSA = \sum_a SA_a \quad \forall a : |q_a| \geq 0.2$$

- **relative hydrophobic surface area (RASA)**

It is the total hydrophobic surface area (TASA) divided by the total molecular solvent-accessible surface area (SASA), that is,

$$RASA = \frac{TASA}{SASA}$$

- **relative polar surface area (RPSA)**

It is the total polar surface area (TPSA) divided by the total molecular solvent-accessible surface area (SASA), that is,

$$RPSA = \frac{TPSA}{SASA}$$

Six additional CPSA descriptors were later proposed as [Aptula, Kühne *et al.*, 2003]

$$\begin{aligned} PPSA_4 &= \frac{Q^+}{A} \cdot \sum_{a^+} SA_a^+ & PNSA_4 &= \frac{Q^-}{A} \cdot \sum_{a^-} SA_a^- \\ PPSA_5 &= \frac{Q^+}{A^+} \cdot \sum_{a^+} SA_a^+ & PNSA_5 &= \frac{Q^-}{A^-} \cdot \sum_{a^-} SA_a^- \\ SPMX &= Q_{\max}^+ \cdot SA_{\max}^+ & SNMX &= Q_{\max}^- \cdot SA_{\max}^- \end{aligned}$$

where A is the total number of atoms, A^+ and A^- the total number of positively and negatively charged atoms, respectively, and the meaning of the other symbols is the same as above.

The set of CPSA descriptors was further developed to account for any particular type of polar interaction such as hydrogen-bonding. **Hydrogen-Bond Charged Partial Surface Area descriptors** (or **HB-CPSA descriptors**) were proposed in analogy with CPSA descriptors [Stanton, Egolf *et al.*, 1992]. Hydrogen-bond donor groups are considered to be any heteroatoms (i.e., O, S, or N) possessing a proton that can be donated. Other types of functional groups such as the alkynes were also included in the donor class. Acceptor groups include any functional group possessing sufficient electron density to participate in a hydrogen bond. To simplify the calculations the halogens, some double and some aromatic bonds were not included in the HB-CPSA descriptors.

Katritzky, Mu *et al.*, [Katritzky, Mu *et al.*, 1996b] later enlarged this set of hydrogen-bonding descriptors. All the H-bond descriptors are assigned zero if no hydrogen atoms in the molecule can be donated; moreover, hydrogen-bond acceptors are usually restricted to oxygen, nitrogen, and sulfur atoms (e.g., carbonyl oxygen atoms except in $-\text{COOR}$, hydroxy oxygen atoms, amino nitrogen atoms, aromatic nitrogens, and mercapto sulfur atoms).

The two simplest HB-CPSA descriptors are the $\rightarrow$ *hydrogen-bond acceptor number HBA* and the $\rightarrow$ *hydrogen-bond donor number HBD*.

Let SA_d and SA_a be the solvent accessible surface areas of hydrogen-bonding donors (d) and acceptors (a), respectively, $SASA$ the solvent-accessible surface area, and q_d and q_a the corresponding partial atomic charges. The HB-CPSA descriptors are then defined as follows (note that the two different symbols encountered in the literature for some are considered as synonymous).

- **RHTA index**

It is the ratio of the number of donor groups (HBD) over the number of acceptor groups (HBA), that is,

$$RHTA = \frac{HBD}{HBA}$$

- **SSAH index** ($\equiv$ *HDSA index*)

It is the sum of the surface areas of the hydrogens, which can be donated:

$$SSAH \equiv HDSA = \sum_d SA_d$$

- **RSAH index**

It is the average surface area of hydrogens, which can be donated:

$$RSAH = \frac{\sum_d SA_d}{HBD}$$

where HBD is the number of hydrogen-bond donors.

- **RSHM index** ($\equiv$ FHDSA index)

It is the fraction of the total molecular surface area associated with hydrogens, which can be donated:

$$RSHM \equiv FHDSA = \frac{\sum_d SA_d}{SASA}$$

- **SSAA index** ($\equiv$ HASA index)

It is the sum of the surface areas of all H-bond acceptor atoms:

$$SSAA \equiv HASA = \sum_a SA_a$$

The **HASA₂ index** is a variant of the HASA index defined as [Katritzky, Lobanov *et al.*, 1998]

$$HASA_2 = \sum_a \sqrt{SA_a}$$

- **RSAA index**

It is the average surface area of H-bond acceptor groups:

$$RSAA = \frac{\sum_a SA_a}{HBA}$$

where HBA is the number of hydrogen-bond acceptors.

- **RSAM index** ($\equiv$ FHASA index)

It is the fraction of the total molecular surface area associated with H-bond acceptor groups:

$$RSAM \equiv FHASA = \frac{\sum_a SA_a}{SASA}$$

Based on the HASA₂ index, the **FHASA₂ index** is defined as [Katritzky, Sild *et al.*, 1998c]

$$FHASA_2 = \frac{\sum_a \sqrt{SA_a}}{SASA}$$

- **HDCA index**

It is the sum of charged surface areas of hydrogens, which can be donated:

$$HDCA = \sum_d q_d \cdot SA_d$$

The charged surface area of hydrogens atoms, called **CSA2_H index**, and the charged surface area of chlorine atoms, called **CSA2_{Cl} index**, are two other similar H-bond descriptors defined as

$$CSA2_H = \sum_h q_h \cdot \sqrt{SA_h} \quad \text{and} \quad CSA2_{Cl} = \sum_{Cl} q_{Cl} \cdot \sqrt{SA_{Cl}}$$

where q_h , q_{Cl} , and SA_h , SA_{Cl} are partial atomic charge and solvent-accessible surface area of hydrogen and chlorine atoms, respectively [Katritzky, Lobanov *et al.*, 1998].

- **FHDCA index**

It is the charged surface area of hydrogens, which can be donated relative to the total molecular surface area:

$$FHDCA = \frac{\sum_d q_d \cdot SA_d}{SASA}$$

- **HDCA₂ index**

It is a hydrogen-bonding descriptor based on solvent-accessible area of hydrogen-bond donor atoms and corresponding partial charges proposed as variant of *FHDCA* index [Katritzky, Mu *et al.*, 1996a]:

$$HDCA_2 = \frac{\sum_d q_d \cdot \sqrt{SA_d}}{\sqrt{SASA}}$$

The summation is performed over the number of simultaneously possible hydrogen bonding donor and acceptor pairs per solute molecule; also hydrogen atoms attached to carbon atoms connected directly to carbonyl or cyano groups are considered as hydrogen bonding donors. The **HDCA₂ index** is another hydrogen-bonding donor descriptor with a definition similar to the *HDCA₂* index [Katritzky, Mu *et al.*, 1996b]:

$$HDSA_2 = \frac{\sum_d q_d \cdot \sqrt{SA_d}}{SASA}$$

where the summation is performed over all possible hydrogen bonding donor sites in a molecule.

- **HACA index** ($\equiv$ *SCAA₁ index*)

It is the sum of charged surface areas of hydrogen-bond acceptors:

$$HACA \equiv SCAA_1 = \sum_a q_a \cdot SA_a$$

An average charged surface area called the **SCAA₂ index** was also calculated as:

$$SCAA_2 = \frac{\sum_a q_a \cdot SA_a}{HBA}$$

where *HBA* is the number of hydrogen-bond acceptors [Turner, Costello *et al.*, 1998; Mitchell and Jurs, 1998b].

- **FHACA index**

It is the charged surface area of hydrogen-bond acceptors relative to the total molecular surface area:

$$FHACA = \frac{\sum_a q_a \cdot SA_a}{SASA}$$

- **HBSA index**

It is the sum of the surface areas of both hydrogens that can be donated and hydrogen acceptor atoms:

$$HBSA = HDSA + HASA$$

- **FHBSA index**

It is the surface area of both hydrogens that can be donated and hydrogen acceptor atoms relative to the total molecular surface area:

$$FHBSA = \frac{HBSA}{SASA}$$

- **HBCA index**

It is the sum of charged surface areas of both hydrogens that can be donated and hydrogen acceptor atoms:

$$HBCA = HDCA + HACA$$

- **FHBCA index**

It is the charged surface area of both hydrogens that can be donated and hydrogen acceptor atoms relative to the total molecular surface area:

$$FHBCA = \frac{HBCA}{SASA}$$

- **CHAA₁ index**

It is the sum of partial charges on hydrogen-bonding acceptor atoms [Mitchell and Jurs, 1998b]:

$$CHAA_1 = \sum_a q_a$$

The average value of CHAA₁ is called the **CHAA₂ index** and is defined as:

$$CHAA_2 = \frac{\sum_a q_a}{HBA}$$

where HBA is the number of hydrogen-bond acceptors.

- **ACGD index**

It is the average difference in charge between all pairs of H-bonding donors.

- **HRPCG index**

It is the relative positive charge (RPCG) restricted to H-bonding donor atoms.

- **HRNCG index**

It is the relative negative charge (RNCG) restricted to H-bonding acceptor atoms.

- **HRPCS index**

It is the relative positive charged surface area (*RPCS*) restricted to H-bonding donor atoms, that is, the positively charged surface area corresponding to the most positively charged atom that is also a possible hydrogen donor.

- **HRNCS index**

It is the relative negative charged surface area (*RNCS*) restricted to H-bonding acceptor atoms, that is, the negatively charged surface area corresponding to the most negatively charged atom that is also a possible hydrogen acceptor.

- **CHGD index**

It is the maximum difference in charge between a hydrogen that can be donated and its covalently-bonded heteroatom.

📖 [Stanton and Jurs, 1992; Nelson and Jurs, 1994; Mitchell and Jurs, 1998a; Eldred, Weikel *et al.*, 1999; Johnson and Jurs, 1999; Schweitzer and Morris, 1999; Katritzky, Maran *et al.*, 2000; De Rienzo, Grant *et al.*, 2002; Eike, Brennecke *et al.*, 2003; Schüürmann, Aptula *et al.*, 2003; Stanton, Mattioni *et al.*, 2004]

- **charge-matching function** → molecular shape analysis
- **charge polarization** → charge descriptors (⊙ total absolute atomic charge)
- **charge-related indices** → charge descriptors
- **charge term matrix** → topological charge indices
- **charge topological index** → charge descriptors (⊙ charge-related indices)
- **charge transfer constant** → electronic substituent constants
- **charge-transfer indices** ≡ *topological charge indices*
- **charge-weighted vertex connectivity indices** → connectivity indices
- **Charton characteristic volume** ≡ *Charton steric constant* → steric descriptors
- **Charton inductive constants** → electronic substituent constants (⊙ inductive electronic constants)
- **Charton steric constant** → steric descriptors
- **Chebyshev distance** ≡ *Lagrange distance* → similarity/diversity (⊙ Table S7)
- **ChemDiverse pharmacophore descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)

■ ChemGPS descriptors

ChemGPS (Chemical Global Positioning System) is a tool that positions novel structures in drug space via PCA-score prediction, providing a unique mapping device for the drug-like chemical space [Oprea and Gottfries, 2001b, 2001a].

Drug space map coordinates are the t-scores extracted via → *Principal Component Analysis*. PCA was performed on a total set of 423 *satellite* and *core* structures described by 72 descriptors representing size, lipophilicity, polarizability, charge, flexibility, rigidity, and hydrogen bond capacity.

Selected molecules include a set of “satellite” structures and a set of representative drugs (“core” structures). Satellites, intentionally placed outside drug space, have extreme values in one or several of the desired properties, while containing drug-like chemical fragments.

- **chemical adjacency matrix** $\equiv$ *atomic weight-weighted adjacency matrix* $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **chemical atom eccentricity** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **CHEMICALC** $\rightarrow$ lipophilicity descriptors ($\odot$ Suzuki–Kudo hydrophobic fragmental constants)
- **chemical descriptors** $\rightarrow$ molecular descriptors
- **chemical distance** $\rightarrow$ bond order indices ($\odot$ conventional bond order)
- **chemical distance degree** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **chemical distance matrix** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **chemical extended connectivity** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **chemical filters** $\equiv$ *functional group filters*
- **chemical formula** $\rightarrow$ molecular descriptors
- **chemical graph** $\rightarrow$ graph
- **chemical hardness** $\equiv$ *absolute hardness* $\rightarrow$ quantum-chemical descriptors ($\odot$ hardness indices)
- **chemical invariance** $\rightarrow$ molecular descriptors ($\odot$ invariance properties of molecular descriptors)
- **Chemically Advanced Template Search descriptors** $\equiv$ *CATS descriptors* $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **chemically intuitive molecular index** $\rightarrow$ spectral indices ($\odot$ Burden eigenvalues)
- **Chemical Shift Sum** $\rightarrow$ spectra descriptors
- **chemical space** $\rightarrow$ Structure/Response Correlations
- **chemodescriptors** $\rightarrow$ molecular descriptors

■ chemoinformatics

As Johann Gasteiger [Gasteiger, 2003b] says in his introduction to the *Handbook of Chemoinformatics*, “*Chemoinformatics is the use of informatics methods to solve chemical problems.*”

Gasteiger continues, “*It is clear that chemistry is a scientific discipline that is largely built on experimental observations and data. The amount of data and information accumulated is, however, enormous, and the size of this mountain . . . is increasing with increasing speed. The problem is, then, to extract knowledge from these data and this information, and use this knowledge to make predictions.*”

The term “chemoinformatics” was coined in 1998–1999 and rapidly gained widespread use and, as F. K. Brown says [Brown, 1998], “*Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.*”

Actually, chemoinformatics is not only related to drug design, but embraces under a unique umbrella all the classical chemical disciplines such as organic chemistry, analytical chemistry, physical chemistry, theoretical chemistry, medicinal chemistry, environmental chemistry, chemometrics, and so on, and more recent fields such as web-chemistry, chemical database managements, and library searching.

We would like also to highlight that in the framework of the chemoinformatics since the past century a fundamental role has been played by disciplines dealing with the mathematical aspects of chemistry such as graph theory, quantum-chemistry, and chemometrics [Balaban, 1978b; Trinajstić and Gutman, 2002]. Also molecular descriptors are of great relevance: they are indeed the basic tool in several chemoinformatics applications such as QSAR/QSPR modeling, drug discovery, similarity/diversity analysis, and library searching.

Some reviews and relevant lectures are reported below for the basic topics involved in chemoinformatics.

Chemoinformatics: [Gasteiger, 2003b, 2006; Gasteiger and Engel, 2003; Leach and Gillet, 2003; Agrafiotis, Bandyopadhyay *et al.*, 2007].

Molecular representation: [Aires-de-Sousa, 2003; Bangov, 2003; Barnard, 2003; Esposito, Hopfinger *et al.*, 2003; Gasteiger, 2003a; Karabunarliev, Nikolova *et al.*, 2003; Rohde, 2003; Sadowski, 2003; Wisniewski, 2003; Xu, 2003].

Molecular descriptors: [Devillers and Balaban, 1999; Karelson, 2000; Todeschini and Consonni, 2000].

Algorithms, multivariate techniques, quality control, and experimental design: [King, Srinivasan *et al.*, 2001; Booth, Isenhour *et al.*, 2003; Eriksson, Antti *et al.*, 2003; Hemmer, 2003; Leardi, 2003; Marsili, 2003; Rose, 2003; Varmuza, 2003; von Homeyer, 2003; Merkwirth, Mauser *et al.*, 2004].

Similarity/diversity analysis: [Farnum, DesJarlais *et al.*, 2003; Willett, 2003a; Maldonado, Doucet *et al.*, 2006].

QSAR and drug design: [Müller, 1997a; Olsson and Oprea, 2001; Xu and Hagler, 2002; Jurs, 2003; Kubinyi, 2003b; Nicklaus, 2003; Oprea, 2003; Selzer, 2003; Steinbeck, 2003; Sottriffer, Stahl *et al.*, 2003; García-Domenech, Gálvez *et al.*, 2008].

Bioinformatics: [Mewes, 2003; Rost, Liu *et al.*, 2003].

Web-chemistry: [Ertl and Jacob, 1997; Ertl, 1998a, 1998b, 2000; Augen, 2002; Ertl and Selzer, 2003; Steinbeck, Han *et al.*, 2003; Tarkhov, 2003; Tetko, Gasteiger *et al.*, 2005].

Database management and retrieval: [Ertl, 2003; Karabunarliev, Nikolova *et al.*, 2003; Neudert and Davies, 2003; Paris, 2003; Voigt, 2003; von Homeyer and Reitz, 2003; Wiggins, 2003; Zass, 2003; Adams and Schubert, 2004; Ósk Jónsdóttir, Jørgensen *et al.*, 2005].

■ chemometrics

Chemometrics is a discipline that deals with mathematical and statistical tools for analysis of complex chemical data [Brereton, 1990; Devillers and Karcher, 1991; Frank and Todeschini, 1994; van de Waterbeemd, 1995; Massart, Vandeginste *et al.*, 1997, 1998; Legendre and Legendre, 1998].

The main characterizing strategies are the multivariate approach to the problem, searching for relevant information, model validation to generate models with predictive power, comparison of the results obtained by using different methods, definition and use of indices capable of measuring the quality of extracted information and the obtained models.

Chemometrics finds a widespread use in QSAR and QSPR studies, in that it provides the basic tools for data analysis and modeling and a battery of different methods. Moreover, a relevant aspect of the chemometric philosophy is the attention it pays to the prediction power of models (estimated by using $\rightarrow$ *validation techniques*), $\rightarrow$ *model complexity*, and the continuous search for suitable parameters to assess the model qualities, such as $\rightarrow$ *classification parameters* and $\rightarrow$ *regression parameters*.

Chemometrics includes several topics of mathematics and statistics; some are listed below in alphabetic order.

• Artificial Neural Networks (ANN)

A set of mathematical methods, models, and algorithms designed to mimic information processing and knowledge acquisition methods of the human brain. ANNs are especially

suitable for dealing with nonlinear relationships and trends and are proposed for facing a large variety of mathematical problems such as data exploration, pattern recognition, modeling of continuous and categorized responses, multiple response problems, etc. [Livingstone, Manallack *et al.*, 1997; Zupan and Gasteiger, 1999; Anzali, Gasteiger *et al.*, 1998a; Niculescu, 2003; Zupan, 2003].

Some historically important artificial neural networks are *Hopfield Networks*, *Perceptron Networks* and *Adaline Networks*, while the most known are *Backpropagation Artificial Neural Networks* (BP-ANN), $\rightarrow$ *Self-Organizing Maps* (SOM), *Counter-Propagation Networks* (CP-ANN), *Radial Basis Function Networks* (RBFN), *Probabilistic Neural Networks* (PNN), *Generalized Regression Neural Networks* (GRNN), *Learning Vector Quantization Networks* (LVQ), and *Adaptive Bidirectional Associative Memory* (ABAM).

📖 Additional references are collected in the thematic bibliography (see Introduction).

• classification

Classification is assignment of objects to one of some classes based on a classification rule. The *classes* are defined *a priori* by groups of objects in a training set belonging to those classes. The goal is to calculate a *classification rule* and, possibly, *class boundaries* based on the training set objects of known classes, and to apply this rule to assign a class to objects of unknown classes [Hand, 1981, 1997; Frank and Friedman, 1989]. Classification methods are suitable for modeling several QSAR responses, such as, for example, active/nonactive compounds, low/medium/high toxic compounds, mutagenic/nonmutagenic compounds.

The most popular classification methods are *Linear Discriminant Analysis* (LDA), *Quadratic Discriminant Analysis* (QDA), *Regularized Discriminant Analysis* (RDA), *Kth Nearest Neighbors* (KNN), *classification tree methods* (such as CART), *Soft-Independent Modeling of Class Analogy* (SIMCA), *potential function classifiers* (PFC), *Nearest Mean Classifier* (NMC), *Weighted Nearest Mean Classifier* (WNMC), *Support Vector Machine* (SVM), and *Classification And Influence Matrix Analysis* (CAIMAN).

Moreover, several classification methods can be found among the artificial neural networks.

Classification model performance is evaluated by $\rightarrow$ *classification parameters*, both for fitting and predictive purposes.

📖 Additional references are collected in the thematic bibliography (see Introduction).

• cluster analysis

A special case of exploratory data analysis aimed at grouping similar objects in the same cluster and less similar objects in different clusters [Massart and Kaufman, 1983; Willett, 1987]. Cluster analysis is based on the evaluation of the $\rightarrow$ *similarity/diversity* of all the pairs of objects of a data set. This information is collected into the $\rightarrow$ *similarity matrix* or $\rightarrow$ *distance matrix*.

Many different methods were designed for cluster analysis; the most popular are the *hierarchical agglomerative methods* (i.e., *average linkage*, *complete linkage*, *single linkage*, *weighted average linkage*, etc.), which are more widely used than the *hierarchical divisive methods*. Other

very popular methods are *nonhierarchical methods*, such as *k-means method* and the *Jarvis–Patrick method*. Among the artificial neural networks dedicated to clustering, → *Self-Organizing Maps* are the most commonly used.

📖 [Dunn III and Wold, 1980; Dean and Callow, 1987; Nakayama, Shigezumi *et al.*, 1988; Willett, 1988; Lawson and Jurs, 1990; Jurs and Lawson, 1991; Good and Kuntz, 1995; Shemetulskis, Dunbar Jr *et al.*, 1995; Brown and Martin, 1996, 1997, 1998; Nouwen, Lindgren *et al.*, 1996; Dunbar Jr, 1997; Junghans and Pretsch, 1997; McGregor and Pallai, 1997; Reynolds, Druker *et al.*, 1998; Rose and Wood, 1998; Reijmers, Wehrens *et al.*, 2001; Rodriguez, Tomas *et al.*, 2005; Stanforth, Kolossov *et al.*, 2007]

• experimental design

Statistical procedures for planning an experiment, that is, collecting appropriate data that, after analysis by statistical methods, result in valid conclusions. The design includes the selection of experimental units, the specification of the experimental conditions, that is, the specification of factors whose effect will be studied on the outcome of the experiment, the specification of the level of the factors involved and the combination of such factors, the selection of response to be measured, and the choice of statistical model to fit the data [Box, Hunter *et al.*, 1978; Carlson, 1992; Livingstone, 1996; Lewis, Mathieu *et al.*, 1999].

An *experiment* consists of recording the values of a set of variables from a measurement process under a given set of experimental conditions.

The most known experimental designs are *complete factorial designs*, *fractional factorial designs*, *Plackett–Burman design*, *Dohelert design*, *composite designs*, and *optimal designs*.

📖 [Borth and McKay, 1985; Bonelli, Cechetti *et al.*, 1991; Pastor and Alvarez-Builla, 1991, 1994; Norinder, 1992; Norinder and Hogberg, 1992; Baroni, Clementi *et al.*, 1993; Baroni, Costantino *et al.*, 1993b; Marsili and Saller, 1993; Cruciani and Clementi, 1994; Rovero, Riganelli *et al.*, 1994; Austel, 1995; Sjöström and Eriksson, 1995; van de Waterbeemd, Costantino *et al.*, 1995; Borth, 1996; Eriksson and Johansson, 1996; Eriksson, Johansson *et al.*, 1997; Giraud, Luttmann *et al.*, 2000; Linusson, Gottfries *et al.*, 2000; Andersson and Lundstedt, 2002; Carro, Campisi *et al.*, 2002; Heimstad and Andersson, 2002; Eriksson, Arnhold *et al.*, 2004; Barroso and Besalú, 2005]

• exploratory data analysis

Exploratory data analysis is a collection of techniques that search for structure in a data set before calculating any statistic model [Krzanowski, 1988]. Its purpose is to obtain information about the data distribution, the presence of outliers and clusters, to disclose relationships and correlations between objects and/or variables. → *Principal component analysis* and *cluster analysis* are the most known techniques for data exploration [Jolliffe, 1986; Jackson, 1991; Basilevsky, 1994].

📖 [Weiner and Weiner, 1973; Stuper and Jurs, 1978; Wold, 1978; Cramer III, 1980a; Henry and Block, 1980a; Streich, Dove *et al.*, 1980; Alunni, Clementi *et al.*, 1983; McCabe, 1984; Takahashi, Miashita *et al.*, 1985; Dunn III and Wold, 1990; Cosentino, Moro *et al.*, 1992;

Livingstone, Evans *et al.*, 1992; Langer and Hoffmann, 1998a; Morais, Ramos *et al.*, 2001; Mazzatorta, Benfenati *et al.*, 2002; Hajduk, Mendoza *et al.*, 2003; Migliavacca, 2003; Restrepo and Villaveces, 2005]

• optimization

This is the procedure that allows to find the optimal value (minimum or maximum) of a numerical function f , called objective function, with respect to a set of parameters $\mathbf{p}$, $f(p_1, p_2, \dots, p_p)$. If the values that the parameters can take on are constrained, the procedure is called *constrained optimization*.

The most popular optimization techniques are *Newton–Raphson optimization*, *steepest ascent optimization*, *steepest descent optimization*, *Simplex optimization*, *Genetic Algorithm optimization*, and *simulated annealing*. More recent optimization techniques are *Particle swarm optimization* [Cedeño and Agrafiotis, 2003; Tang, Zhou *et al.*, 2007] and *ant colony optimization* [Izrailev and Agrafiotis, 2001a, 2001b]. Moreover $\rightarrow$ *variable reduction* and $\rightarrow$ *variable selection* are also among the optimization techniques.

📖 [Holland, 1975; Papadopoulos and Dean, 1991; Carlson, 1992; Hall, 1995; Kalivas, 1995; Handschuh, Wagener *et al.*, 1998; Sundaram and Venkatasubramanian, 1998; Wehrens, Pretsch *et al.*, 1998]

• ranking methods

Ranking methods are mathematical approaches that provide an ordering of the elements of a system. Ordering is one of the possible ways to analyze data and get an overview over the elements of a system [Pavan and Todeschini, 2008]. Ranking methods are largely used in **Multicriteria decision making** (MCDM) to take decisions about the studied objects (events, molecules, cases, scenarios, etc.) on the basis of more than one criterion [Hendriks, de Boer *et al.*, 1992; Carlson, 1992].

The different kinds of ranking methods available can be roughly classified as total ranking methods and partial-order ranking methods, according to the specific order they provide. These methods are the ones needed to support and solve (a) decision problems, that is, in defining a rank order of the available options, such as different procedures in analytical chemistry, (b) setting priorities, that is, to point out the most dangerous chemicals in a series of compounds, (c) defining global indices, that is, deriving environmental or material global quality indices from their multivariate characterization, and (d) evaluating characteristics of molecular descriptors, that is, analyzing the ranking of a series of compounds with respect to their branching description.

In the ranking techniques, the objects to be ordered can be any kind of objects, such as, for example, available alternatives, scenarios, chemicals, and molecular descriptors.

Specifically, ranking methods may be used to organize chemical information by harmonizing structural information, experimental knowledge, and other specific characteristics of the problem in analysis, such as environmental or health parameters.

Each i th object is represented by a set of p variables f_{ij} , ($j = 1, p$), that, in this framework, are also called criteria. In **total ranking methods**, these variables are joined into a global index, after some scaling procedure or some arbitrary transformation function, eventually using different weights for each criterion, in such a way that the actual value f_{ij} of each i th object for the j th

criterion assumes a value between 0 (worst case) and 1 (optimality). Then, based on the values of the calculated global index, the objects can be ordered as the following:

$$a \geq b \geq c \geq \dots \geq z$$

Simple additive ranking, desirability functions [Harrington, 1965], utility functions, and dominance functions are among the most used total ranking methods.

In **partial-order ranking methods**, a new relationship, which introduces the concept of noncomparability between two objects, is added to the classical ordering relationships. The partial-order ranking methods do not produce a global index useful for a total ranking, but use directly the original variables characterizing each object. The $\rightarrow$ *Hasse diagram* is an effective graphical tool to represent partial ordering. Partial ordering has been used to describe $\rightarrow$ *DNA sequences*.

Examples of applications of partial-order ranking methods and other theoretical aspects are reported in [Walczak and Massart, 1999; Klein and Bytautas, 2000; Carlsen, Sørensen *et al.*, 2001; Carlsen, Lerche *et al.*, 2002; Lerche, Sørensen *et al.*, 2003; Sørensen, Brüggemann *et al.*, 2003; Carlsen, 2004; Pavan and Todeschini, 2004; Voigt, Brüggemann *et al.*, 2004; Carlsen, 2005; Ivanciuc, Ivanciuc *et al.*, 2005; Pavan, Consonni *et al.*, 2005; Randić, Lerš *et al.*, 2005b; Todeschini, Consonni *et al.*, 2006]. Other applications of ranking methods can be found in [Bangov, 1988; Willett, 1988; Eriksson, Jonsson *et al.*, 1990; Ginn, Turner *et al.*, 1997; Randić, 2001e; Balaban, Mills *et al.*, 2002; Gramatica, Pilutti *et al.*, 2002, 2005; Russom, Breton *et al.*, 2003; Wilton and Willett, 2003; Hemmateenejad, 2004, 2005; Pavan, Mauri *et al.*, 2004; Papa, Battaini *et al.*, 2005; Batista and Bajorath, 2007; Gramatica and Papa, 2007; Todeschini, Ballabio *et al.*, 2007; Vogt, Godden *et al.*, 2007].

DART (*Decision Analysis by Ranking Techniques*) is a free available software implementing both partial-order ranking and several total ranking methods [DART – Milano Chemometrics, 2007].

• regression analysis

A set of statistical methods using a mathematical equation to model the relationship between an observed or measured response and one or more predictor variables. The goal of this analysis is twofold: modeling and predicting. The relationship is described in algebraic form as

$$y = f(x) + e \quad \text{or} \quad y = \mathbf{X} \cdot \mathbf{b}$$

where x denotes the predictor variable(s), y the response variable(s), $f(x)$ the systematic part of the model, and e the random error, also called model error or residual; y and $\mathbf{b}$ are the vectors of the responses and regression coefficients to be estimated, respectively; the matrix $\mathbf{X}$ is usually called *model matrix*, that is, its columns are the independent variables used in the regression model.

The mathematical equation used to describe the relationship between response and predictor variables is called *regression model* [Frank and Friedman, 1993; Wold, 1995; Ryan, 1997; Draper and Smith, 1998].

Regression analysis includes not only the estimation of model $\rightarrow$ *regression parameters*, but also the calculation of $\rightarrow$ *goodness of fit* and $\rightarrow$ *goodness of prediction* statistics, *regression diagnostics*, *residual analysis*, and *influence analysis* [Atkinson, 1985].

In particular, the **leverage matrix** $\mathbf{H}$, also called *influence matrix*, is an important tool in regression diagnostics containing information on the independent variables on which the model is built.

Let $\mathbf{X}$ be a matrix with n rows and p' columns, where p' is the number of model parameters. The leverage matrix $\mathbf{H}$ is a symmetric $n \times n$ matrix defined as

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T$$

where the matrix $\mathbf{X}$ is the model matrix. Moreover, a column where all the values are equal to one is added to the model matrix if the model is not constrained in the origin of the independent variables but an offset is allowed. To distinguish the two cases, a parameter c is used; $c = 1$ for the former and $c = 0$ for the latter.

The main properties of the leverage matrix are

$$\begin{aligned} \text{(a)} \quad & \frac{c}{n} \leq h_{ii} \leq 1 \quad \text{(b)} \quad \sum_{i=1}^n h_{ii} = p' \quad \text{(c)} \quad \bar{h} = \frac{\sum_{i=1}^n h_{ii}}{n} = \frac{p'}{n} \\ \text{(d)} \quad & \sum_{j=1}^n h_{ij} = c \quad \forall i \end{aligned}$$

where $\bar{h}$ is the average value of the leverage.

The leverage matrix is related to the response vector $\mathbf{y}$ by the following relationship:

$$\hat{\mathbf{y}} = \mathbf{H} \cdot \mathbf{y}$$

where $\hat{\mathbf{y}}$ is the calculated response vector from the model.

Usually the diagonal elements h_{ii} of the matrix $\mathbf{H}$ are those used for regression diagnostics: the i th object whose diagonal element h_{ii} is greater than two or three time the average value $\bar{h}$ can be considered as having a great influence (leverage) on the model.

Besides the well-known *Ordinary Least Squares regression (OLS)*, *biased regression*, *nonlinear regression*, and *robust regression* models are also important. The most popular biased methods are *Principal Component Regression (PCR)*, *Partial Least Squares regression (PLS)*, *Ridge Regression (RR)*, *Continuum Regression (CR)*, and *StepWise Regression (SWR)*.

Among the nonlinear methods, there are, besides *nonlinear least squares regression*, that is, *polynomial regression*, the *nonlinear PLS* method, *Alternating Conditional Expectations (ACE)*, *SMART*, and *MARS*. Moreover, some Artificial Neural Networks techniques have been specifically designed for nonlinear regression problems, such as the *back-propagation method*.

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **CHGD index** → charged partial surface area descriptors
- **CHI chirality descriptor** → chirality descriptors
- **CHI index** $\equiv$ *chromatographic hydrophobicity index* → chromatographic descriptors
- **Chi operator** → connectivity indices
- **chiral A_{xi} indices** → spectral indices ($\odot$ A_{xi} eigenvalue indices)
- **chiral connectivity indices** → chirality descriptors ($\odot$ topological chirality descriptors)
- **chiral factors** → weighted matrices ($\odot$ weighted distance matrices)

- **Chirality Codes** → chirality descriptors
- **chirality correction factor** → chirality descriptors (⊙ topological chirality descriptors)

■ chirality descriptors

A n -dimensional object is called *chiral* if it is nonsuperimposable on its mirror image by any rotation in the n -dimensional space. *Chirality* is the property of chiral objects and was perceived by Lord Kelvin in 1884 [Kelvin, 1904]: “I call any geometrical figure, or group of points, *chiral*, and say that it has *chirality*, if its image in a plane mirror, ideally realized, cannot be brought to coincide with itself.”

If looked in an isolation, → *physico-chemical properties* (and mathematical) of a chiral molecule and its antipodal counterpart all coincide. However, when chiral structures are considered in an environment, their behavior can be different, such as it occurs, for example, when a chiral molecule interacts with a receptor. Thus, chirality descriptors are useful for modeling properties related to interactions involving chiral centers [Aires-de-Sousa, 2003].

Two general classes of chirality measures have been recognized: in the first, the degree of chirality expresses the extent to which a chiral object differs from an achiral reference object, while in the second it expresses the extent to which two enantiomorphs differ from each other [Buda, Auf der Heyde *et al.*, 1992].

Chirality measures of the first class are the **Ruch's chirality functions** [Ruch, 1972], according to which a chiral molecule is represented by an achiral skeleton with attached four “ligands”, for example, chemical groups a , b , c , and d , each of them characterized by a specific parameter λ . Polynomial functions, such as

$$F = (\lambda_a - \lambda_b) \cdot (\lambda_a - \lambda_c) \cdot (\lambda_a - \lambda_d) \cdot (\lambda_b - \lambda_c) \cdot (\lambda_b - \lambda_d) \cdot (\lambda_c - \lambda_d)$$

transform these parameters into a chirality measure, being, for two enantiomers R and S, $F(R) = -F(S)$. A specific application of this approach was proposed by Lukovits and Linert [Lukovits and Linert, 2001] using the → *valence connectivity index* $^1\chi^v$ (→ *Chi chirality descriptor*) to describe chirality.

The **Hausdorff chirality measure** is a chirality measure of the second class [Buda and Mislow, 1992]. Let Q and Q' denote two enantiomorphous, nonempty, and bounded sets of points defined in the geometrical space (x, y, z) . Let $d(q, q')$ denotes the distance between two points: $q \in Q$ and $q' \in Q'$. Then, the Hausdorff distance h between sets Q and Q' is defined as

$$h(Q, Q') = h(Q', Q) = \max[\rho(Q, Q'), \rho(Q', Q)]$$

where ρ is defined as:

$$\rho(Q, Q') = \max_{q \in Q} \{ \min_{q' \in Q'} (d_{qq'}) \} \quad \rho(Q', Q) = \max_{q' \in Q'} \{ \min_{q \in Q} (d_{q'q}) \}$$

The Hausdorff distance $h(Q, Q')$ between Q and Q' corresponds to the smallest number $\delta = h(Q, Q')$ that has the following properties: (a) a spherical ball of radius δ centered at any point of Q contains at least one point of Q' and (b) a spherical ball of radius δ centered at any point of Q' contains at least one point of Q . It is obvious that $h(Q, Q') = 0$ only if $Q = Q'$.

This means that the Hausdorff distance between two sets of points, Q and Q' , representing geometric objects, can be zero only if these two objects are identical, that is, achiral mirror images.

The value of the Hausdorff distance between a geometric object Q and its mirror image Q' depends not only on the shape of these objects but also on their size and their relative

orientations in the geometrical space. By rotating and translating one enantiomorph with respect to the other, one can find the minimal value $h_{\min}(Q, Q')$ corresponding to the optimal overlap.

The Hausdorff chirality measure is finally defined as

$$H(Q) = \frac{h_{\min}(Q, Q')}{d_{\max}(Q)}$$

where $d_{\max}(Q)$ denotes the diameter of Q , that is, the largest distance between any two points of Q . The Hausdorff chirality measure does not depend on the size of Q and Q' and their relative position, and it can be easily shown that it has all the attributes required for a degree of chirality.

The interest in using chirality descriptors in QSAR/QSPR modeling is increasing and hence some chirality descriptors are explained below. $\rightarrow$ *Schultz weighted distance matrices* have been also devised for obtaining the $\rightarrow$ *chiral modification number* that is added to any topological index to discriminate cis/trans isomers. Other interesting methods to quantify chirality are the Kuz'min's method based on the *dissymmetry functions* [Kuz'min, Stel'makh *et al.*, 1992a, 1992b; Kutulya, Kuz'min *et al.*, 1992], the Mezey's method [Mezey, 1997b] based on the $\rightarrow$ *Mezey 3D shape analysis*, $\rightarrow$ *MARCH-INSIDE descriptors*, $\rightarrow$ *chiral TOMOCOMD descriptors*, and the spectral $\rightarrow$ *chiral A_{xi} indices* [Xu, Zhang *et al.*, 2006].

• Seri-Levy chirality coefficients

These are chirality descriptors based on the maximum overlapping of the van der Waals volumes of the two enantiomers R and S in 3D space, which are embedded into a gridded box [Seri-Levy, Salter *et al.*, 1994; Seri-Levy, West *et al.*, 1994]. A shape similarity coefficient S_{RS} is calculated as

$$S_{RS} = \frac{n_{RS}}{\sqrt{n_R \cdot n_S}}$$

where n_{RS} is the number of grid points falling inside both enantiomers, n_R and n_S are the number of grid points included within the volume of enantiomers R and S , respectively.

From the shape similarity coefficient, a shape chirality coefficient was defined as:

$$S'_{RS} = 1 - S_{RS}$$

• Continuous Chirality Measure (CCM)

The continuous chirality measure is an example of first class chirality measure based on the general definition of *continuous symmetry measure*; it is defined as [Zabrodsky and Avnir, 1995]

$$S(G) = \frac{100}{A} \cdot \sum_{i=1}^A \|\mathbf{p}_i - \hat{\mathbf{p}}_i\|^2 \quad 0 < S \leq 100$$

where G is a given symmetry group, $\mathbf{p}_i$ the coordinate vector of the i th atom of the original chiral configuration, $\hat{\mathbf{p}}_i$ the coordinate vector of the corresponding atom in the nearest G -symmetric configuration, and A the number of atoms of the molecule. In practice, since the minimal requirement for an object to be achiral is that it possesses either a reflection mirror (σ), an inversion center (i), or a higher order improper rotation axis (S_{2n}), the function $S(G)$ has to be screened over symmetry groups having these elements. The continuous chirality measure is the total (normalized) distance of the original chiral configuration from the considered G -symmetry configuration, bounded between 0 and 100.

• Randić chirality index

This is a topological index that was proposed to discriminate between a chiral molecule and its mirror image; it is restricted to molecules embedded in 2D space, such as benzenoids, and is based on $\rightarrow$ *periphery codes* [Randić, 1998a, 2001a].

Starting from a selected i th atom on the molecule periphery, the $\rightarrow$ *vertex degrees* δ of the periphery atoms define an ordered code for the molecule; two different codes are obtained whether the clockwise or the anticlockwise direction is chosen. These codes referring to the i th atom are then transformed by making partial sums, adding successively elements of the series; the two obtained codes encode information on the asymmetry of the molecular periphery.

To obtain a single value descriptor D_i for the i th atom, the differences between the corresponding elements in the anticlockwise and clockwise codes are computed and then added:

$$D_i = \sum_{j=1}^A (AD_{ij} - CD_{ij})$$

where AD_{ij} represents the j th element in the i th anticlockwise code and CD_{ij} represents the j th element in the i th clockwise code. This procedure is repeated for all the atoms on the molecule periphery. Finally, the Randić chirality index is defined as

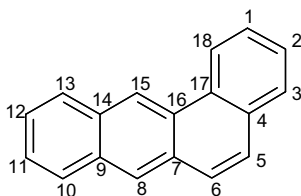
$$CH_R^k = \frac{1}{A^k} \cdot \sum_{i=1}^A D_i^k$$

where k is an odd power exponent and A is the number of atoms. Using different odd powers, a sequence of chirality indices can be calculated to better characterize the molecule and its mirror image.

For achiral molecules, all the chirality indices (and the corresponding vector elements) equal zero.

Example C8

Randić chirality index for benzoanthracene.



clockwise direction

labels	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
atom 1	2	2	2	3	2	2	3	2	3	2	2	2	2	3	2	3	3	2

anticlockwise direction																		
labels	1	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
atom 1	2	2	3	3	2	3	2	2	2	2	3	2	3	2	2	3	2	2

↓

clockwise direction CD_{1j}

anticlockwise direction AD_{1j}																		
atom 1	2	4	6	9	11	13	16	18	21	23	25	27	29	32	34	37	40	42

$$D_1 = \sum_{j=1}^{18} (AD_j - CD_j) = +14$$

↓ repeating the procedure for all the atoms

Randic chirality index CH_R^k					
k	3	5	7	9	11
benzanthracene	+ 2.17284	+ 5.07240	+ 10.97954	+ 23.70935	+ 51.05061
mirror image	- 2.17284	- 5.07240	- 10.97954	- 23.70935	- 51.05061

• Moreau chirality index

This is a molecular descriptor proposed with the aim of quantifying the chirality of a molecule by means of the positive or negative quantification of the chirality of the environment of the atoms in the molecule for any scalar atomic property [Moreau, 1997]. The basic idea is that unsymmetrical environments are not the privilege of atoms in chiral molecules, however they are the most frequent situations. The Moreau chirality index is defined as

$$CH_M = \sum_{i=1}^A AS_i$$

where the summation runs over all atoms in the molecule and AS_i is the measure of the chirality of the environment of the i th atom given by the following expression:

$$AS_i = 10^3 \cdot C_i \cdot \{XYZ\}_i \cdot S_i = 10^3 \cdot \frac{(\lambda_1 - \lambda_2) \cdot (\lambda_2 - \lambda_3) \cdot \lambda_3}{(\sum_k \lambda_k)^3} \cdot \frac{X_i \cdot Y_i \cdot Z_i}{(e_X \cdot e_Y \cdot e_Z)} \cdot (v_1, v_2, v_3)$$

where λ_1 , λ_2 , and λ_3 are the eigenvalues of the square symmetric matrix $\mathbf{C}$ closely related to the covariance matrix of the Cartesian coordinates (x_j, y_j, z_j) of the atoms in the environment of the i th atom and defined as

$$\mathbf{C} = \begin{vmatrix} \sum_j p_j \cdot w_j \cdot x_j^2 & \sum_j p_j \cdot w_j \cdot x_j \cdot y_j & \sum_j p_j \cdot w_j \cdot x_j \cdot z_j \\ \sum_j p_j \cdot w_j \cdot x_j \cdot y_j & \sum_j p_j \cdot w_j \cdot y_j^2 & \sum_j p_j \cdot w_j \cdot y_j \cdot z_j \\ \sum_j p_j \cdot w_j \cdot x_j \cdot z_j & \sum_j p_j \cdot w_j \cdot y_j \cdot z_j & \sum_j p_j \cdot w_j \cdot z_j^2 \end{vmatrix}$$

where summations are over all atoms in the environment of i and p_j and w_j are parameters defined below.

The origin of the coordinates is the barycenter of the environment of the considered atom and the eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ associated with the eigenvalues are vectors that define the three principal axes of this environment. The considered i th atom can be included or not in its environment. Each atom of the environment is assigned an atomic property p_j (e.g., unitary property, atomic mass, atomic electronegativity, and atomic van der Waals volume) and a weight w_j , which is a function of the distance of the j th environment atom from the i th atom.

In this approach, the principal planes of the environment are taken as the best approximation of the potential symmetry planes and the asymmetry of the environment as seen from the considered i th atom is defined as proportional to the distance from i to the symmetry plane.

The coefficient C_i in the expression of AS_i accounts for the asymmetry of the environment and is equal to zero when two eigenvalues are equal, or the third eigenvalue equals zero, or the last two eigenvalues equal zero.

The term $\{XYZ\}_i$ is the product of the coordinates of the i th atom with respect to the principal axes normalized by the product of the half-thicknesses e_x , e_y , e_z of the “slab” which approximates the set of weighted environment atoms. This term indicates that asymmetry is positive or negative according to the octant in which the i th atom is.

The term S is the box-product of the three eigenvectors and is equal to $+1$ or -1 depending on the handedness of the principal axis space. It was introduced in the expression of AS to have an intrinsic measure of chirality.

• topological chirality descriptors

The idea of modifying graph invariants to make them chirality sensitive was proposed by Schultz *et al.* in 1992 [Schultz, Schultz *et al.*, 1995], introducing a $\rightarrow$ *chiral factor* equal to $+1$ or -1 for any atom in R- or S-configuration, respectively, and assigning a value of 0 to all the other atoms.

Developing this idea, several series of topological chirality descriptors were introduced by using a **chirality correction factor**, denoted as c , applied to the $\rightarrow$ *vertex degree* of chiral atoms in a $\rightarrow$ *H-depleted molecular graph* [Golbraikh, Bonchev *et al.*, 2001a, 2001b; Golbraikh and Tropsha, 2003]. These descriptors include modified $\rightarrow$ *Zagreb indices*, $\rightarrow$ *connectivity indices*, $\rightarrow$ *extended connectivity indices*, $\rightarrow$ *overall connectivity indices*, and $\rightarrow$ *topological charge indices*.

For each asymmetric atom in R-configuration, the vertex degree δ_i is substituted with $(\delta_i + c)$ and for each atom in S-configuration with $(\delta_i - c)$. This transformation is equivalent to making main diagonal elements a_{ii} of the $\rightarrow$ *adjacency matrix* $\mathbf{A}$ equal to $+c$ or $-c$ for all chiral atoms in R- or S-configuration, respectively. For achiral atom, the chirality correction factor equals zero.

According to this approach, the $\rightarrow$ Randić connectivity index can be transformed into the corresponding chirality index as

$${}^1\chi = \sum_b (\delta_i \cdot \delta_j)_b^{-1/2} \Rightarrow {}^1\chi^{chir} = \sum_b [(\delta_i \pm c) \cdot (\delta_j \pm c)]_b^{-1/2}$$

where summation goes over all edges in the molecular graph and i and j refer to the vertices, which are connected by an edge.

In general, values of $|c| < \delta_i$ were assumed. The chirality correction was also applied to $\rightarrow$ valence vertex degrees from which valence connectivity indices are derived.

Chirality correction can be a real number (chirality descriptors of class I) or an imaginary number (chirality descriptors of class II). In the latter case, chirality descriptors are complex numbers.

The **chiral connectivity indices** were calculated by analogy with the $\rightarrow$ Kier–Hall connectivity indices by using vertex degrees modified by a chirality correction $c = \pm 1$ [Xu, Zhang *et al.*, 2006].

• Chi chirality descriptor (χ^c)

This chirality descriptor is derived from the Ruch's chirality functions applied to the first-order $\rightarrow$ valence connectivity index ${}^1\chi^v$. Separate values of the valence connectivity index are calculated for the four atoms/substituents a , b , c , and d bonded to the chiral atom [Lukovits and Linert, 2001]. The chirality correction χ^c is calculated by the following function F:

$$F \equiv \chi^c = ({}^1\chi_a^v - {}^1\chi_b^v) \cdot ({}^1\chi_a^v - {}^1\chi_c^v) \cdot ({}^1\chi_a^v - {}^1\chi_d^v) \cdot ({}^1\chi_b^v - {}^1\chi_c^v) \cdot ({}^1\chi_b^v - {}^1\chi_d^v) \cdot ({}^1\chi_c^v - {}^1\chi_d^v)$$

that, for two enantiomers R and S, has the following property: $F(R) = -F(S)$.

The chirality descriptor χ^- is then defined as

$$\chi^- = {}^1\chi^v \pm \chi^c$$

where ${}^1\chi^v$ is the valence connectivity index for the whole molecule and χ^c the chirality correction.

A drawback of this index is that index ${}^1\chi^v$ is often degenerate leading to the same value for different substituent groups (e.g., halogens), resulting into a zero correction for chirality. This drawback may be overcome using more discriminant vertex degrees, such as $\rightarrow$ perturbation delta values, $\rightarrow$ Yang vertex degree, or the δ^{CT} proposed by the Authors.

${}^1\chi^v$ values for some substituents are reported in $\rightarrow$ connectivity descriptors (Table C6).

• Chirality Codes (f_{CICC} , f_{CDCC})

The conformational-independent chirality code f_{CICC} is a modification of the $\rightarrow$ radial distribution function code $g(R)$ to account for chirality:

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot e^{-\beta \cdot (R-r_{ij})^2} \Rightarrow f_{CICC}(R) = \sum_{i=1}^A \sum_{j=1}^{A-1} \sum_{k=1}^{A-2} \sum_{l=1}^{A-3} S_{ijkl} \cdot e^{-\beta \cdot (R-E_{ijkl})^2}$$

where β is a smoothing factor and A the number of atoms [Aires-de-Sousa and Gasteiger, 2001, 2002; Aires-de-Sousa, 2003]. The term E_{ijkl} was introduced to account for the stereochemical situation of the chiral center; this term considers atoms i , j , k , and l , each of them belonging to a different neighborhood of the four atoms A, B, C, and D that are directly bonded to the chiral

center. This term is defined as

$$E_{ijkl} = \frac{w_i \cdot w_j}{r_{ij}} + \frac{w_i \cdot w_k}{r_{ik}} + \frac{w_i \cdot w_l}{r_{il}} + \frac{w_j \cdot w_k}{r_{jk}} + \frac{w_j \cdot w_l}{r_{jl}} + \frac{w_k \cdot w_l}{r_{kl}}$$

where w is an atomic property and r the distance calculated as the sum of the geometric distances along the shortest path joining two atoms. Further, the chirality signal S_{ijkl} can attain values $+1$ or -1 . For the computation of S_{ijkl} , atoms i, j, k , and l are ranked according to the decreasing atomic property w . When the property of two atoms is the same, the properties of the neighbors (A, B, C, or D) are used for ranking. The (x, y, z) coordinates of A are then used for atom i , those of B for j , those of C for k , and those of D for l . The first three atoms, in the order defined by the ranking define a plane. If they are ordered clockwise and the fourth atom is behind the plane, the chirality signal is set at $+1$; if the geometric arrangement is opposite, $S_{ijkl} = -1$.

The two values, E and S , calculated for all the combinations of the four atoms are then combined to generate the chirality code $f_{\text{CICC}}(R)$, where the function is calculated at a number of discrete points R with defined intervals to obtain the same number of descriptors, irrespective of the size of the molecule. The actual range of R is chosen according to the range of the studied properties related to the range of observed interatomic distances for the data set molecules.

The number of discrete points determines the resolution of the chirality code.

Moreover, the **Conformational-Dependent Chirality Code** (f_{CDCC}) was defined to account for conformational behavior of molecules, replacing the chirality signal S_{ijkl} by a conformational-dependent geometric parameter C_{ijkl} [Caetano, Aires-de-Sousa *et al.*, 2005]. The f_{CDCC} code is calculated as

$$f_{\text{CDCC}}(R) = \sum_{i=1}^A \sum_{j=1}^{A-1} \sum_{k=1}^{A-2} \sum_{l=1}^{A-3} C_{ijkl} \cdot e^{-\beta \cdot (R - E_{ijkl})^2}$$

where A is the number atoms. The parameter C_{ijkl} takes real values and is defined according to the expression:

$$C_{ijkl} = \frac{x_j \cdot y_k \cdot z_l}{x_j \cdot y_k + x_j \cdot |z_l| + y_k \cdot |z_l|}$$

that is, C_{ijkl} is defined for the combination of four atoms i, j, k , and l and x, y and z are the atomic Cartesian coordinates. The Cartesian coordinates are defined in such a way that atom i is at position $(0, 0, 0)$, atom j lies on the positive side of the x -axis, and atom k on the xy plane and having a positive y coordinate. Therefore, C_{ijkl} will have opposite values for enantiomers, because C_{ijkl} will have either a positive or negative value depending on whether atom l is above ($z_l > 0$) or below ($z_l < 0$) the plane formed by atoms i, j , and k .

• Ursu–Diudea chirality (χ_{1234})

The chirality of an ordered quadruple of atoms numbered 1,2,3,4 is measured in terms of their (x, y, z) Cartesian coordinates, adopting some geometrical constraints, by the sign of the following determinant [Ursu and Diudea, 2005; Ursu, Diudea *et al.*, 2006]:

$$\chi_{1234} = \text{sgn} \left(\det \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{pmatrix} \right)$$

• **Relative Chirality Index (^VRCI)**

The Relative Chirality Index is calculated from a series expansion of the $\rightarrow$ *valence vertex degree*, calculated for the four atoms/groups attached to the chiral carbon, where atom/groups priorities (a, b, c, d) are given according to the Cahn–Ingold–Prelog rule [Natarajan, Basak *et al.*, 2007] (Figure C2).

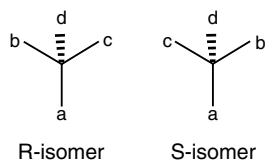


Figure C2 R- and S-configurations.

To calculate the relative chirality index, the least important chemical group (d) is placed at the rear, and the clockwise and anticlockwise arrangement of the other three groups (a, b, c) are used to represent the R- and S-configurations, respectively.

The groups/atoms a, b, c , and d are then assigned $\rightarrow$ *valence vertex degrees* (δ^v). When the group has more than one atom, δ^v for the group a, b , or c is calculated considering the relative proximities of the atoms to the chiral neighbor, and decreasing importance with increasing topological distance was assigned while calculating the contribution of atoms other than hydrogen in a group. The group delta value for any group attached to a chiral carbon is, then, calculated as

$$\delta_i^v = \delta_{n1}^v + \frac{\delta_{n2}^v}{2} + \frac{\delta_{n3}^v}{4} + \frac{\delta_{n4}^v}{8} + \dots$$

where $n1$ is the atom attached directly to the chiral center (nearest neighbor), $n2$ is separated by one atom, $n3$ by two atoms, etc. Relative chirality indices (^VRCI) for a pair of enantiomers (R and S) are calculated as:

$$^V\text{RCI}_R = \delta_a^v \cdot [1 + (1 + \delta_b^v) + (1 + \delta_b^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

$$^V\text{RCI}_S = \delta_a^v \cdot [1 + (1 + \delta_c^v) + (1 + \delta_c^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

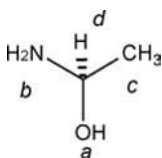
To obtain ^VRCI for molecules containing more than one chiral center, root-mean-square of ^VRCI for all the chiral atoms is calculated as

$$^V\text{RCI} = \sqrt{\frac{1}{n_{\text{chi}}} \sum_{i=1}^{n_{\text{chi}}} (^V\text{RCI}_i)^2}$$

where n_{chi} is the number of chiral centers.

Example C9

Relative chirality indices and chirality correction factors for the two enantiomers shown below, together with valence vertex degrees and first-order valence connectivity index of the functional groups.



$$\delta^v(\text{OH}) = 5 \quad {}^1\chi^v(\text{OH}) = 0.25820$$

$$\delta^v(\text{NH}_2) = 3 \quad {}^1\chi^v(\text{NH}_2) = 0.33333$$

$$\delta^v(\text{CH}_3) = 1 \quad {}^1\chi^v(\text{CH}_3) = 0.57735$$

$$\delta^v(\text{H}) = 0 \quad {}^1\chi^v(\text{H}) = 0$$

$${}^v\text{RCI}_R = \delta_a^v \cdot [1 + (1 + \delta_b^v) + (1 + \delta_b^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

$$= 5 \cdot [1 + (1 + 3) + (1 + 3 + 3 \cdot 1) + 0] = 60$$

$${}^v\text{RCI}_S = \delta_a^v \cdot [1 + (1 + \delta_c^v) + (1 + \delta_c^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

$$= 5[1 + (1 + 1) + (1 + 1 + 1 \cdot 3) + 0] = 40$$

The chirality correction factor χ^c for the calculation of the $\rightarrow$ *Chi chirality index* is 0.00029.

📖 [King, 1991; Gilat, 1994; Liang and Mislow, 1994; Flapan, 1995; Franke, Rose *et al.*, 1995; Winberg and Mislow, 1995; De Julián-Ortiz, García-Domenech *et al.*, 1996; Fujita, 1996; Petitjean, 1996; Randić and Mezey, 1996; Randić and Razinger, 1996; Balaban, 1997b; Gutman and Pyka, 1997; Klein and Babic, 1997; Mislow, 1997; De Julián-Ortiz, de Gregorio Alapont *et al.*, 1998; Keinan and Avnir, 1998; Nemba and Balaban, 1998; Randić, 1998a; Golbraikh, Bonchev *et al.*, 2002; González Díaz, Sánchez *et al.*, 2003; Wildman and Crippen, 2003; Kovatcheva, Golbraikh *et al.*, 2004, 2005; Marrero-Ponce, González Díaz *et al.*, 2004; Marrero-Ponce and Castillo-Garit, 2005]

- **chiral modification number** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **chiral TOMOCOMD descriptors** $\rightarrow$ TOMOCOMD descriptors
- **CHI-square statistics** $\rightarrow$ statistical indices ($\odot$ concentration indices)
- **chord distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)

■ chromatic decomposition

A decomposition $\mathcal{A}(V_1, V_2, \dots, V_G)$ of the set V of the graph G vertices into $G \rightarrow$ *equivalence classes* is said **chromatic decomposition** of G (or **vertex chromatic decomposition**) if, for any pair of vertices v_i and v_j belonging to V_g , the edge e_{ij} connecting the considered vertices does not belong to the set of edges $\mathcal{E}$ of the graph; it means that two vertices belonging to the same chromatic class V_g cannot be adjacent [Bonchev, 1983].

A decomposition $\mathcal{B}(\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_G)$ of the set $\mathcal{E}$ of the graph G edges into G equivalence classes is said **edge chromatic decomposition** of G if any pair of edges e_{ij} and e_{kl} belonging to $\mathcal{E}_g$ does not belong to the set ${}^2\mathcal{P}$ of the second-order paths of the graph (i.e., the two edges are not adjacent).

A **graph coloring** of vertices (or edges) is an assignment of a minimal number of different colors to the vertices (or edges) of G such that no two adjacent vertices (or edges) have the same color. Graph coloring produces a $\rightarrow$ *chromatic graph*.

The subsets V_g are called **color classes**. The simplest descriptor that can be defined by a vertex chromatic decomposition is called **chromatic number** $k(G)$ (or **vertex chromatic number**, $v_k(G)$) and is the smallest number of color equivalence classes (i.e., G). In general, there is not a unique chromatic decomposition of a graph with the smallest number of colors. Analogously, the descriptor obtained by an edge chromatic decomposition is called **edge chromatic number**, denoted as $^E k(G)$.

The **chromatic information index** (or **vertex chromatic information index**) [Mowshowitz, 1968d] is the minimum value of the $\rightarrow$ *mean information content* of all possible vertex chromatic decompositions with a number of colors equal to the vertex chromatic number $k(G)$ and is defined as

$$^V \bar{I}_{CHR} = \min \left(- \sum_{g=1}^{k(G)} \frac{|V_g|}{A} \log_2 \frac{|V_g|}{A} \right)$$

where $|V_g|$ is the number of vertices (i.e., the cardinality of g th set) within the same equivalence class for the decomposition and A is the number of graph vertices.

The **edge chromatic information index** is the minimum value of the mean information content of all possible edge chromatic decompositions having a number of colors equal to the edge chromatic number $^E k(G)$ and is defined as

$$^E \bar{I}_{CHR} = \min \left(- \sum_{g=1}^{^E k(G)} \frac{|E_g|}{B} \log_2 \frac{|E_g|}{B} \right)$$

where $|E_g|$ is the number of edges within the same equivalence class for the decomposition and B is the number of graph edges.

- **chromatic graph** $\rightarrow$ graph
- **chromatic information index** $\rightarrow$ chromatic decomposition
- **chromatic number** $\rightarrow$ chromatic decomposition

■ chromatographic descriptors

These are experimental quantities derived from chromatographic techniques, that is, from gas chromatography (GC), high-performance liquid chromatography (HPLC), thin-layer chromatography (TLC), and paper chromatography (PC) [Kaliszan, 1987, 1992] or structural indices used to predict experimental chromatographic parameters from molecular structure.

The most important ones are listed below.

• retention time (t_R)

It is the characteristic time it takes to a compound to pass through a chromatographic system (e.g., from the column inlet to the detector) under fixed conditions. The **adjusted retention time** of a compound, denoted by t'_R , is the difference between the total retention time t_R and the retention time of an unretained compound t_M ; the **relative retention time**, denoted as RRT , is the ratio of the adjusted retention time of a compound over that of a reference compound.

Based on the retention times in HPLC systems, a **chromatographic hydrophobicity index (CHI)** was defined aimed at correlating chromatographic retention times with lipophilicity. This is defined as [Valkó, Bevan *et al.*, 1997; Valkó, Plass *et al.*, 1998]

$$CHI = a \cdot t_R + b$$

where the parameters a and b are dependent on the flow rate, column length, gradient time, column, etc. and are experimentally determined; they allow to transfer different chromatographic measurements into a unique scale. Another index derived from HPLC is the **ϕ_0 index**, which is defined as the percentage (by volume) of acetonitrile required to achieve an equal distribution of a compound between the mobile and stationary phases. For most compounds this is a physically attainable volume percent of organic phase with a value between 0 and 100%. By plotting the $\log k'$ values, k' being the $\rightarrow$ *capacity factor*, as a function of the organic solvent concentration, the ϕ_0 value can be obtained from the slope and the intercept of the straight interpolation line as $\phi_0 = -b/a$, where a and b are the slope and intercept of the straight line, respectively.

- **capacity factor (k')**

Also called **phase capacity ratio** or **retention factor**, it is a measure of the degree of retention of a compound in a chromatographic column, as

$$k' = \frac{t_R - t_M}{t_M} = \frac{V_R - V_M}{V_M}$$

where t_R and V_R are the retention time and retention volume, respectively, of the compound; t_M and V_M , named dead time and dead volume, are respectively the retention time and the retention volume of an unretained compound. The quantity $\log k'$ can be considered analogous to the $\rightarrow$ *Bate-Smith-Westall retention index* R_M (see below) and, like this, is related to $\rightarrow$ *partition coefficients*.

In micellar electrokinetic chromatography and microemulsion electrokinetic chromatography, the retention factor k' of a neutral compound is defined as [Muijselaar, Claessens *et al.*, 1994]

$$k' = \frac{t_R - t_M}{t_M} \left(1 - \frac{t_R}{t_m} \right)^{-1}$$

where t_M , t_R , and t_m are the migration times of electroosmotic flow, compound, and microemulsion, respectively.

From this retention factor, the **Migration Index (MI)** for a compound in microemulsion electrokinetic chromatography was defined as [Ishihama, Oda *et al.*, 1996]

$$MI = a \cdot \log \left(\frac{\mu_{aq} - \mu_{eff}}{\mu_{eff} - \mu_{me}} \right) + b = a \cdot \log k' + b$$

where μ_{aq} and μ_{me} are the electrophoretic mobilities of the compound in the aqueous phase and microemulsion phase, respectively, μ_{eff} the effective mobility in the microemulsion solution, and a and b the slope and the intercept of a calibration line relative to $\log k'$ values of reference solutes such as alkyl benzene and their migration indices.

The Migration Index scale can be applied to all neutral compounds that migrate in the range t_M and t_m and this might be independent of the volume of the microemulsion. The Migration

Index scale can be used as a measure of hydrophobicity of compounds for QSAR/QSPR modeling studies [Ishihama, Oda *et al.*, 1996; Fatemi, 2003].

- **Kovats retention index (I_i)**

This is an index characteristic of a gas-chromatographed compound on a given column at a definite temperature defined as

$$I_i = 100 \cdot \frac{\log t'_{R_i} - \log t'_{R(N_C)}}{\log t'_{R(N_C+1)} - \log t'_{R(N_C)}} + 100 \cdot N_C$$

where $t'_{R(N_C)}$ is the $\rightarrow$ *adjusted retention time* of a homologue standard with a number of carbon atoms equal to N_C ; $t'_{R(N_C+1)}$ an analogous parameter for another standard with carbon number $N_C + 1$; t'_{R_i} the adjusted retention time of the i th compound [Kováts, 1968]. The measured total retention time t_R is a sum of two factors $t_R = t_M + t'_R$, where t_M is the initial dead time and t'_R is the adjusted retention time of the compound.

- **Bate-Smith-Westall retention index (R_M)**

This is a retention index derived from thin-layer and paper chromatography defined as [Bate-Smith and Westall, 1950]

$$R_M = \log \left(\frac{1}{R_f} - 1 \right)$$

where R_f is the **retardation factor**, which is the ratio of the migration distance of the compound over that of the solvent front.

The quantity R_M is proportional to the partition coefficient $\rightarrow \log P$ and has been used in its place in many QSAR models.

- **semiempirical topological index (I_{ET})**

The semiempirical topological index (the name is not the best choice) was designed to predict the chromatographic $\rightarrow$ *Kovats retention index* of organic molecules, based on experimental chromatographic measurements [Heinzen, Soares *et al.*, 1999; da Silva Junkes, Amboni *et al.*, 2003b].

This is derived from the $\rightarrow$ *H-depleted molecular graph* of a molecule as

$$I_{ET} = \sum_{i=1}^A (C_i + \delta_i)$$

where A is the number of graph vertices and C_i is the carbon atom contribution of the i th molecular fragment (Table C4); δ_i is the contribution of the carbon atoms bonded to the i th carbon atom, calculated as

$$\delta_i = \sum_{j=1}^A a_{ij} \cdot \log(C_j)$$

where the summation goes over all graph vertices but only contributions from vertices adjacent to the i th vertex are accounted for, a_{ij} being the elements of the $\rightarrow$ *adjacency matrix* A .

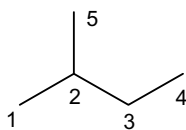
In this approach, the molecular fragment including the functional group and the carbon atom directly attached to the functional group are considered as a single vertex in the molecular graph.

Table C4 Values of the fragmental constant C_i for carbon atoms and functional groups of esters, aldehydes, ketones, and alcohols.

Chemical class	Fragment	Fragment position	C_i
Linear and branched alkanes	$-\text{CH}_3$		1.0
	$-\text{CH}_2-$		0.9
	$-\text{CH}<$		0.8
	$>\text{C}<$		0.7
Alkenes	$=\text{CH}_2; =\text{CH}-$	1C	0.8975
	$=\text{CH}-\text{trans}$	2C	0.895
	$=\text{CH}-\text{cis}$	2C	0.910
	$=\text{CH}-\text{trans}$	3C	0.875
	$=\text{CH}-\text{cis}$	3C	0.885
	$=\text{CH}-\text{trans}$	4C	0.865
	$=\text{CH}-\text{cis}$	4C	0.870
	$=\text{CH}-\text{trans}$	5C	0.865
	$=\text{CH}-\text{cis}$	5C	0.855
	$=\text{CH}-\text{trans}$	6C	0.860
	$=\text{CH}-\text{cis}$	6C	0.850
	$=\text{CH}-\text{trans}$	7C	0.8575
	$=\text{CH}-\text{cis}$	7C	0.845
Alcohols	$-\text{CH}_2-\text{OH}$		2.63
	$>\text{CH}-\text{OH}$	2nd position	1.79
	$>\text{CH}-\text{OH}$	3rd position	1.78
	$>\text{CH}-\text{OH}$	Middle	1.68
	$-\text{CH}<$	α OH	0.75
	$-\text{CH}<$	β OH	0.73
	$>\text{C}<$	α OH	0.61
	$>\text{C}<$	β OH	0.63
Aldehydes and ketones	$\text{HC}=\text{O}$	Aldehyde	2.094
	$\text{C}=\text{O}$	2nd position	1.71
	$\text{C}=\text{O}$	3rd position	1.69
	$\text{C}=\text{O}$	Middle	1.60
	$-\text{CH}<$	α $\text{C}=\text{O}$	0.73
	$-\text{CH}<$	β $\text{C}=\text{O}$	0.70
	$-\text{CH}<$	γ $\text{C}=\text{O}$	0.765
	$>\text{C}<$	α or β $\text{C}=\text{O}$	0.61

Example C10

Semiempirical topological index for 2-methylbutane.



$$\delta_1 = \log 0.8 = -0.0969$$

$$\delta_2 = \log 0.9 + \log 1.0 + \log 1.0 = -0.0458$$

$$\delta_3 = \log 0.8 + \log 1.0 = -0.0969$$

$$\delta_4 = \log 0.9 = -0.0458$$

$$\delta_5 = \log 0.8 = -0.0969$$

$$\begin{aligned} I_{\text{ET}} &= (1.0 - 0.0969) + (0.8 - 0.0458) + (0.9 - 0.0969) + (1.0 - 0.0458) + (1.0 - 0.0969) \\ &= 4.3177 \end{aligned}$$

📖 [Amboni, da Silva Junkes *et al.*, 2002b, 2002a; da Silva Junkes, Amboni *et al.*, 2002, 2003a, 2004, 2007; da Silva Junkes, Silva Arruda *et al.*, 2005]

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **chromatographic hydrophobicity index** → chromatographic descriptors (⊙ retention time)
- **CID'/CID index** → bond order indices (⊙ graphical bond order)
- **CIM index** ≡ *Chemically Intuitive Molecular index* → spectral indices (⊙ Burden eigenvalues)
- **circuit** ≡ *cyclic path* → graph
- **circular substructure descriptors** → substructure descriptors
- **CIRD indices** → distance matrix
- **CIRS indices** → distance matrix
- **CIRS' indices** → distance matrix
- **cis/trans binary factor** → *cis/trans* descriptors

■ cis/trans descriptors

cis/trans isomerism is usually easily distinguished by using → *geometrical descriptors*, that is, descriptors derived from 3D molecular structures or structures embedded in a 3D space [Randić, Jerman-Blazic *et al.*, 1990]. Otherwise, the simplest way to distinguish cis/trans isomers is the **cis/trans binary factor**, which takes value −1 for *cis*-isomers and +1 for *trans*-isomers [Lekishvili, 1997].

However, when molecular descriptors are derived from molecular graphs, cis/trans isomerism is not usually recognized and some molecular descriptors were proposed to discriminate between cis/trans isomers, such as the → *corrected electron charge density connectivity index*, and → *periphery codes*. → *Weighted matrices* were also devised for obtaining the → *geometric modification number* that is added to any topological index to discriminate cis/trans isomers.

Another topological descriptor specifically proposed for cis/trans isomerism is the **Pogliani cis/trans connectivity index** χ_{CT} defined in terms of the → *Randić connectivity index* χ as

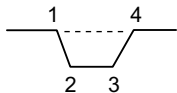
$$\chi_{\text{CT}} = \chi - \chi_{\text{CIS}} = \chi - \sum_k (\delta_1^r \cdot \delta_2 \cdot \delta_3 \cdot \delta_4^r)_k^{-1.5}$$

where the summation runs over all the cis-butadienic or cis-2-butenic fragments in the graph; δ is the → *vertex degree* and δ^r is the raised vertex degree obtained by joining the two *cis* vertices by a

virtual bond, forming a virtual four-membered ring [Pogliani, 1994b] (Example C11). For all-trans molecular graphs, $\chi_{CT} = \chi$.

Example C11

Calculation of the cis correction factor for the Pogliani cis/trans connectivity index.



$$\delta_1 = 2 + 1; \quad \delta_2 = 2; \quad \delta_3 = 2; \quad \delta_4 = 2 + 1$$

$$\chi_{CIS} = 0.046$$

📖 [Balaban, 1976b, 1998]

- **city-block distance** $\equiv$ *Manhattan distance* $\rightarrow$ similarity/diversity (☉ Table S7)
- **Ciubotariu shape indices** $\rightarrow$ shape descriptors
- **Clark distance** $\rightarrow$ similarity/diversity (☉ Table S7)
- **classical QSAR** $\rightarrow$ structure/response correlations
- **classification** $\rightarrow$ chemometrics

■ classification parameters

Statistical indices used to evaluate the performance of classification models [Frank and Todeschini, 1994]. They are derived from two kinds of statistics, called $\rightarrow$ *goodness of fit* and $\rightarrow$ *goodness of prediction*, thus distinguishing if the classification results are obtained as true predictions or not.

All the classification parameters can be derived from the **confusion matrix**, where the rows represent the known true classes and the columns the classes assigned by the classification method. It is a non-symmetric matrix of size $G \times G$, where G is the number of classes. For example, for a three-class problem ($G = 3$), the confusion matrix is:

		<i>assigned classes</i>			
<i>true classes</i>	<i>class</i>	A'	B'	C'	n_g
	A	c_{11}	c_{12}	c_{13}	n_a
	B	c_{21}	c_{22}	c_{23}	n_b
	C	c_{31}	c_{32}	c_{33}	n_c
	n'_g	n'_a	n'_b	n'_c	n

where A, B, and C represent labels for the true classes and A', B', and C' labels for the assigned classes. n_g represents the total number of objects effectively belonging to the g -th class and n'_g the total number of objects assigned by the classification model to the g -th class. The diagonal elements c_{gg} represent the correctly classified objects, while the off-diagonal elements c_{gk} represent the objects erroneously classified from class g to class k . Usually, two confusion matrices are obtained, one in the fitting and one after a validation procedure.

From the confusion matrix entries, the following parameters are defined:

Nonerror rate (NER), also called **overall accuracy** or simply *accuracy*, is the simplest measure of the quality of a classification model; usually expressed as a percentage, it is defined as

$$NER\% = \frac{\sum_g c_{gg}}{n} \times 100$$

where c_{gg} are the diagonal elements of the confusion matrix and n the total number of objects.

The complementary quantity is called **error rate (ER)** and is defined as

$$ER\% = \frac{n - \sum_g c_{gg}}{n} \times 100 = 100 - NER\%$$

To evaluate the efficiency of a classification model the error rate can be compared with the *no-model error rate (NOMER)*, that represents the error rate without a classification model and is calculated considering all the objects of smaller classes as erroneously classified in the largest class containing n_M objects:

$$NOMER\% = \frac{n - n_M}{n} \times 100$$

Another classification reference parameter is the *random classification error*, which is the error rate obtained if the objects are randomly assigned to the classes. It is defined as:

$$RER\% = \frac{1}{n} \cdot \left[\sum_{g=1}^G (n - n_g) \cdot p_g \right] \times 100$$

where n_g is the number of objects belonging to the g -th class, p_g is g -th class a-priori probability and n the total number of objects. $RER\%$ is equal to $NOMER\%$ if all the classes contain the same number of objects, i.e. $RER\% = NOMER\% = (1 - 1/G) \times 100$, which also corresponds to the error rate obtained by a complete random assignment.

Misclassification risk (MR). This is defined as:

$$MR\% = \sum_g \frac{(\sum_k L_{gk} \cdot c_{gk}) \cdot p_g}{n_g} \times 100$$

where p_g is the *prior class probability*, defined *a priori*, usually as $p_g = 1/G$ or $p_g = n_g/n$, where G is the total number of classes and n_g is the number of objects belonging to the g th class. L'_{gk} are elements of the *loss matrix* L , which is a user-defined nonsymmetric penalty matrix for classification errors, whose diagonal elements are zero, that is, no penalty is applied for correct classification, and the off-diagonal elements are the costs of the classification errors.

Sensitivity (Sn). A parameter that characterizes the ability of a classifier to correctly catch objects of the g th class, defined as

$$Sn_g = \frac{c_{gg}}{n_g} \times 100$$

Specificity (Sp). A parameter which characterizes the ability of the g -th class to reject objects of the other classes after the application of a classifier and defined as:

$$Sp_g = \left(1 - \frac{n'_g - c_{gg}}{n - n_g} \right) \times 100$$

where n'_g is the number of objects assigned to the g -th class.

Precision (Pr). It is a parameter which characterizes the purity of a class after the application of a classifier. It can be simply measured as the ratio of the number of objects assigned to the estimated g -th class and correctly classified (c_{gg}) over the total number of objects assigned to that class:

$$Pr_g = \frac{c_{gg}}{n'_g} \times 100$$

The degree of purity of a class can also be measured by the $\rightarrow$ *Shannon's entropy* and the $\rightarrow$ *Gini index*. In particular, using the Shannon's entropy, the **information gain in classification**, denoted as IG and usually expressed as percent, is calculated as:

$$IG\% = \frac{1}{H_0} \cdot \left[H_0 - \sum_{k=1}^G \frac{n'_k}{n} \cdot \sum_{g=1}^G \left(-\frac{n_{gk}}{n'_k} \cdot \log_2 \frac{n_{gk}}{n'_k} \right) \right] \times 100 = \frac{1}{H_0} \cdot \left[H_0 - \sum_{k=1}^G \frac{n'_k}{n} \cdot H'_k \right] \times 100$$

where n is the total number of objects, n'_k/n the proportion of objects in each final class (or in a node, for classification tree algorithms), n_{gk} the number of objects of the g th class present in the k th class (or node), H'_k the final entropy in each class (or node), and the summation over all the classes is the residual entropy [A-Razzak and Glen, 1992]. H_0 is the initial entropy, that is, the entropy before the classification:

$$H_0 = \sum_{g=1}^G \left[-\frac{n_g}{n} \cdot \log_2 \left(\frac{n_g}{n} \right) \right]$$

In the case of a perfect classification, the residual entropy, that is, the second term in the IG expression, is equal to zero and $IG\% = 100\%$.

Example C12

In a three-class problem ($G = 3$) of 30 objects ($n = 30$), after the application of a classification algorithm, the following confusion matrix is obtained:

Class	1'	2'	3'	n_g
1	9	1	0	10
2	2	8	2	12
3	1	2	5	8
n_k	12	11	7	30

The objects in the three classes are the following: $n_1 = 10$, $n_2 = 12$, and $n_3 = 8$. Moreover, a unitary loss matrix is assumed and the *a-priori* probabilities p_g of each class are assumed equal to $1/G$.

$$RER\% = \frac{\left[\left(\frac{30-10}{30} \right) \cdot 10 \right] + \left[\left(\frac{30-12}{30} \right) \cdot 12 \right] + \left[\left(\frac{30-8}{30} \right) \cdot 8 \right]}{30} \times 100 = 65.8\%$$

$$NOMER\% = \frac{30-12}{30} \times 100 = 60.0\%$$

$$NER\% = \frac{9+8+5}{30} \times 100 = 73.3\% \quad ER\% = 100 - 73.3 = 26.7\%$$

$$MR\% = \left[\frac{(0 \times 9 + 1 \times 1 + 1 \times 0) \cdot \frac{1}{3}}{10} + \frac{(1 \times 2 + 0 \times 8 + 1 \times 2) \cdot \frac{1}{3}}{12} + \frac{(1 \times 1 + 1 \times 2 + 0 \times 5) \cdot \frac{1}{3}}{8} \right] \times 100$$

$$= (0.033 + 0.111 + 0.125) \times 100 = 26.9\%$$

$$Sn_1 = \frac{9}{10} = 0.900 \quad Sn_2 = \frac{8}{12} = 0.667 \quad Sn_3 = \frac{5}{8} = 0.625$$

$$Sp_1 = \frac{17}{30-10} = 0.850 \quad Sp_2 = \frac{15}{30-12} = 0.833 \quad Sp_3 = \frac{20}{30-8} = 0.909$$

$$Pr_1 = \frac{9}{12} = 0.750 \quad Pr_2 = \frac{8}{11} = 0.727 \quad Pr_3 = \frac{5}{7} = 0.714$$

$$H_0 = -\frac{10}{30} \cdot \log_2 \left(\frac{10}{30} \right) - \frac{12}{30} \cdot \log_2 \left(\frac{12}{30} \right) - \frac{8}{30} \cdot \log_2 \left(\frac{8}{30} \right) = 1.566$$

$$H'_1 = -\frac{9}{12} \cdot \log_2 \left(\frac{9}{12} \right) - \frac{2}{12} \cdot \log_2 \left(\frac{2}{12} \right) - \frac{1}{12} \cdot \log_2 \left(\frac{1}{12} \right) = 1.041$$

$$H'_2 = -\frac{1}{11} \cdot \log_2 \left(\frac{1}{11} \right) - \frac{8}{11} \cdot \log_2 \left(\frac{8}{11} \right) - \frac{2}{11} \cdot \log_2 \left(\frac{2}{11} \right) = 1.096$$

$$H'_3 = -0 - \frac{2}{7} \cdot \log_2 \left(\frac{2}{7} \right) - \frac{5}{7} \cdot \log_2 \left(\frac{5}{7} \right) = 0.863$$

$$IG\% = \frac{1.566 - [1.041 \cdot \frac{12}{30} + 1.096 \cdot \frac{11}{30} + 0.863 \cdot \frac{7}{30}]}{1.566} \times 100 = 34.9\%$$

The misclassification risk calculated using *a-priori* probabilities defined by the relative frequencies of the classes, that is, $p_1 = 10/30$, $p_2 = 12/30$, and $p_3 = 8/30$, is $MR\% = 17.7\%$.

For two-class problems (the most common ones), classification parameters can be defined using $\rightarrow$ *binary distance measures*, based on the frequencies a , b , c , and d , which in this case may be interpreted as true positive (TP), false negative (FN), false positive (FP), and true negative (TN), respectively.

Let n be the number of objects, P the number of objects belonging to the class P and N the number of objects of the class N, the following frequency table can be constructed:

	Class P'	Class N'	
Class P	TP	FN	P
Class N	FP	TN	N
	P'	N'	n

where P' is the number of objects the classifier assigns to the class P and N' the number of objects assigned to the class N, and the following relationship holds $P + N = P' + N' = n$.

The nonerror rate is then defined as

$$NER\% = \frac{TP + TN}{TP + TN + FP + FN}$$

and the **Pearson coefficient** Φ (also called **Matthews correlation index**, *MCC*, [Matthews, 1975]), which is the most used global binary classification measure, is defined as

$$\Phi \equiv MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}}$$

(See also binary similarity coefficients in $\rightarrow$ *similarity/diversity*.)

Specific characteristics of binary classifications can also be highlighted by the following parameters:

$$\begin{aligned} Sn \equiv TPR &= \frac{TP}{TP + FN} & FNR &= \frac{FN}{TP + FN} = 1 - Sn & PPV &= \frac{TP}{TP + FP} \\ Sp \equiv TNR &= \frac{TN}{TN + FP} & FPR &= \frac{FP}{TN + FP} = 1 - Sp & NPV &= \frac{TN}{TN + FN} \end{aligned}$$

Sn being the sensitivity (or the **true positive rate**, *TPR* or **recall**), Sp the specificity (or the **true negative rate**, *TNR*), FNR the **false negative rate**, FPR the **false positive rate**, PPV the **positive predictive value** (or precision), and NPV the **negative predictive value**.

Moreover, derived from sensitivity and positive predictive value, the **F-measure** was also proposed as [Cannon, Amini *et al.*, 2007]

$$F\text{-measure} = \frac{2 \cdot Sn \cdot PPV}{Sn + PPV}$$

The **Receiver Operator Characteristic curve (ROC curve)** is a graphical plot of the sensitivity Sn versus false positive rate FPR for a binary classifier system as its discrimination threshold is varied. The ROC curve can also be represented equivalently by plotting the fraction of true positives (TP) versus the fraction of false positives (FP) (Figure C3). ROC analysis provides tools to select possibly optimal classification models.

For binary classification, **weighted classification accuracy (WCA)** was also defined as [Jensen, Refsgaard *et al.*, 2005]

$$WCA = \frac{4 \cdot Sn + 1 \cdot Sp}{8 \cdot FPR + 2 \cdot FNR + 2 \cdot NCR}$$

where *NCR* stands for the nonclassified rate, defined as

$$NCR = \frac{NCP + NCN}{TP + TN + FP + FN + NCP + NCN}$$

where *NCP* and *NCN* are the number of nonclassified objects belonging to the classes P and N, respectively. The coefficients of the *WCA* index were defined to deal with the specific classification purposes of the authors and may be modified depending on the problem; however, for a perfect classification result, *WCA* suffers from singularity. Then, a modified general form of this index [Authors, This book], ranging between 0 and 1 and similar to the $\rightarrow$ *Baroni–Urbani association index* defined for $\rightarrow$ *binary distances*, may be the following:

$$\text{mod-WCA} = \frac{a \cdot Sn + b \cdot Sp}{a \cdot Sn + b \cdot Sp + c \cdot FPR + d \cdot FNR + e \cdot NCR}$$

where a , b , c , d , and e are the weighting coefficients to be estimated or defined depending on each specific problem.

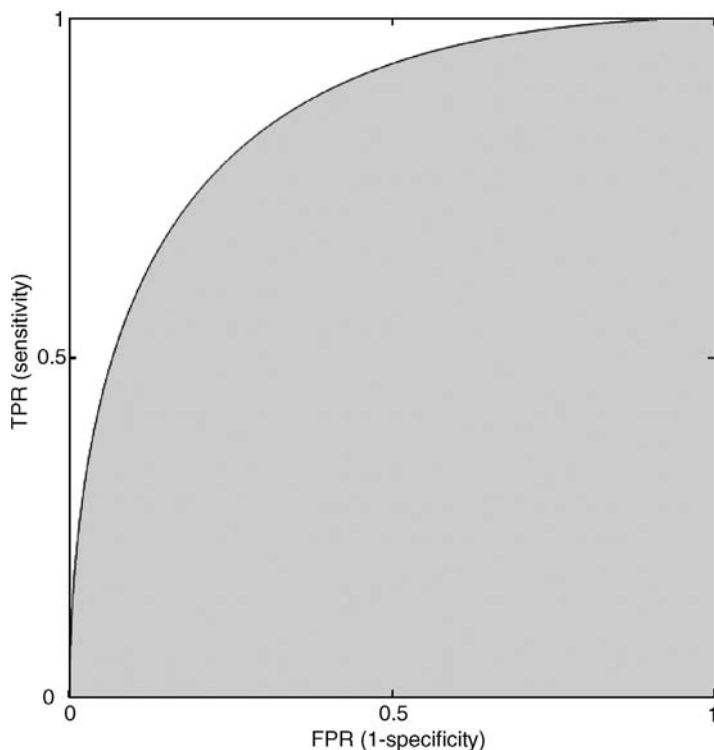


Figure C3 ROC curve.

- **class unfolding** → data set
- **clique** → graph
- **CLOGP** → lipophilicity descriptors (© Leo–Hansch hydrophobic fragmental constants)
- **closeness centrality** → center of a graph
- **cloud point** → technological properties
- **Cluj difference matrix** → Cluj matrices
- **Cluj-detour index** → Cluj matrices
- **Cluj-detour matrix** → Cluj matrices
- **Cluj-distance index** → Cluj matrices
- **Cluj-distance matrix** → Cluj matrices
- **Cluj-Ilmenau index** → Omega polynomial

■ Cluj matrices (CJ)

These are square unsymmetrical matrices $A \times A$ (A being the number of graph vertices), denoted by **UCJ**, defined following the principle of single endpoint characterization of a path; symmetric Cluj matrices, denoted by **SCJ**, are derived from the unsymmetrical Cluj matrices **UCJ** [Diudea, 1996b, 1997a, 1997b]. Several indices can be calculated from Cluj matrices, either directly by the → *orthogonal Wiener operator* from the unsymmetrical matrices or as the half-sum of entries in the symmetric matrices by the → *Wiener operator*.

A Cluj fragment, denoted by $CJ_{ij,p_{ij}}$, collects vertices lying closer to vertex v_i than to vertex v_j , the endpoints of a path p_{ij} . In other words, such a fragment collects the vertex proximity of the i th vertex against any j th vertex, joined by the path p_{ij} , with the distances measured in the subgraph $G-p_{ij}$, obtained by deleting the edges and any internal vertices of the considered path p_{ij} in the $\rightarrow$ *H-depleted molecular graph* G ; the vertices v_i and v_j are not deleted. The Cluj fragment is formally defined as

$$CJ_{ij,p_{ij}} = \{v|v \in V(G-p_{ij}); d_{iv}(G-p_{ij}) < d_{jv}(G-p_{ij})\}$$

that is, the set of vertices closer to v_i than v_j in the component of the subgraph $G-p_{ij}$ containing v_i . The focused vertex v_i is included in the Cluj fragment. $V(G-p_{ij})$ is the set of the subgraph vertices and d is the $\rightarrow$ *topological distance*.

In cycle-containing graphs, more than one path could join the pair v_i and v_j , thus resulting more than one Cluj fragment related to the i th vertex (with respect to the j th vertex and the given path p_{ij}). Therefore, by definition, the off-diagonal entries in the Cluj matrix are taken as the cardinality of the largest Cluj fragment, that is, the fragment with the maximum number of vertices:

$$[UCJ]_{ij} = \begin{cases} \max_{p_{ij}} |CJ_{ij,p_{ij}}| & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

For acyclic graphs, only one Cluj fragment exists for each pair of vertices and, accordingly, the Cluj matrix entry is its cardinality. Moreover, for these graphs, the Cluj fragment cardinality coincides with the number of paths going to the j th vertex through v_i . Diagonal entries are always assumed to be zero.

When the path p_{ij} belongs to the set of topological distances $D(G)$, that is, it is the shortest path connecting vertices v_i and v_j , then, the suffix **D** is added to the matrix symbol, as **UCJD** and **SCJD**, and the matrix is properly called **Cluj-distance matrix**. When the path p_{ij} belongs to the set of detours $\Delta(G)$, that is, it is the longest path connecting vertices v_i and v_j , then, the matrix symbol is **UCJA** or **SCJA**, and the matrix is called **Cluj-detour matrix** [Diudea, Pârv *et al.*, 1997a; Diudea, Katona *et al.*, 1998].

The Cluj matrices are defined for any graph and are, in general, unsymmetrical, except for some symmetric graphs. They can be symmetrized by the $\rightarrow$ *Hadamard matrix product* with their transpose:

$$SCJ = UCJ \otimes UCJ^T$$

where UCJ^T is a transposed unsymmetrical Cluj matrix and **SCJ** is the corresponding symmetric Cluj matrix.

The Cluj matrices defined above, both symmetric and unsymmetrical, can be either **path-Cluj matrices** (**UCJ_p** and **SCJ_p**) when all the pairs of vertices of the graph are accounted for in the matrix calculation or **edge-Cluj matrices** (**UCJ_e** and **SCJ_e**) if the only nonzero elements correspond to edges, that is, only pairs of adjacent vertices are accounted for. The edge-Cluj matrices can be obtained by the Hadamard product of the path-Cluj matrices and the $\rightarrow$ *adjacency matrix* **A**:

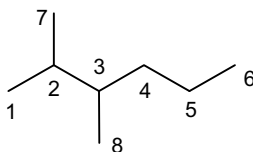
$$SCJ_e = SCJ_p \otimes A \quad UCJ_e = UCJ_p \otimes A$$

In trees, since there exists only one path joining any pair of vertices, Cluj-distance and Cluj-detour matrices coincide; moreover, symmetric Cluj matrices, **SCJD** and **SCJA**, are equal to the $\rightarrow$ *Wiener matrix* **W** (**SCJD_e** = **SCJA_e** = **W_e** and **SCJD_p** = **SCJA_p** = **W_p**). For cyclic graphs, Cluj-distance and Cluj-detour matrices are different, while Wiener matrices are not defined.

Moreover, the relationship for edge-matrices **SCJ_e** = **SZ_e** holds for any graph, **SZ_e** being the $\rightarrow$ *Szeged matrix* defined only accounting for edges, while the path-matrices are different, that is, **SCJD_p** $\neq$ **SZ_p** and **SCJA_p** $\neq$ **SZ_p**.

Example C13

Distance matrix **D**, distance-path matrix **D_p**, unsymmetrical path-Cluj-distance matrix **UCJD_p**, symmetric path-Cluj-distance matrix **SCJD_p**, symmetric edge-Cluj-distance matrix **SCJD_e**, and expanded distance unsymmetrical path-Cluj-distance matrix **D_UCJD_p** for 2,3-dimethylhexane. **W_e** and **W_p** are the edge- and path-Wiener matrices, respectively. **VS_i** and **CS_j** are the row and column sums, respectively.



D									
	1	2	3	4	5	6	7	8	VS_i
1	0	1	2	3	4	5	2	3	20
2	1	0	1	2	3	4	1	2	14
3	2	1	0	1	2	3	2	1	12
4	3	2	1	0	1	2	3	2	14
5	4	3	2	1	0	1	4	3	18
6	5	4	3	2	1	0	5	4	24
7	2	1	2	3	4	5	0	3	20
8	3	2	1	2	3	4	3	0	18
CS_j	20	14	12	14	18	24	20	18	140

Wiener index (**W**)= 70

UCJD_p									
	1	2	3	4	5	6	7	8	VS_i
1	0	1	1	1	1	1	1	1	7
2	7	0	3	3	3	3	7	3	29
3	5	5	0	5	5	5	5	7	37
4	3	3	3	0	6	6	3	3	27
5	2	2	2	2	0	7	2	2	19
6	1	1	1	1	1	0	1	1	7
7	1	1	1	1	1	1	0	1	7
8	1	1	1	1	1	1	1	0	7
CS_j	20	14	12	14	18	24	20	18	140

Wiener index (**W**)= 70

D_p									
	1	2	3	4	5	6	7	8	VS_i
1	0	1	3	6	10	15	3	6	44
2	1	0	1	3	6	10	1	3	25
3	3	1	0	1	3	6	3	1	18
4	6	3	1	0	1	3	6	3	23
5	10	6	3	1	0	1	10	6	37
6	15	10	6	3	1	0	15	10	60
7	3	1	3	6	10	15	0	6	44
8	6	3	1	3	6	10	6	0	35
CS_j	44	25	18	23	37	60	44	35	286

Hyper-distance-path
index (**D_p**)= 143

SCJD_p = W_p									
	1	2	3	4	5	6	7	8	VS_i
1	0	7	5	3	2	1	1	1	20
2	7	0	15	9	6	3	7	3	50
3	5	15	0	15	10	5	5	7	62
4	3	9	15	0	12	6	3	3	51
5	2	6	10	12	0	7	2	2	41
6	1	3	5	6	7	0	1	1	24
7	1	7	5	3	2	1	0	1	20
8	1	3	7	3	2	1	1	0	18
CS_j	20	50	62	51	41	24	20	18	286

Hyper-Wiener index
(**WW**)= 143

Hyper-Cluj-distance
index (**CJ/D_p**)= 143

SCJD _e = W _e										D_UCJD _p									
	1	2	3	4	5	6	7	8	VS _i		1	2	3	4	5	6	7	8	VS _i
1	0	7	0	0	0	0	0	0	7	1	0	1	2	3	4	5	2	3	20
2	7	0	15	0	0	0	7	0	29	2	7	0	3	6	9	12	7	6	50
3	0	15	0	15	0	0	0	7	37	3	10	5	0	5	10	15	10	7	62
4	0	0	15	0	12	0	0	0	27	4	9	6	3	0	6	12	9	6	51
5	0	0	0	12	0	7	0	0	19	5	8	6	4	2	0	7	8	6	41
6	0	0	0	0	7	0	0	0	7	6	5	4	3	2	1	0	5	4	24
7	0	7	0	0	0	0	0	0	7	7	2	1	2	3	4	5	0	3	20
8	0	0	7	0	0	0	0	0	7	8	3	2	1	2	3	4	3	0	18
CS _j	7	29	37	27	19	7	7	7	140	CS _j	44	25	18	23	37	60	44	35	286
Wiener index (W)= 70 Cluj-distance index (CJD _e)= 70										D ^U CJD _p = 143									

For acyclic graphs, the main properties of the unsymmetrical path-Cluj matrix $\mathbf{UCJ}_p$ are

- the row sums of $\mathbf{UCJ}_p$ are equal to the corresponding row sums of the $\rightarrow$ *edge-Wiener matrix* $\mathbf{W}_e$, namely:

$$VS_i(\mathbf{UCJ}_p) = VS_i(\mathbf{W}_e)$$

where VS , which stands for vertex sum, is the $\rightarrow$ *row sum operator*.

- the column sums of $\mathbf{UCJ}_p$ are equal to the corresponding column sums (and row sums) of the $\rightarrow$ *distance matrix* $\mathbf{D}$, namely:

$$CS_j(\mathbf{UCJ}_p) = CS_j(\mathbf{D}) = VS_i(\mathbf{D}) \quad \text{for } i = j$$

where CS_j is the $\rightarrow$ *column sum operator*.

- from the previous relationships it follows:

$$\sum_{i=1}^A VS_i(\mathbf{UCJ}_p) = \sum_{i=1}^A VS_i(\mathbf{W}_e) = \sum_{i=1}^A VS_i(\mathbf{D}) = \sum_{j=1}^A CS_j(\mathbf{UCJ}_p) = 2 \cdot W$$

where W is the $\rightarrow$ *Wiener index*.

$\rightarrow$ *Expanded distance Cluj matrices* were also proposed [Diudea and Gutman, 1998] as a generalization of the expanded distance matrix and calculated by the $\rightarrow$ *Hadamard matrix product* between the unsymmetrical path-Cluj matrices $\mathbf{UCJ}_p$ and the $\rightarrow$ *distance matrix* $\mathbf{D}$:

$$\mathbf{D\_UCJ}_p = \mathbf{D} \otimes \mathbf{UCJ}_p$$

where $\mathbf{UCJ}$ refers to both Cluj-distance and Cluj-detour matrix. Moreover, the topological distance matrix $\mathbf{D}$ can be replaced by the $\rightarrow$ *geometry matrix* $\mathbf{G}$, which collects the 3D interatomic distances, to generate expanded geometric distance Cluj matrices accounting for conformational variability and stereoisomers.

In trees, these expanded distance matrices show the two following properties:

$$\begin{aligned} VS_i(\mathbf{D\_UCJ}_p) &= VS_i(\mathbf{W}_p) \\ CS_j(\mathbf{D\_UCJ}_p) &= CS_j(\mathbf{D}_p) \end{aligned}$$

where VS_i is the $\rightarrow$ row sum operator and CS_j is the $\rightarrow$ column sum operator, $\mathbf{W}_p$ is the $\rightarrow$ path-Wiener matrix and $\mathbf{D}_p$ the $\rightarrow$ distance-path matrix.

Cluj indices are $\rightarrow$ Wiener-type indices calculated either on symmetric (**SCJ**) or unsymmetrical (**UCJ**) Cluj matrices as [Diudea, 1997d; Diudea, Pârv *et al.*, 1997b; Diudea and Gutman, 1998; Katona and Diudea, 2003]

$$\begin{aligned} Wi(\mathbf{SCJ}_e) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{SCJ}_e]_{ij} = Wi^\perp(\mathbf{UCJ}_e) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{UCJ}_e]_{ij} \cdot [\mathbf{UCJ}_e]_{ji} \\ Wi(\mathbf{SCJ}_p) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{SCJ}_p]_{ij} = Wi^\perp(\mathbf{UCJ}_p) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{UCJ}_p]_{ij} \cdot [\mathbf{UCJ}_p]_{ji} \end{aligned}$$

where **CJ** refers to both Cluj-distance and Cluj-detour matrix, **CJ_e** and **CJ_p** indicate the corresponding edge-Cluj and path-Cluj matrices, respectively. Wi is the $\rightarrow$ Wiener operator and $Wi^\perp$ the $\rightarrow$ orthogonal Wiener operator. A is the number of graph vertices.

Therefore, $Wi(\mathbf{SCJD}_e) = Wi^\perp(\mathbf{UCJD}_e)$ is the **Cluj-distance index** (CJD_e), $Wi(\mathbf{SCJD}_p) = Wi^\perp(\mathbf{UCJD}_p)$ is the **hyper-Cluj-distance index** (CJD_p), $Wi(\mathbf{SCJ}\Delta_e) = Wi^\perp(\mathbf{UCJ}\Delta_e)$ is the **Cluj-detour index** ($CJ\Delta_e$), and $Wi(\mathbf{SCJ}\Delta_p) = Wi^\perp(\mathbf{UCJ}\Delta_p)$ is the **hyper-Cluj-detour index** ($CJ\Delta_p$).

Note that the $\rightarrow$ Szeged index SZ_e equals the Cluj-distance index CJD_e for any graph. Moreover, for acyclic graphs, the following relationships hold:

$$\begin{aligned} Wi(\mathbf{UCJD}_p) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{UCJD}_p]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{W}_e]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{D}]_{ij} = W \\ Wi(\mathbf{D\_UCJD}_p) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{D\_UCJD}_p]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{W}_p]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{D}_p]_{ij} = WW \end{aligned}$$

where $\mathbf{UCJD}_p$ is the unsymmetrical path-Cluj-distance matrix, $\mathbf{D\_UCJD}_p$ the corresponding expanded Cluj-distance matrix, $\mathbf{W}_e$ and $\mathbf{W}_p$ the edge-Wiener and path-Wiener matrices, respectively, $\mathbf{D}$ the topological distance matrix, and $\mathbf{D}_p$ the distance-path matrix.

Thus, $Wi(\mathbf{D\_UCJD}_p)$ reduces to the $\rightarrow$ hyper-Wiener index WW , calculated as the half sum of entries in the matrix $\mathbf{D\_UCJD}_p$. This matrix is a direct proof of the finding that, in acyclic graphs, the sum of all internal paths (given by $\mathbf{D}_p$) equals the sum of all external paths (given by $\mathbf{W}_p$) with respect to all pairs (i, j) in the graph [Klein, Lukovits *et al.*, 1995]. The matrix $\mathbf{D\_UCJD}_p$ offers an alternative definition of the hyper-Wiener index.

Moreover, from unsymmetrical Cluj matrices **UCJ** other invariants are the $\rightarrow$ matrix sum indices MS calculated as:

$$MS(\mathbf{UCJ}) = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{UCJ}]_{ij}$$

where the summation goes over all the matrix elements.

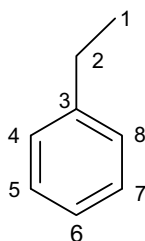
In a bipartite graph, the sum of all edge-counted vertex proximities $MS(\mathbf{UCJD}_e)$ equals the product $A \times B$ of the number of vertices and edges in the graph (e.g., $8 \times 7 = 56$ for 2,3-dimethylhexane in Example C13).

In a tree graph, the sum of all path-counted vertex proximities $MS(\mathbf{UCJD}_p)$ is twice the sum of all distances in the graph or twice the $\rightarrow$ Wiener index W . Moreover, in trees, the $\rightarrow$ PI index,

which represents the edge-counted nonequidistant edges, can be calculated as $MS(UCJD_e)$ from the $\rightarrow$ line graph of the considered molecular graph.

Example C14

Unsymmetrical and symmetric path-Cluj-distance ($UCJD_p$, $SCJD_p$) and path-Cluj-detour ($UCJA_p$, $SCJA_p$) matrices for ethylbenzene. Elements of the corresponding edge-Cluj matrices are highlighted in bold face. VS_i and CS_j are the row sum and column sums, respectively.



$UCJD_p$										$SCJD_p$									
	1	2	3	4	5	6	7	8	VS_i		1	2	3	4	5	6	7	8	VS_i
1	0	1	1	1	1	1	1	1	7	1	0	7	6	4	3	3	3	4	30
2	7	0	2	2	2	2	2	2	19	2	7	0	12	8	6	4	6	6	49
3	6	6	0	5	4	4	4	5	34	3	6	12	0	15	8	8	8	15	72
4	4	4	3	0	5	4	4	2	26	4	4	8	15	0	15	8	8	4	62
5	3	3	2	3	0	5	2	2	20	5	3	6	8	15	0	15	4	8	59
6	3	2	2	2	3	0	3	2	17	6	3	4	8	8	15	0	15	8	61
7	3	3	2	2	2	5	0	3	20	7	3	6	8	8	4	15	0	15	59
8	4	3	3	2	4	4	5	0	25	8	4	6	15	4	8	8	15	0	60
CS_j	30	22	15	17	21	25	21	17	168	CS_j	30	49	72	62	59	61	59	60	452

Cluj-distance index CJD_e :

$$Wi^{\perp}(UCJD_e) = Wi(SCJD_e) = 109$$

Hyper-Cluj-distance index CJD_p :

$$Wi^{\perp}(UCJD_p) = Wi(SCJD_p) = 226$$

$UCJA_p$										$SCJA_p$									
	1	2	3	4	5	6	7	8	VS_i		1	2	3	4	5	6	7	8	VS_i
1	0	1	1	1	1	1	1	1	7	1	0	7	6	1	2	3	2	1	22
2	7	0	2	2	2	2	2	2	19	2	7	0	12	2	4	4	4	2	35
3	6	6	0	3	3	4	3	3	28	3	6	12	0	3	3	8	3	3	38
4	1	1	1	0	1	1	4	1	10	4	1	2	3	0	1	1	8	1	17
5	2	2	1	1	0	1	1	2	10	5	2	4	3	1	0	1	1	8	20
6	3	2	2	1	1	0	1	1	11	6	3	4	8	1	1	0	1	1	19
7	2	2	1	2	1	1	0	1	10	7	2	4	3	8	1	1	0	1	20
8	1	1	1	1	4	1	1	0	10	8	1	2	3	1	8	1	1	0	17
CS_j	22	15	9	11	13	11	13	11	105	CS_j	22	35	38	17	20	19	20	17	188

Cluj-detour index CJA_e :

$$Wi^{\perp}(UCJA_e) = Wi(SCJA_e) = 29$$

Hyper-Cluj-detour index CJA_p :

$$Wi^{\perp}(UCJA_p) = Wi(SCJA_p) = 94$$

Other Cluj indices can be derived from the Cluj polynomials. The **Cluj polynomials** are $\rightarrow$ *counting polynomials* defined on the basis of Cluj matrices as [Diudea, 2002a; Diudea, Vizitiu *et al.*, 2007]

$$CJ(G; x) = \sum_k m(G; k) \cdot x^k$$

where the coefficients $m(G; k)$ are calculated from the entries of the Cluj matrices as the frequency of occurrence of each value k . Cluj polynomials can be calculated both from edge-Cluj matrices (i.e., *Cluj-edge polynomial*) and path-Cluj matrices (i.e., *Cluj-path polynomial*).

Derived from Cluj matrices, the **reciprocal Cluj matrices** CJ^{-1} are the matrices whose elements are the reciprocal of the corresponding Cluj matrix elements [Diudea, 1997c; Diudea, Katona *et al.*, 1998]. $\rightarrow$ *Harary indices* and $\rightarrow$ *hyper-Harary indices* are defined for these matrices applying Wiener and orthogonal Wiener operators.

The **Cluj difference matrix**, denoted as CJ_{Δ} , is obtained in the same way as the Cluj matrix CJ , but path contributions are calculated only on paths larger than one [Diudea, 1996b, 1997a; Ivanciuc, Ivanciuc *et al.*, 1997]. The Cluj difference matrix is calculated as difference between path-Cluj and edge-Cluj matrices:

$$CJ_{\Delta} = CJ_p - CJ_e$$

📖 [Diudea and Randić, 1997; Diudea, 1997b; Gutman and Diudea, 1998; Jäntschi, Katona *et al.*, 2000; Ardelan, Katona *et al.*, 2001; Ursu, Don *et al.*, 2004]

- **Cluj polynomials** $\rightarrow$ Cluj matrices
- **cluster analysis** $\rightarrow$ chemometrics
- **cluster analysis feature selection** $\rightarrow$ variable reduction

■ cluster expansion of chemical graphs

Given a $\rightarrow$ *molecular graph* G , where vertices are labeled by the chemical element of the corresponding atoms, cluster expansion in the additive form is among the $\rightarrow$ *group contribution methods* expressing a molecular property Φ as a sum of contributions of all the connected subgraphs of G , that is,

$$\Phi(G) = \sum_k \phi_k(G') N_k$$

where ϕ_k is the contribution to the molecular property of the k th fragment and k runs over all the connected subgraphs G' . N_k are called **embedding frequencies** and are the number of times a given substructure (cluster) appears in a chemically isomorphic subgraph within the molecular graph. In practice, embedding frequencies are $\rightarrow$ *count descriptors* such as atom-type counts, two-atom fragment counts, etc. [Smolenskii, 1964; Gordon and Kennedy, 1973; Essam, Kennedy *et al.*, 1977; Klein, 1986; Schmalz, Klein *et al.*, 1992]. The property contributions of the fragments are estimated by multivariate regression analysis.

Usually this method is used on a $\rightarrow$ *H-depleted molecular graph*, truncated expansions being obtained considering only fragments up to a user-defined size. Some methods for $\rightarrow$ *log P* estimation are based on cluster expansion. Moreover, a new method for the calculation of embedding frequencies for acyclic trees based on $\rightarrow$ *spectral moments of iterated line graph sequence* was proposed by [Gutman, Popovic *et al.*, 1998; Estrada, 1999c].

📖 [Schmalz, Živković *et al.*, 1987; Poshusta and McHughes, 1989; McHughes and Poshusta, 1990; Baskin, Skvortsova *et al.*, 1995; Grassy, Trape *et al.*, 1995; Kvasnička and Pospichal, 1995; Klein, Schmalz *et al.*, 1999]

- **clustering coefficient of a vertex** → adjacency matrix
- **cluster significance analysis** → variable selection
- **cluster subgraph** → molecular graph
- **CMC** $\equiv$ *critical micelle concentration* → technological properties
- **CMC index** → similarity/diversity
- **CMD index** → similarity/diversity

■ CODESSA descriptors

Among the several CODESSA descriptors, implemented in the homonymous software CODESSA (*Comprehensive DEscriptors for Structural and Statistical Analysis*) [Katritzky and Gordeeva, 1993; CODESSA – Katritzky, Lobanov *et al.*, 1996; Katritzky, Lobanov *et al.*, 1996], are → *molecular weight*, → *molecular volume*, → *count descriptors*, → *topological indices*, → *charge descriptors*, → *shadow indices*, → *charged partial surface area descriptors*, → *quantum-chemical descriptors*, and → *electric polarization descriptors*.

The software CODESSA allows to perform QSAR analysis starting from the calculation of theoretical molecular descriptors up to the evaluation of the best multivariate linear models based on → *variable selection*.

A stepwise selection procedure is adopted to search for QSPR/QSAR models after the preliminary exclusion of → *constant and near-constant variables*. The → *pair correlation cut-off selection* of variables is then performed to avoid highly correlated descriptor variables within the model.

Several molecular properties have been modeled by CODESSA descriptors, such as chromatographic indices [Katritzky, Ignatchenko *et al.*, 1994; Pompe and Novič, 1999], boiling [Katritzky, Mu *et al.*, 1996b; Katritzky, Lobanov *et al.*, 1998; Ivanciuc, Ivanciuc *et al.*, 1998b] and melting points [Katritzky, Maran *et al.*, 1997], critical temperatures [Katritzky, Mu *et al.*, 1998], gas solubilities [Katritzky, Mu *et al.*, 1996a; Huibers and Katritzky, 1998; Katritzky, Wang *et al.*, 1998], critical micelle concentrations [Huibers, Lobanov *et al.*, 1996, 1997], → *solvent polarity scales* [Katritzky, Mu *et al.*, 1997], and mutagenic activities [Maran, Karelson *et al.*, 1999].

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **coefficient of alienation** $\equiv$ *coefficient of nondetermination* → regression parameters
- **coefficient of determination** → regression parameters
- **coefficient of divergence** $\equiv$ *Clark distance* → similarity/diversity (Table S7)
- **coefficient of nondetermination** → regression parameters
- **coefficient of variation** → statistical indices ($\odot$ indices of dispersion)
- **color classes** → chromatic decomposition
- **column sum operator** → algebraic operators
- **column sum vector** → algebraic operators ($\odot$ column sum operator)
- **combinatorial Laplacian matrix** $\equiv$ *Laplacian matrix*
- **combinatorial matrices** → matrices of molecules

■ combined descriptors

These are fixed combinations of selected descriptors accounting for molecular properties of interest. The simplest combined descriptors are the differences and average values of → *basis descriptors* such as → *connectivity indices* or → *path numbers*, and the ratios of different

descriptors defined with the aim of normalization to obtain, for example, size-independent indices. Moreover, optimal linear combinations of highly correlated descriptors are combined descriptors calculated so as to reduce the number of independent variables (e.g., $\rightarrow$ *principal properties*).

Simple sums of different molecular descriptors were proposed as $\rightarrow$ *superindices* to obtain highly discriminant indices; particular superindices were suggested to account for $\rightarrow$ *molecular complexity*.

Examples of combined descriptors are reported below.

Difference indices were proposed as the difference between topological descriptors obtained from the $\rightarrow$ *distance matrix* $\mathbf{D}$ and the $\rightarrow$ *detour matrix* $\mathbf{\Delta}$; they are defined as [Castro, Tueros *et al.*, 2000]

$$\Delta\mathcal{D} = \mathcal{D}(\mathbf{D}) - \mathcal{D}(\mathbf{\Delta})$$

where $\mathcal{D}(\mathbf{D})$ and $\mathcal{D}(\mathbf{\Delta})$ indicate any molecular descriptor obtained from distance and detour matrix, respectively. Difference indices were calculated for $\rightarrow$ *Wiener index*, $\rightarrow$ *Zagreb indices*, and $\rightarrow$ *Schultz molecular topological index*; they equal zero for any acyclic graph, since distance and detour matrices coincide.

A special case of difference indices are the **differential descriptors**, which are $\rightarrow$ *molecular descriptors* or $\rightarrow$ *substituent descriptors* calculated by difference between a compound (or functional group, fragment) and a $\rightarrow$ *reference structure* or a $\rightarrow$ *hyperstructure*. Examples of differential descriptors are those obtained by the $\rightarrow$ *minimal topological difference* (MTD) and $\rightarrow$ *molecular shape analysis* (MSA), as well as some descriptors among the $\rightarrow$ *ETA indices*.

Differential connectivity indices (or **connectivity differences**) are defined as the difference between connectivity indices ${}^m\chi$ and $\rightarrow$ *valence connectivity indices* ${}^m\chi^v$ [Hall and Kier, 1986; Kier and Hall, 1991; Gálvez, García-Domenech *et al.*, 1995; Llacer, Gálvez *et al.*, 2006]:

$${}^m\Delta\chi_t = {}^m\chi_t - {}^m\chi_t^v$$

where the superscript m denotes the order of connectivity indices and the subscript t the type of $\rightarrow$ *molecular subgraph*. These are descriptors proposed to encode electronic information in terms of π and lone pair electrons on that part of the molecule defined by m and t ; moreover, it was found that such descriptors are related to differences in inductive and mesomeric effects [Gálvez, García-Domenech *et al.*, 1994].

Distance measure connectivity indices (DM) are derived from the set of molecular connectivity indices by means of the definition of the $\rightarrow$ *Minkowski distance* [Balaban, Ciubotariu *et al.*, 1990]. They are calculated as

$$DM^k = \sum_{j=1}^{14} [({}^m\chi_t - {}^m\chi_t(R))^k]^{1/k}$$

where the summation goes over all the connectivity indices of different type t up to the sixth order ($m = 6$); k is an integer parameter ranging from 1 to 5 ($k = 1$ is the $\rightarrow$ *Manhattan distance*, $k = 2$ is the $\rightarrow$ *Euclidean distance*); ${}^m\chi_t$ and ${}^m\chi_t(R)$ are the connectivity indices for the considered molecule and a reference molecule R , respectively. DM^k indices can be interpreted as a 14-dimensional measure of the structural diversity of the compound from the reference compound. Methane was proposed as the reference structure, having all the connectivity indices equal to zero.

To account for nondispersive force effects, the **relative valence connectivity indices** to nonpolar compounds were defined as

$$\Delta\chi_{np} = \chi_{np}^v - \chi^v$$

where the nonpolar connectivity index χ_{np}^v is calculated substituting oxygen and nitrogen atoms in the considered molecule by carbon atoms but keeping the number of bonds to all nonhydrogen atoms constant [Bahnick and Doucette, 1988; Schramke, Murphy *et al.*, 1999]. Exceptions to bond constancy were made by replacing the carbonyl group with C—C instead of C=C unless the oxygen atom was directly bonded to an unsaturated ring system (uracils), or the nitrile group with C=C. The difference between connectivity indices of adjacent order was also proposed to model surface tension (${}^2\chi - {}^3\chi$) and critical temperature (${}^1\chi - {}^2\chi$) of alkanes [Randić and Basak, 1999].

A **topological Hammett function** σ_t was also defined by the most significant differences between the $\rightarrow$ *connectivity indices* as

$$\sigma_t = b_0 + b_1 \cdot ({}^4\chi_p - {}^4\chi_p^v) + b_2 \cdot ({}^4\chi_{pc} - {}^4\chi_{pc}^v)$$

where ${}^4\chi_p$, ${}^4\chi_p^v$, ${}^4\chi_{pc}$, and ${}^4\chi_{pc}^v$ are the fourth-order atom and valence $\rightarrow$ *connectivity indices* for path (*p*) and path-cluster (*pc*) graph decompositions; b_j are estimated regression coefficients.

The **L index** was proposed as the molecular descriptor defined as the simple linear combination of molecular $\rightarrow$ *path counts* of order one ${}^1P$ (the number of bonds), order two ${}^2P$ (the $\rightarrow$ *connection number* N_2), and order three ${}^3P$:

$$L = 2 \cdot {}^1P + {}^2P - {}^3P - 2$$

It was found to correlate the sum of ${}^{13}\text{C}$ atomic chemical shifts in alkanes [Miyashita, Okuyama *et al.*, 1989].

Moreover, path count differences ${}^1P - {}^2P$ and ${}^2P - {}^3P$ [Randić and Trinajstić, 1988] and connectivity differences ${}^1\chi - {}^2\chi$ and ${}^2\chi - {}^3\chi$ are often encountered in QSAR modeling; the following path count combination $P_0 + P_1 + P_2 + P_3$ was also found as the critical parameter in the correlation of carbon-13 $\rightarrow$ *chemical shift sums* in alkanes [Miyashita, Okuyama *et al.*, 1989].

Other examples of combined descriptors are the **connectivity quotients** defined as [Gálvez, García-Domenech *et al.*, 1995; Llacer, Gálvez *et al.*, 2006]

$${}^mC_t = \frac{{}^m\chi_t}{{}^m\chi_t^v}$$

where ${}^m\chi$ are the simple $\rightarrow$ *connectivity indices* and ${}^m\chi^v$ the $\rightarrow$ *valence connectivity indices*. Examples of other connectivity quotients are

$${}^1C = \frac{{}^1\chi}{{}^1\chi^v + 1} \quad {}^4C_p = \frac{{}^4\chi_p}{{}^4\chi_p^v + 1} \quad \chi^{23} = \frac{{}^4\chi_p + {}^3\chi}{2}$$

where the first two were proposed by Gálvez [Gálvez, Gomez-Lechón *et al.*, 1996], and the last one was found to correlate well with the van der Waals area [Randić, 1991g].

Semiempirical molecular connectivity terms *X* are special combinations of $\rightarrow$ *connectivity indices* that make use of empirical parameters, dielectric constants, molar masses, and other

ad hoc related parameters accounting for noncovalent interactions [Pogliani, 1997a, 1999a, 1999c]; an example is

$$X = \frac{{}^1\chi}{{}^2\chi + b \cdot {}^3\chi}$$

where b is a parameter to be optimized. These connectivity terms are derived by a trial-and-error procedure based on connectivity indices of lower order.

Other examples of combined descriptors are ratios of some $\rightarrow$ *count descriptors* used by [Zheng, Luo *et al.*, 2005] to define $\rightarrow$ *property filters* and the following:

$$\frac{W}{Z} \quad \frac{CID}{{}^1\chi} \quad \frac{{}^4\chi_{pc}}{MW} \quad \frac{N_X}{MW}$$

where W is the $\rightarrow$ *Wiener index*, Z the $\rightarrow$ *Hosoya Z index*, CID the $\rightarrow$ *connectivity ID number*, N_X the number of atoms of type X , and MW the molecular weight [Boethling and Sabljic, 1989].

Examples of combined descriptors using products are the contributions $q_a \cdot SA_a$, largely used in $\rightarrow$ *CPSA descriptors*, where q and SA are partial charges and atomic surface areas, respectively [Bakken and Jurs, 1999a].

📖 [Stanton and Jurs, 1992; Randić, 1993a]

- **combined matrices** $\rightarrow$ matrices of molecules
- **CoMFA** $\equiv$ *Comparative Molecular Field Analysis* $\rightarrow$ grid-based QSAR techniques
- **CoMFA descriptors** $\rightarrow$ grid-based QSAR techniques (⊙ *Comparative Molecular Field Analysis*)
- **CoMFA fields** $\rightarrow$ molecular interaction fields
- **CoMFA lattice** $\rightarrow$ grid-based QSAR techniques (⊙ *Comparative Molecular Field Analysis*)
- **CoMMA** $\equiv$ *Comparative Molecular Moment Analysis*
- **CoMMA descriptors** $\rightarrow$ comparative molecular moment analysis
- **common overlap length** $\rightarrow$ molecular shape analysis (⊙ *common overlap steric volume*)
- **common overlap surface** $\rightarrow$ molecular shape analysis (⊙ *common overlap steric volume*)
- **common overlap steric volume** $\rightarrow$ molecular shape analysis
- **compactness** $\rightarrow$ distance matrix
- **Comparative Molecular Field Analysis** $\rightarrow$ grid-based QSAR techniques

■ Comparative Molecular Moment Analysis (CoMMA)

The Comparative Molecular Moment Analysis method based on the 3D $\rightarrow$ *molecular geometry* calculates different molecular moments with respect to the $\rightarrow$ *center of mass*, center of charge, and $\rightarrow$ *center-of-dipole* of the molecule [Silverman and Platt, 1996; Silverman, Pitman *et al.*, 1998].

CoMMA descriptors are the following 14 molecular descriptors:

$$\{MW; I_x, I_y, I_z; \mu; Q; \mu_x, \mu_y, \mu_z; d_x, d_y, d_z; Q_{xx}, Q_{yy}\}$$

The first descriptor MW is the $\rightarrow$ *molecular weight*, that is, the zero-order molecular moment with respect to the center of mass. The three $\rightarrow$ *principal moments of inertia* I are the second-order moments with respect to the center of mass. μ and Q are the magnitudes of $\rightarrow$ *dipole*

moment and $\rightarrow$ *quadrupole moment* that are the first- and the second-order moments with respect to the center of charge, respectively. The dipole moment components μ_x , μ_y , and μ_z and the components of displacement d between the center of mass and the center of dipole are calculated with respect to the $\rightarrow$ *principal inertia axes*. Finally, the quadrupole components Q_{xx} and Q_{yy} are calculated with respect to a translated initial reference frame whose origin coincides with the center-of-dipole (Table C5).

By calculating molecular descriptors based on 3D geometry without a common orientation frame, the Comparative Molecular Moment Analysis overcomes the problems due to the molecule alignment.

To extend the CoMMA approach to account for the lipophilicity of the molecule, the $\rightarrow$ *Leo–Hansch hydrophobic fragmental constants* [Abraham and Leo, 1987] have been proposed as a set of atomic lipophilic weights for the calculation of lipophilic molecular multipole moments, called **hydropoles** [Burden and Winkler, 1999a].

Table C5 Molecular moments of order zero, one, and two. A is the number of atoms, MW the molecular weight, q the atomic charges, μ the total dipole moment, and f the hydrophobic atomic constants.

Moment order	Unit	Mass	Charge	Lipophilicity
0	A	MW	$\sum_i q_i$	$\sum_i f_i$
1	0	0	μ	Lipophilic dipole moment
2	Moments of geometry	Moments of inertia	Electrostatic quadrupole moments	Lipophilic quadrupole moments

📖 [Silverman, Pitman *et al.*, 1998, 1999; Silverman, Platt *et al.*, 1998; Burden and Winkler, 1999a; Silverman, 2000a, 2000b; Pitman, Huber *et al.*, 2001; Kovatcheva, Golbraikh *et al.*, 2004; Can, Dimoglo *et al.*, 2005]

- **Comparative Molecular Similarity Indices Analysis** $\rightarrow$ grid-based QSAR techniques
- **Comparative Molecular Surface Analysis** $\rightarrow$ grid-based QSAR techniques
- **Comparative Receptor Surface Analysis** $\equiv$ *CoRSA*
- **Comparative Spectral Analysis** $\rightarrow$ spectra descriptors
- **Comparative Structurally Assigned Spectral Analysis** $\rightarrow$ spectra descriptors
- **Compass descriptors** $\rightarrow$ Compass method

■ Compass method

A QSAR method based on the search for the best model predicting compound activity and likely bioactive conformations and alignments from a set of physical properties measured only near the surface of the molecules [Jain, Koile *et al.*, 1994; Jain, Dietterich *et al.*, 1994; Jain, Harris *et al.*, 1995]. The basic assumption is that the enthalpy of ligand-target binding depends on the interactions occurring at the ligand-target interface. Therefore, the main features characterizing the Compass method are the definition of descriptors related to surface properties, an automatic selection of the optimal molecular conformation and alignment, and the use of $\rightarrow$ *artificial neural networks* with back-propagation to take into account also nonlinear structure-activity relationships.

The method is based on three fundamental phases. The first phase consists in the generation of low-energy conformations for each molecule and in the choice of one conformer as the one most likely to be bioactive; all selected conformers are aligned along with the identified pharmacophore or a substructure common to all molecules in the data set. A molecule *pose* is a conformation of the molecule in a particular alignment.

The second phase proceeds iteratively through three steps. (a) For each molecule pose **Compass descriptors** are calculated as $\rightarrow$ *geometric distances* representing the surface shape or polar functionalities of the pose in the proximity of a given point in the space; compass steric descriptors measure distances from sampling points to the van der Waals surface of a molecule, while donor/acceptor ability descriptors measure the distance from a sampling point to the nearest H-bond donor or acceptor group. Few sampling points are scattered on a surface $2.0 \text{ \AA}$ outside the average van der Waals envelope of the $\rightarrow$ *hypermolecule* obtained by alignment in an invariant and common reference frame (Figure C4). (b) A neural network model is built relating the structural features (Compass descriptors) of molecule poses to biological activity. The network is trained by the backpropagation algorithm and is constituted by three layers with Gaussian input units and standard sigmoid units in the hidden layer. (c) In the third step the model is used to realign the molecules to find better poses, which are then used to give an improved model until convergence is reached.

The third step predicts the activity and bioactive pose of a new molecule.

With respect to $\rightarrow$ CoMFA, the Compass method effectively reduces the number of descriptors, performing a physico-chemically based $\rightarrow$ *variable reduction* and overcomes the problem of guessing the best conformation and alignment of the molecules.

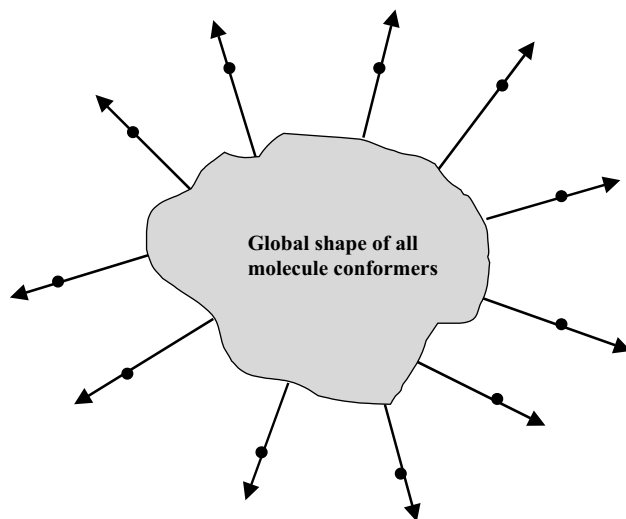


Figure C4 Compass descriptors arising from the molecular surface.

Morphological similarity is a 3D molecular similarity method based on surface shape and charge characteristics of compounds [Jain, 2000]. As in the Compass method, distances to molecular surface from weighted observation points on a uniform grid are calculated. Morphological similarity is defined as a Gaussian function of the differences in molecular

surface distances of two molecules. At each point, a weight is defined in such a way that only grid points that are on the outside of one or another of two molecules contribute to the measure of similarity. Moreover, at each point, for each molecule, in a particular conformation and alignment, three distances are computed: the minimum distance to the van der Waals surface, the minimum distance to a hydrogen-bond acceptor or negatively charged atom, and the minimum distance to a hydrogen-bond donor or positively charged atom. In addition, a directionality term is computed that corresponds to the directional concordance of the vector from an observation point to the polar atom and the atom's favored interaction vector.

- **complementary distance matrix** → distance matrix
- **complementary information content** → indices of neighborhood symmetry
- **complementary Wiener indices** → distance matrix
- **complement Balaban index** → distance matrix
- **complement Barysz distance matrix** → weighted matrices ($\odot$ weighted distance matrices)
- **complement matrices** → matrices of molecules
- **complement Wiener index** → distance matrix
- **complete centric index** → centric indices
- **complete graph** → graph
- **complexity indices** $\equiv$ *molecular complexity indices* → molecular complexity
- **composite ETA index** → ETA indices
- **composite nuclear potential** → quantum-chemical descriptors
- **composite reference ETA index** → ETA indices
- **composition indices** $\equiv$ *atomic composition indices*
- **Compressed Feature Matrix** → substructure descriptors ($\odot$ pharmacophore-based descriptors)

■ computational chemistry

In a broad sense, the term computational chemistry includes several fields such as quantum chemistry, statistical molecular mechanics, molecular modeling, approaches based on → *graph invariants*, molecular graphics and visualization, evaluation of experimental data in X-ray crystallography, NMR spectroscopy, and, in general, spectroscopic techniques; moreover, in this broad sense, analysis, exploration, and modeling performed by → *chemometrics* on experimental data, searching for → *structure-response correlations*, information retrieval from chemical databases, and expert chemical systems are also included in computational chemistry, as constitutive parts of → *chemoinformatics*.

Theoretical chemistry and, especially, quantum chemistry constitute the basic core of computational chemistry and their success covers the field of molecular geometries and energies, reactivity, spectroscopic properties, behavior of electrons in atoms and molecules, and various other fundamental chemical topics [Lipkowitz and Boyd, 1990]. Therefore, the term computational chemistry is also used in a more restricted sense to denote the mathematical approaches and their software implementations to the calculation of molecular properties from theoretical chemistry. → *Quantum-chemical descriptors* are derived from computational chemistry in this restricted sense.

Together with the many methods based on quantum chemistry, other important and effective approaches to computational chemistry are those called *Empirical Force-Field methods* (EFF methods), based on a mechanistic view of the molecule in terms of force constants of bonds,

bending, torsion, and other special interaction terms. The set of force constants constitutes a field of empirical parameters used for the calculation of molecular geometries and energies.

Calculations based on computational chemistry methods can be performed by means of software packages, such as *AMPAC* [AMPAC, 2005], *GAMESS* [GAMESS, 2005], *GAUSSIAN* [GAUSSIAN03 – Pople and *et al.*, 1990], *JAGUAR* [Jaguar – Schrödinger, 1990], *MOLPRO* [MOLPRO – Werner, Knowles *et al.*, 1991], *MOPAC* [MOPAC – Air Force Academy, 1999], *NWCHEM* [NWChem – EMSL, 1990], *SPARTAN* [SPARTAN, 2005], and *TURBOMOLE* [TURBOMOLE, 2007].

📖 [Lewis, 1916, 1923; Mulliken, 1928a, 1928b, 1955a; Hückel, 1930, 1932; Pauling, 1932, 1939; Pauling and Wilson, 1935; Coulson, 1939, 1960; Eyring, Walter *et al.*, 1944; Streitweiser, 1961; Dewar, 1969; Murrell and Harget, 1972; Lowe, 1978; Löw and Saller, 1988; Parr and Yang, 1989; Stewart, 1990; Leach, 1996; Szabo and Ostlund, 1996; Jorgensen, Olsen *et al.*, 2000]

- **Computer-Aided Drug Design** → drug design
- **Computer-Aided Molecular Design** → drug design
- **Computer-Aided Molecular modeling** → drug design
- **COMSA** $\equiv$ *Comparative Molecular Surface Analysis* → topological feature maps
- **CoMSIA** $\equiv$ *Comparative Molecular Similarity Indices Analysis* → grid-based QSAR techniques
- **CON index** → statistical indices (⊙ concentration indices)
- **concentration indices** → statistical indices
- **conditional Wiener index** → Wiener index
- **conductance matrix** → resistance matrix
- **Conformational-Dependent Chirality Code** → chirality descriptors (⊙ Chirality Codes)
- **conformational global sensitivity** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **Conformational-Independent Chirality Code** → chirality descriptors (⊙ Chirality Codes)
- **conformational invariance** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **conformational pairwise sensitivity** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **Conformation Energy Profile** → 4D-Molecular Similarity Analysis
- **confusion matrix** → classification parameters
- **congenericity principle** → Structure/Response Correlations
- **conjugation** → delocalization degree indices
- **connected graph** → graph
- **connectedness index** → Wiener index
- **connection** → edge adjacency matrix
- **connection number** → edge adjacency matrix
- **connection orbital information content** → orbital information indices
- **connective eccentricity index** → eccentricity-based Madan indices (⊙ Table E1)
- **connectivity bond layer matrix** → layer matrices
- **connectivity differences** $\equiv$ *differential connectivity indices* → combined descriptors
- **connectivity ID number** $\equiv$ *Randić connectivity ID number* → ID numbers
- **connectivity index** $\equiv$ *Randić connectivity index* → connectivity indices

■ connectivity indices

Connectivity indices are among the most popular $\rightarrow$ *topological indices* and are calculated from the $\rightarrow$ *vertex degree* δ of the atoms in the $\rightarrow$ *H-depleted molecular graph*. The **Randić connectivity index** was the first connectivity index proposed [Randić, 1975b, 2008; Li and Gutman, 2006]; it is also called **connectivity index** or **branching index**, and is defined as

$$\chi_R \equiv {}^1\chi = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i \cdot \delta_j)^{-1/2} = \sum_{b=1}^B (\delta_{b(1)} \cdot \delta_{b(2)})_b^{-1/2}$$

where the first summation goes over all the pairs of vertices v_i and v_j in the molecular graph, but only contributions from pairs of adjacent vertices are accounted for, a_{ij} being the elements of the $\rightarrow$ *adjacency matrix* A ; the second summation goes over all the edges in the molecular graph. A and B are the total number of vertices and edges in the graph, respectively; δ_i and δ_j are the vertex degrees of the vertices v_i and v_j ; the subscripts $b(1)$ and $b(2)$ represent the two vertices connected by the edge b .

The Randić connectivity index is closely related to the $\rightarrow$ *second Zagreb index* M_2 and was proposed as measure of $\rightarrow$ *molecular branching*.

The term $(\delta_i \cdot \delta_j)^{-1/2}$ for each pair of adjacent vertices is called **edge connectivity** and can be used to characterize edges as a primitive $\rightarrow$ *bond order* accounting for bond accessibility, that is, the accessibility of a bond to encounter another bond in intermolecular interactions, as the reciprocal of the vertex degree δ is the fraction of the total number of nonhydrogen sigma electrons contributed to each bond formed with a particular atom [Kier and Hall, 2000]. This interpretation places emphasis on the bimolecular encounter possibility among molecules, reflecting collective influence of the bond accessibilities of each molecule with other molecules in its immediate environment. Therefore, the Randić connectivity index ${}^1\chi$ can be interpreted as the contribution of one molecule to the bimolecular interaction arising from the encounters of bonds of two identical molecules:

$${}^1\chi = \sqrt{\sum_{b=1}^B \sum_{b'=1}^B (\delta_i \cdot \delta_j)_b^{-1/2} \cdot (\delta_k \cdot \delta_l)_{b'}^{-1/2}}$$

where the two summations run over all the bonds of the molecules and δ are the vertex degrees.

Important papers about characteristics and meaning of the connectivity indices are: [Kier and Hall, 2000, 2002; Hall and Kier, 2001; Randić, 2001g; Estrada, 2002b].

$\rightarrow$ *Information connectivity indices* based on the partition of the edges in the graph according to the equivalence and the magnitude of their edge connectivity values were derived.

The mean Randić connectivity index (or mean Randić branching index) is defined as

$$\bar{\chi}_R = \frac{\chi_R}{B}$$

where B is the number of edges in the molecular graph.

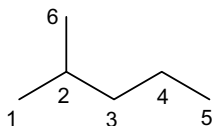
A variant of the Randić connectivity index was also proposed as

$$\chi'_R = A \cdot \chi_R$$

where A is the number of graph vertices [Mihalić, Nikolić *et al.*, 1992].

Example C15

Connectivity indices of order 0, 1, and 2 for 2-methylpentane.



Atoms	1	2	3	4	5	6
δ_i	1	3	2	2	1	1

$${}^0\chi = \delta_1^{-1/2} + \delta_2^{-1/2} + \delta_3^{-1/2} + \delta_4^{-1/2} + \delta_5^{-1/2} + \delta_6^{-1/2} =$$

$$= 1^{-1/2} + 3^{-1/2} + 2^{-1/2} + 2^{-1/2} + 1^{-1/2} + 1^{-1/2} = 4.992$$

$${}^1\chi = (\delta_1 \times \delta_2)^{-1/2} + (\delta_2 \times \delta_3)^{-1/2} + (\delta_3 \times \delta_4)^{-1/2} + (\delta_4 \times \delta_5)^{-1/2} + (\delta_2 \times \delta_6)^{-1/2} =$$

$$= (1 \times 3)^{-1/2} + (3 \times 2)^{-1/2} + (2 \times 2)^{-1/2} + (2 \times 1)^{-1/2} + (3 \times 1)^{-1/2} = 2.770$$

$${}^2\chi = (\delta_1 \times \delta_2 \times \delta_3)^{-1/2} + (\delta_2 \times \delta_3 \times \delta_4)^{-1/2} + (\delta_3 \times \delta_4 \times \delta_5)^{-1/2} + (\delta_1 \times \delta_2 \times \delta_6)^{-1/2} +$$

$$+ (\delta_3 \times \delta_2 \times \delta_6)^{-1/2} =$$

$$= (1 \times 3 \times 2)^{-1/2} + (3 \times 2 \times 2)^{-1/2} + (2 \times 2 \times 1)^{-1/2} + (1 \times 3 \times 1)^{-1/2} + (2 \times 3 \times 1)^{-1/2}$$

$$= 2.183$$

Kier and Hall defined [Kier and Hall, 1986; Kier and Hall, 1977b] a general scheme based on the Randić index to calculate also zero-order and higher order descriptors; these are called **Molecular Connectivity Indices (MCIs)**, also known as **Kier–Hall connectivity indices**. They are calculated by the following:

$${}^0\chi = \sum_{i=1}^A \delta_i^{-1/2} \quad {}^1\chi = \sum_{b=1}^B (\delta_i \cdot \delta_j)_b^{-1/2} \quad {}^2\chi = \sum_{k=1}^{2P} (\delta_i \cdot \delta_l \cdot \delta_j)_k^{-1/2}$$

$${}^m\chi_t = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i \right)_k^{-1/2}$$

where k runs over all of the m th order subgraphs constituted by n atoms ($n = m + 1$ for acyclic subgraphs); K is the total number of m th order subgraphs present in the molecular graph and in the case of the path subgraphs equals the m th order path count ${}^mP$. The product is over the simple vertex degrees δ of all the vertices involved in each subgraph. The subscript “ t ” for the connectivity indices refers to the type of $\rightarrow$ *molecular subgraph* and is “ ch ” for chain or ring, “ pc ” for path-cluster, “ c ” for cluster, and “ p ” for path (that can also be omitted). Obviously, the first-order Kier–Hall connectivity index is the Randić connectivity index.

By replacing the vertex degree δ by the $\rightarrow$ *valence vertex degree* δ^v in the formulas reported above, similar **valence connectivity indices** were proposed [Kier and Hall, 1981, 1983b], denoted by ${}^m\chi_i^v$, able to account for the presence of heteroatoms in the molecule as well as double and triple bonds (Table C6).

Table C6 Values of the first-order Kier–Hall connectivity index for some substituent groups attached to a Carbon atom with a valence vertex degree equal to 3.

Substituent	${}^1\chi^v$	Substituent	${}^1\chi^v$	Substituent	${}^1\chi^v$
–H	0	–COOH	0.7164	–CH ₂ CH ₂ S CH ₃	1.762
–CH ₃	0.5773	–NH ₂	0.3333	–CH ₂ CH ₂ COOH	1.6900
–OH	0.2582	–CH ₂ OH	0.7240	–COH	0.5690

Analogously, **bond order-weighted vertex connectivity indices**, denoted by ${}^m\chi_i^b$, were also defined by using the $\rightarrow$ *bond vertex degree* δ^b instead of the simple vertex degree δ to specifically account for multiplicity in the molecule. To derive these connectivity indices, either the $\rightarrow$ *conventional bond order* or quantum-chemical derived $\rightarrow$ *bond orders* can be used [Estrada and Montero, 1993; Estrada and Molina, 2001a; Jalbout and Li, 2003c]. Moreover, connectivity indices were also calculated using the $\rightarrow$ *Z-delta number* δ^Z and therefore denoted by ${}^m\chi^Z$ [Pogliani, 1999b]. Another set of modified valence connectivity indices was proposed based on the $\rightarrow$ *Li valence vertex degree* δ^{Li} , used in place of the original valence vertex degree [Li, Jalbout *et al.*, 2003].

The inverse-square-root function was selected for the Randić connectivity index, and later used for the most general connectivity indices, because it provided high correlation with properties of isomeric alkane series, thus showing high sensitivity to variation in molecular structure. However, it was observed that it is not a very effective connectivity measure in nonisomeric molecule series as it shows two opposing trends: to increase with molecular size and to decrease with $\rightarrow$ *molecular complexity* [Bonchev, 2001a]. Substituting the inverse-square-root function in favor of the total adjacency function, the $\rightarrow$ *overall connectivity indices* were proposed as a measure of topological complexity.

Table C7 Some connectivity and valence connectivity indices for the data set of phenethylamines (Appendix C– Set 2).

Mol.	X	Y	${}^0\chi$	${}^1\chi$	${}^2\chi$	${}^3\chi$	${}^4\chi$	${}^5\chi$	${}^0\chi^v$	${}^1\chi^v$	${}^2\chi^v$	${}^3\chi^v$	${}^4\chi^v$	${}^5\chi^v$
1	H	H	8.975	5.698	5.005	3.298	2.639	1.702	9.082	4.952	4.246	2.505	1.972	0.825
2	H	F	9.845	6.092	5.627	3.708	2.791	1.914	9.383	5.052	4.387	2.575	1.982	0.824
3	H	Cl	9.845	6.092	5.627	3.708	2.791	1.914	10.139	5.43	4.823	2.827	2.108	0.969
4	H	Br	9.845	6.092	5.627	3.708	2.791	1.914	10.969	5.845	5.302	3.104	2.246	1.129
5	H	I	9.845	6.092	5.627	3.708	2.791	1.914	11.541	6.131	5.632	3.295	2.342	1.239
6	H	Me	9.845	6.092	5.627	3.708	2.791	1.914	10.005	5.363	4.746	2.783	2.086	0.943
7	F	H	9.845	6.092	5.639	3.625	2.934	1.978	9.383	5.052	4.39	2.554	1.991	0.887
8	Cl	H	9.845	6.092	5.639	3.625	2.934	1.978	10.139	5.43	4.827	2.789	2.19	1.128
9	Br	H	9.845	6.092	5.639	3.625	2.934	1.978	10.969	5.845	5.306	3.047	2.408	1.393
10	I	H	9.845	6.092	5.639	3.625	2.934	1.978	11.541	6.131	5.636	3.225	2.559	1.575
11	Me	H	9.845	6.092	5.639	3.625	2.934	1.978	10.005	5.363	4.749	2.747	2.155	1.086

(Continued)

Table C7 (Continued)

Mol.	X	Y	$^0\chi$	$^1\chi$	$^2\chi$	$^3\chi$	$^4\chi$	$^5\chi$	$^0\chi^v$	$^1\chi^v$	$^2\chi^v$	$^3\chi^v$	$^4\chi^v$	$^5\chi^v$
12	Cl	F	10.715	6.503	6.135	4.287	3.042	2.161	10.44	5.535	4.916	2.939	2.186	1.122
13	Br	F	10.715	6.503	6.135	4.287	3.042	2.161	11.27	5.95	5.363	3.257	2.394	1.381
14	Me	F	10.715	6.503	6.135	4.287	3.042	2.161	10.306	5.468	4.844	2.887	2.153	1.08
15	Cl	Cl	10.715	6.503	6.135	4.287	3.042	2.161	11.195	5.913	5.323	3.388	2.304	1.258
16	Br	Cl	10.715	6.503	6.135	4.287	3.042	2.161	12.026	6.328	5.77	3.863	2.511	1.517
17	Me	Cl	10.715	6.503	6.135	4.287	3.042	2.161	11.062	5.846	5.251	3.311	2.27	1.216
18	Cl	Br	10.715	6.503	6.135	4.287	3.042	2.161	12.026	6.328	5.77	3.882	2.433	1.407
19	Br	Br	10.715	6.503	6.135	4.287	3.042	2.161	12.856	6.743	6.217	4.529	2.64	1.666
20	Me	Br	10.715	6.503	6.135	4.287	3.042	2.161	11.892	6.261	5.698	3.777	2.399	1.365
21	Me	Me	10.715	6.503	6.135	4.287	3.042	2.161	10.928	5.78	5.179	3.236	2.249	1.192
22	Br	Me	10.715	6.503	6.135	4.287	3.042	2.161	11.892	6.261	5.698	3.756	2.49	1.493

Connectivity-like indices are molecular descriptors calculated applying the same mathematical formula as the connectivity indices, but substituting the vertex degree δ with any $\rightarrow$ *local vertex invariant* (LOVI):

$${}^m\text{Chi}_t(\mathcal{L}) = \sum_{k=1}^K \left(\prod_{i=1}^n \mathcal{L}_i \right)_k^{-1/2}$$

where $\mathcal{L}_i$ is the general symbol for local vertex invariants, the summation goes over all the subgraphs of type t constituted by n atoms and m edges; K is the total number of such m th order subgraphs present in the molecular graph, and each subgraph is weighted by the product of the local invariants associated to the vertices contained in the subgraph. Connectivity-like indices may also be calculated by replacing local vertex invariants $\mathcal{L}_i$ with $\rightarrow$ *atomic properties* P_i .

The general formula for the calculation of connectivity-like indices, which uses the row sums VS_i of a $\rightarrow$ *graph-theoretical matrix* as the local vertex invariants, was called by Ivanciuc **Chi operator** [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc, 2001c]. Specifically, for any square symmetric ($A \times A$) matrix $\mathbf{M}(w)$ representing a molecular graph with A vertices and a $\rightarrow$ *weighting scheme* w , the *Chi operator* is defined as

$${}^m\text{Chi}(\mathbf{M}; w) = \sum_{k=1}^K \left(\prod_{i=1}^n VS_i(\mathbf{M}, w) \right)_k^{-1/2}$$

where VS_i indicates the $\rightarrow$ *row sum operator*.

Randić-like indices are connectivity-like indices defined for graph edges and calculated by using the same mathematical formula as the $\rightarrow$ *Randić connectivity index* ${}^1\chi$, but replacing the vertex degree δ with any $\rightarrow$ *local vertex invariants* $\mathcal{L}$:

$${}^1\chi(\mathcal{L}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\mathcal{L}_i \cdot \mathcal{L}_j)^{-1/2} = \sum_{b=1}^B (\mathcal{L}_{b(1)} \cdot \mathcal{L}_{b(2)})_b^{-1/2}$$

where A and B are the total number of vertices and edges in the graph, respectively; a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to one for pairs of adjacent vertices, and zero otherwise; the subscripts $b(1)$ and $b(2)$ represent the two vertices connected by the edge b . Note

that, in the left expression, the summation goes over all pairs of vertices in the graph but the only nonvanishing contributions are from the pairs of adjacent vertices for which elements a_{ij} equal one. Several mathematical properties of Randić-like indices were investigated [Gutman, 2002a; Li and Gutman, 2006].

Moreover, **generalized connectivity indices** are a generalization of the Kier–Hall connectivity indices in terms of a variable exponent λ as:

$${}^m\chi_t = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i \right)_k^\lambda$$

where λ is any real exponent. If $\lambda = 1$ and $m = 1$, the $\rightarrow$ *second Zagreb index* M_2 is obtained; values of $\lambda = -1$ and $\lambda = 1/2$ were considered by Altenburg [Altenburg, 1980], and values of $\lambda = -1/3$ and $\lambda = -1/4$ were also investigated [Randić, Hansen *et al.*, 1988; Estrada, 1995c; Amić, Beslo *et al.*, 1998; Ivanciuc, Ivanciuc *et al.*, 2002e].

Related to Randić-like indices are the $\rightarrow$ *Balaban-like indices*, which only differ for the normalization factor.

Some connectivity-like indices are reported below. Other connectivity-like indices reported elsewhere are $\rightarrow$ *JJ indices* derived from the $\rightarrow$ *Wiener matrix*, $\rightarrow$ *electronegativity-based connectivity indices*, $\rightarrow$ *extended edge connectivity indices*, $\rightarrow$ *chiral connectivity indices*, $\rightarrow$ *variable connectivity indices*, $\rightarrow$ *line graph connectivity indices*, and $\rightarrow$ *line graph Randić connectivity index*.

• Evans extended connectivity indices

Two molecular descriptors proposed to generalize the Randić connectivity index, defined as [Evans, Lynch *et al.*, 1978]

$$\chi_2^{ext} = \sum_b [(^2f_i n_i \cdot ^2f_j n_j)_b \cdot \pi_b^*]^{-1/2} \quad \text{and} \quad \chi_3^{ext} = \sum_b [(^3f_i n_i \cdot ^3f_j n_j)_b \cdot \pi_b^*]^{-1/2}$$

where i and j are indices for the two adjacent vertices incident to the edge b , n_i is an integer describing the i th atom type, ^2f_i and ^3f_i are the second and third order $\rightarrow$ *vertex distance counts* of the i th vertex, that is, the number of the vertices at topological distances 2 and 3 from the i th vertex, respectively. π_b^* is the $\rightarrow$ *conventional bond order* and the summation runs over all edges.

• environment connectivity descriptors

These are Randić-like indices of molecular fragments calculated on fragment atoms and then first neighbors [Jurs, Chou *et al.*, 1979]. The value of the Randić connectivity index for a given fragment represents the immediate surroundings of the substructure as embedded within the molecule. If the fragment is not present in the molecule, zero value is given.

Environment descriptors are closely related to $\rightarrow$ *substructure descriptors*, differing from the latter in using real values in place of binary variables or counts. The set of fragments is defined by the user depending on the data set and the specific problem.

• Fragment Molecular Connectivity indices (FMC)

These are first-order connectivity indices computed for predefined positions on molecular fragments in congeneric series [Takahashi, Miashita *et al.*, 1985]. By superimposition of all congeneric compounds, a template structure is derived whose vertices define the positions for the *FMC* indices; the vertices of the common parent structure are not considered in defining the

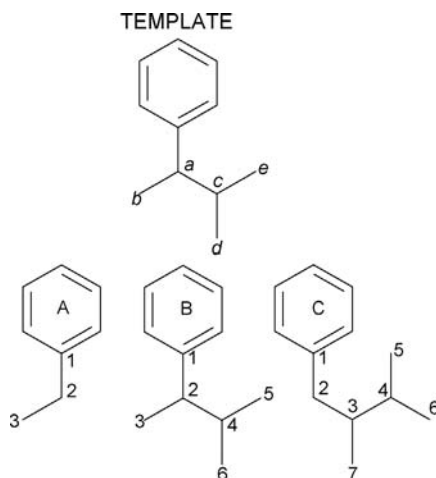
positions. For a k th position the corresponding fragment connectivity index is defined as

$$FMC_k = \sum_i (\delta_k \cdot \delta_i)^{-1/2}$$

where δ can be the simple $\rightarrow$ *vertex degree* or the $\rightarrow$ *valence vertex degree*, and i denotes each vertex joint to the vertex in the k th position (the link vertex of the parent molecule is also considered). By definition, FMC is equal to zero if there is no atom of the substituent in the considered position. Each molecule is finally described by a number of FMC values, corresponding to the number of predefined positions.

Example C16

Fragment molecular connectivity indices.



$$FMC_a(A) = (\delta_2 \cdot \delta_1)^{-1/2} + (\delta_2 \cdot \delta_3)^{-1/2} = (2 \cdot 3)^{-1/2} + (2 \cdot 1)^{-1/2} = 1.115$$

$$FMC_b(A) = (\delta_3 \cdot \delta_2)^{-1/2} = (1 \cdot 2)^{-1/2} = 0.707 \quad FMC_c(A) = FMC_d(A) = FMC_e(A) = 0$$

$$FMC_a(B) = (\delta_2 \cdot \delta_1)^{-1/2} + (\delta_2 \cdot \delta_3)^{-1/2} + (\delta_2 \cdot \delta_4)^{-1/2} = (3 \cdot 3)^{-1/2} + (3 \cdot 1)^{-1/2} + (3 \cdot 3)^{-1/2} = 1.244$$

$$FMC_b(B) = (\delta_3 \cdot \delta_2)^{-1/2} = (1 \cdot 3)^{-1/2} = 0.577 \quad FMC_d(B) = FMC_e(B) = 0.577$$

$$FMC_c(B) = (\delta_4 \cdot \delta_2)^{-1/2} + (\delta_4 \cdot \delta_5)^{-1/2} + (\delta_4 \cdot \delta_6)^{-1/2} = (3 \cdot 3)^{-1/2} + (3 \cdot 1)^{-1/2} + (3 \cdot 1)^{-1/2} = 1.488$$

$$FMC_a(C) = (\delta_2 \cdot \delta_1)^{-1/2} + (\delta_2 \cdot \delta_3)^{-1/2} = (2 \cdot 3)^{-1/2} + (2 \cdot 3)^{-1/2} = 0.816$$

$$FMC_c(C) = (\delta_3 \cdot \delta_2)^{-1/2} + (\delta_3 \cdot \delta_4)^{-1/2} + (\delta_3 \cdot \delta_7)^{-1/2} = (3 \cdot 2)^{-1/2} + (3 \cdot 3)^{-1/2} + (3 \cdot 1)^{-1/2} = 1.319$$

Molecule	FMC_a	FMC_b	FMC_c	FMC_d	FMC_e
A	1.115	0.707	0	0	0
B	1.244	0.577	1.488	0.577	0.577
C	0.816	0	1.319	0.577	1.488

• walk connectivity indices

These Randić-like indices are defined by using the $\rightarrow$ atomic walk counts as the local vertex invariants in place of the vertex degrees and applying the Randić-type formula as [Razinger, 1986]

$$\chi^W = \sum_{b=1}^B (awcs_i \cdot awcs_j)_b^{-1/2} \quad \chi^{kW} = \sum_{b=1}^B (awc_i^{(k)} \cdot awc_j^{(k)})_b^{-1/2}$$

where the summations run over all edges in the H-depleted molecular graph and the subscripts i and j refer to the two vertices incident to the considered edge; the first index χ^W is calculated from the $\rightarrow$ atomic walk count sum $awcs$, that is, considering all walks of any length from the vertex, while the second index χ^{kW} is calculated counting only the walks of length k from each vertex ($awc^{(k)}$). In particular, the **longest walk connectivity index** χ^{LW} was proposed as a highly discriminant descriptor, defined as

$$\chi^{LW} = \sum_{b=1}^B (awc_i^{(A-1)} \cdot awc_j^{(A-1)})_b^{-1/2}$$

where only the longest walks of length $A-1$ from each vertex are counted, A being the total number of graph vertices.

The **Randić–Razinger index** χ_i^{kW} is a $\rightarrow$ local vertex invariant, defined as [Diudea, Minailiuc *et al.*, 1997a]

$$\chi_i^{kW} = \sum_{j=1}^A a_{ij} (awc_i^{(k)} \cdot awc_j^{(k)})^{-1/2}$$

where $awc_i^{(k)}$ and $awc_j^{(k)}$ are the $\rightarrow$ atomic walk counts of order k for vertices v_i and v_j ; the summation goes over all the vertices, but accounts only for contributions from vertices adjacent to v_i , a_{ij} being the elements of the adjacency matrix. It can be noted that the sum of these LOVIs over all the vertices corresponds to twice the corresponding walk connectivity index:

$$2 \cdot \chi^{kW} = \sum_{i=1}^A \chi_i^{kW}$$

• Kupchik modified connectivity indices

These are modifications of the Randić connectivity index defined in such a way as to account for the presence of heteroatoms in the molecule [Kupchik, 1986, 1988, 1989]:

$${}^1\chi^r = \sum_{b=1}^B (\delta_i^{\text{het}} \cdot \delta_j^{\text{het}})_b^{-1/2} \quad \text{and} \quad {}^1\chi^b = \sum_{b=1}^B \frac{r_{ij}}{r_{CC}} \cdot (\delta_i \cdot \delta_j)_b^{-1/2}$$

where the summations run over all the edges in the molecular graph and i, j denote the vertices incident with the considered edge; r_{ij} is the bond length and r_{CC} a standard carbon–carbon bond

length (1.54 Å); δ is the simple $\rightarrow$ vertex degree, that is, the number of first neighbors. The $\rightarrow$ Kupchik vertex degree δ^{het} is calculated as

$$\delta_i^{\text{het}} = \frac{R_C}{R_i} \cdot (Z_i^v - h_i)$$

where R_i and R_C are the covalent radius of the i th atom and the carbon atom, respectively; Z_i^v is the atomic number and h_i the number of hydrogen atoms bonded to i th vertex.

The first index was later called **radius-corrected connectivity index** and the second one **bond-length-corrected connectivity index** [Sun, Huang *et al.*, 1996].

These modified connectivity indices were found to be related to the $\rightarrow$ molar refractivity of alkanes, alkylsilanes, and alkylgermanes.

- **perturbation connectivity indices** (${}^m\chi_t^p$)

These are connectivity-like indices based on the $\rightarrow$ perturbation delta value δ^p and defined as [Gombar, Kumar *et al.*, 1987]

$${}^m\chi_t^p = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i^p \right)_k^{-1/2}$$

where

$$\delta_i^p = \delta_i^v + \sum_{j=1}^A a_{ij} \cdot \gamma_{ij} \cdot \delta_j^v$$

is the perturbation delta value. The summation in the first formula goes over all of the m th order subgraphs of type t containing n atoms; K is the total number of m th order subgraphs; a_{ij} are the elements of the $\rightarrow$ adjacency matrix equal to one for adjacent vertices and otherwise zero. Perturbation delta values are obtained from $\rightarrow$ valence vertex degrees δ^v modified by atomic environment.

- **3D-connectivity indices** (${}^m\chi\chi_t$)

These are connectivity-like indices derived from the $\rightarrow$ geometry matrix $\mathbf{G}$; they are defined using the $\rightarrow$ geometric distance degree ${}^G\sigma$ in place of the topological vertex degree δ [Randić, 1988a, 1988b; Randić, Jerman-Blazic *et al.*, 1990]:

$${}^m\chi\chi_t = \sum_{k=1}^K \left(\prod_{i=1}^n {}^G\sigma_i \right)_k^{-1/2}$$

where k runs over all of the m th order subgraphs constituted by n vertices; K is the total number of m th order subgraphs. The subscript “ t ” refers to the type of molecular subgraph.

- **total structure connectivity index**

This is an extremal connectivity index contemporarily accounting for all the vertices in the graph as [Needham, Wei *et al.*, 1988]

$$\chi_T = \left(\prod_{i=1}^A \delta_i \right)^{-1/2}$$

Note that the total structure connectivity index is the square root of the $\rightarrow$ *simple topological index* proposed by Narumi for measuring molecular branching.

• **local connectivity indices** (${}^m\bar{\chi}_i$) ($\equiv$ *atomic connectivity indices*)

Computed for individual vertices in a graph, they were developed by equally partitioning each term ${}^mw_{ij} = (\delta_i \cdot \delta_k \cdot \dots \cdot \delta_j)^{-1/2}$ of the connectivity indices among all of the vertices along the path i - j of m th order, as [Balaban, Catana *et al.*, 1990]

$${}^m\bar{\chi}_i = \frac{1}{m+1} \cdot \sum_{j=1} {}^mP_{ij} {}^mw_{ij}$$

where ${}^mw_{ij}$ is the $\rightarrow$ *path connectivity*, the index j represents the terminal vertex v_j of a path and the summation runs over all the paths ${}^mP_{ij}$ of length m starting from vertex v_i ; $m+1$ is the number of vertices along each path of length m .

For example, the first-order local connectivity index for the i th vertex is defined as

$${}^1\bar{\chi}_i = \frac{1}{2} \cdot \sum_{j=1}^{\delta_i} {}^1w_{ij}$$

where δ_i is the vertex degree of the i th vertex, that is, the number of edges incident to the i th vertex. The zero-order local connectivity index is simply defined as

$${}^0\bar{\chi}_i = (\delta_i)^{-1/2}$$

By summing these local connectivity indices over all the nonhydrogen atoms, the Kier–Hall connectivity indices are reproduced.

• **H₁ topological index**

This is a Randić-like index defined as [Li and You, 1993a, 1993b; Li Zhang *et al.*, 1995]

$$\begin{aligned} H_1 &= \left(\sum_{b=1}^B \frac{1}{(1+\Delta_b)\sqrt{\delta_i \cdot \delta_j}} \right)^2 = \\ &= \sum_{b=1}^B \left(\frac{1}{(1+\Delta_b)\sqrt{\delta_i \cdot \delta_j}} \right)_b^2 + \sum_{b=1}^{B-1} \sum_{b'=b+1}^B \left(\frac{1}{(1+\Delta_b)\sqrt{\delta_i \cdot \delta_j}} \right)_b \cdot \left(\frac{1}{(1+\Delta_{b'})\sqrt{\delta_i \cdot \delta_j}} \right)_{b'} \end{aligned}$$

where the summations run over all B edges in the $\rightarrow$ *H-filled molecular graph*; δ_i and δ_j are the $\rightarrow$ *vertex degree* of the two vertices incident to the considered b th edge. Δ_b is a bond parameter representing the interaction between the two bonded vertices i and j and calculated as the following:

$$\Delta_b = \alpha \cdot (\text{IP}_i - \text{EA}_j)_b + (1-\alpha) \cdot (\text{IP}_i - \text{EA}_j)_b$$

where IP and EA are the $\rightarrow$ *ionization potential* and $\rightarrow$ *electron affinity*, respectively. The first term in the equation represents the electron transfer interaction from the HOAO (Highest Occupied Atomic Orbital) of the i th atom to the LUAO (Lowest Unoccupied Atomic Orbital) of the j th atom, the second term represents the feedback interaction from the HOAO of the j th

atom to LUAO of the i th atom. The parameter α is used to modulate the importance of the two kinds of interaction; it is generally taken equal to 0.5.

- **modified Randić index** (${}^1\chi_{\text{mod}}$)

This is a molecular descriptor based on atomic properties, accounting for valence electrons and connectivities in the H-depleted molecular graph, calculated by using a Randić-like formula [Lohninger, 1993]:

$${}^1\chi_{\text{mod}} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot \frac{Z_i}{\sqrt{\delta_i \cdot \delta_j}}$$

where the summation goes over all the pairs of vertices in the molecular graph; the only nonvanishing terms are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the $\rightarrow$ adjacency matrix; δ is the $\rightarrow$ vertex degree and Z the atomic number.

- **charge-weighted vertex connectivity indices** (${}^m\Omega_i(q)$)

Very similar to the $\rightarrow$ electronic-topological descriptors, charge-weighted vertex connectivity indices are connectivity-like indices obtained by replacing the simple vertex degree with a charge-related atomic quantity calculated by $\rightarrow$ computational chemistry [Estrada and Montero, 1993; Estrada, 1995d; Estrada and Molina, 2001a]:

$${}^m\Omega_i(q) = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i(q) \right)_k^{-1/2}$$

where $\delta_i(q) = q_i - h_i$ is the **electron charge density weight**, q_i being the electron charge density on the i th atom and h_i the number of hydrogen atoms bonded to it. Another set of connectivity-like indices, called **corrected charge-weighted vertex connectivity indices**, and denoted by ${}^m\Omega_i^c(q)$, was also defined based on a local vertex invariant corrected for hydrogen atomic charges as

$$\delta_i^c(q) = q_i - \sum_j q_j^{\text{H}}$$

where q_j^{H} is the electron charge density of j th hydrogen atom bonded to the i th atom. This quantity was called **corrected electron charge density weight**.

- **atomic molecular connectivity index** (χ^c) ($\equiv$ molecular connectivity topochemical index)

This index was designed as an extension of the Randić connectivity index to take into account the relative size of heteroatoms in a H-depleted molecular graph. It is based on the $\rightarrow$ Madan vertex degree derived from the $\rightarrow$ chemical adjacency matrix [Goel and Madan, 1995]:

$$\chi^c = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i^c \cdot \delta_j^c)^{-1/2}$$

where the only nonvanishing terms in the summation are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the $\rightarrow$ adjacency matrix; δ^c is the Madan vertex

degree, calculated by summing up relative atomic weights of all the adjacent atoms, assuming the carbon atom weight as the reference.

This index was applied in a structure-activity studies on anti-inflammatory activity of pyrazole carboxylic acid hydrazide analogs. It was demonstrated that, despite the overall classification accuracy was above 80%, this descriptor did not perform better than the popular valence connectivity index.

- **Euclidean connectivity index (χ^E)**

Derived from the $\rightarrow$ geometry matrix G , it is defined by using a Randić-like formula applied to $\rightarrow$ geometric distance degrees ${}^G\sigma$ used in place of the topological vertex degrees δ [Balasubramanian, 1995b]:

$$\chi^E = \sum_{i=1}^{A-1} \sum_{j=i+1}^A ({}^G\sigma_i \cdot {}^G\sigma_j)^{-1/2}$$

This index discriminates the geometrical isomers and can be considered as a measure of the compactness of a molecule in the 3D space. Note that all possible atom pairs are considered instead of the pairs of bonded atoms because in 3D space there exists an Euclidean distance between every pair of atoms.

- **local Balaban index**

This is a local vertex invariant, denoted as J_i and defined by using a Randić-like formula [Diudea, Minailiuc *et al.*, 1997a]:

$$J_i = \sum_{j=1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2}$$

where σ_i and σ_j are the $\rightarrow$ vertex distance degrees of vertices v_i and v_j ; the summation goes over all the vertices, but accounts only for contributions from vertices adjacent to v_i , a_{ij} being the elements of the adjacency matrix. The sum of the local Balaban index over all the graph vertices is related to the $\rightarrow$ Balaban distance connectivity index J by the following relation:

$$\frac{2 \cdot J \cdot (C + 1)}{B} = \sum_{i=1}^A J_i$$

where C and B are the $\rightarrow$ cyclomatic number and the number of edges, respectively.

- **solvation connectivity indices (${}^m\chi^s$)**

These are molecular descriptors defined to model solvation entropy and describe dispersion interactions in solution [Zefirov and Palyulin, 2001]. Taking into account the characteristic dimension of the molecules by atomic parameters, they are defined as

$${}^m\chi^s = \frac{1}{2^{m+1}} \cdot \sum_{k=1}^K \frac{\left(\prod_{i=1}^n L_i \right)_k}{\left(\prod_{i=1}^n \delta_i \right)_k^{1/2}}$$

where L is the principal quantum number (2 for C, N, O atoms; 3 for Si, S, Cl, . . .) associated to a vertex in the k th subgraph and δ the corresponding $\rightarrow$ vertex degree; K is the total number of m th order subgraphs; n is the number of vertices in the subgraph. The normalization factor $1/(2^{m+1})$ is defined in such a way that the indices ${}^m\chi$ and ${}^m\chi^s$ for compounds containing only second-row atoms coincide.

For example, the first-order solvation connectivity index is

$${}^1\chi^s = \frac{1}{4} \cdot \sum_{b=1}^B \frac{(L_i \cdot L_j)_b}{(\delta_i \cdot \delta_j)_b^{1/2}}$$

where the summation goes over all the B edges; L_i and L_j are the principal quantum numbers of the two vertices incident to the considered edge. This index coincides with the Randić connectivity index ${}^1\chi$ for the hydrocarbons, being $L = 2$ for all the atoms.

These molecular descriptors are defined for a $\rightarrow$ *H-depleted molecular graph*; furthermore, fluorine atoms are not included in the graph, their dimension being very close to that of the hydrogen atom.

• Yang connectivity index (${}^1\chi^Y$)

This is a Randić-like index based on the $\rightarrow$ Yang vertex degree [Jiang, Liu *et al.*, 2003]. It is defined as

$${}^1\chi^Y = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i^Y \cdot \delta_j^Y)^{-1/2}$$

where δ^Y is the vertex degree based on the $\rightarrow$ Yang's electronegativity force gauge; the only nonvanishing terms in the summation are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the $\rightarrow$ adjacency matrix;

A widespread use of connectivity descriptors in modeling a lot of molecular properties is recognizable in the literature since 1975.

 Additional references are collected in the thematic bibliography (see Introduction).

- **connectivity-like indices** $\rightarrow$ connectivity indices
- **connectivity matrix** $\equiv$ atom connectivity matrix $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **connectivity quotients** $\rightarrow$ combined descriptors
- **connectivity table** $\rightarrow$ molecular geometry
- **connectivity valence layer matrix** $\rightarrow$ layer matrices
- **Connolly surface area** $\rightarrow$ molecular surface ($\odot$ solvent-accessible molecular surface)
- **consecutive AT numbers** $\rightarrow$ vertex degree
- **consensus analysis** $\rightarrow$ structure/response correlations
- **consensus binding free energy** $\rightarrow$ scoring functions
- **consensus fingerprint** $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **constant interval reciprocal indices** $\rightarrow$ distance matrix
- **constant and near-constant variables** $\rightarrow$ variable reduction

■ constitutional descriptors

These are the most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its → *molecular geometry* or topology.

The most common constitutional descriptors are number of atoms (→ *atom number*), number of bonds (→ *bond number*), absolute and relative numbers of specific atom-types (→ *count descriptors*), absolute and relative numbers of single, double, triple, and aromatic bonds, number of rings (→ *cyclomatic number*), number of rings divided by the number of atoms or bonds, number of benzene rings, number of benzene rings divided by the number of atoms, → *molecular weight* and → *average molecular weight*, → *atomic composition indices*, → *information index on size*, etc.

These descriptors are insensitive to any conformational change, do not distinguish among isomers, and are either → *0D-descriptors* or → *1D-descriptors*.

- **constitutional graph** ≡ *molecular graph*
- **contact surface** → molecular surface (⊙ solvent-accessible molecular surface)
- **contingency coefficient** → statistical indices (⊙ concentration indices)
- **Continuous Chirality Measure** → chirality descriptors
- **continuous wavelet transforms** → spectra descriptors
- **contour length** → size descriptors (⊙ Kuhn length)
- **contour profiles** → molecular profiles
- **conventional bond order** → bond order indices
- **conventional bond order ID number** → ID numbers
- **core count** → ETA indices
- **Corey–Pauling–Koltun volume** → volume descriptors
- **corrected charge-weighted vertex connectivity indices** → connectivity indices (⊙ charge-weighted vertex connectivity indices)
- **corrected electron charge density weight** → connectivity indices (⊙ charge-weighted vertex connectivity indices)
- **corrected second moments** ≡ *topological atomic valencies* → self-returning walk counts
- **corrected structure count** ≡ *algebraic structure count* → Kekulé number
- **corrected Taft steric constant** → steric descriptors (⊙ Taft steric constant)
- **correlation distance** → similarity/diversity (Table S7)
- **correlation matrix** → statistical indices (⊙ correlation measures)
- **correlation measures** → statistical indices
- **correlation weights** → variable descriptors
- **Correlation Weights of the Local Invariants of Molecular Graphs** → variable descriptors

■ CoRSA (≡ *Comparative Receptor Surface Analysis*)

CoRSA is a 3D QSAR approach applied to compute structure-activity models whenever the structure of the biological target is not known [Ivanciuc, Ivanciuc *et al.*, 2000a, 2000b; Hirashima, Kuwano *et al.*, 2003]. Using the common steric and electrostatic features of the most active members of a series of compounds, CoRSA generates a virtual receptor model, represented as points on a surface complementary to the van der Waals surface of the set of compounds. The structural descriptors of the model are represented by the total interaction energies between each surface point of the virtual receptor and all atoms in a molecule. These descriptors are used in a Partial Least Squares (PLS) regression to generate a structure-activity model.

The development of a CoRSA model consists of the following seven steps. (1) The geometry of all molecules in the data set is optimized with molecular mechanics or quantum mechanics methods. (2) All optimized molecules are aligned (superimposed) using some pharmacophore hypothesis. The CoRSA model depends on the molecule alignment and errors in this step may lead to models that have a low predictive power. (3) A subset of the most active molecules is selected to generate the virtual receptor model; these compounds form the receptor generation set (RGS) of molecules. The central assumption is the complementarity between the shape and properties of these molecules and the virtual receptor. (4) The virtual receptor is generated using information on the geometry, volume, atomic charges, hydrophobicity, hydrogen-bonding, or other properties of the selected molecules. Unlike real receptors, the virtual receptor is not formed by atoms, but by a three-dimensional receptor surface represented by points having certain properties. The coordinates of these points are generated from the shape field of the RGS molecules.

Two field functions are used to create the shape of the virtual receptor, namely the van der Waals field function and the Wyvill field function. Each field source corresponds to an atom. The van der Waals field function generated by the atom i at distance r_{ij} is

$$V_i^{vdw} = r_{ij} - R_i^{vdw}$$

where r_{ij} is the distance from the atom i to the grid point j and R_i^{vdw} is the van der Waals radius of the atom i . This field function, which is computed for every grid point, has the property that inside the van der Waals volume the value is negative, outside the volume the value is positive, and at the van der Waals surface the value $V(r)$ is zero. If a grid point contains a shape field value computed for a different atom, the smaller of the two values is assigned to that grid point. The **Wyvill function** is a bounded function that decays completely in a finite distance R . The Wyvill field function generated by the atom i at distance r_{ij} is

$$V_i^{Wv} = 1 - \frac{4 \cdot r_{ij}^6}{9 \cdot R^6} + \frac{17 \cdot r_{ij}^4}{9 \cdot R^4} - \frac{22 \cdot r_{ij}^2}{9 \cdot R^2}$$

where r_{ij} is the distance from the atom i to the grid point j . A field value is the sum of the field values contributed by each atom; if a grid point is outside of R , its shape field value is not computed. The value of R depends on the atom type, and usually it is twice the van der Waals radius of the atom i . The Wyvill function has the properties that $V(0) = 1$, $V(R) = 0$, and $V(R/2) = 1/2$.

Using the shape field values the marching cubes isosurface algorithm produces a set of triangulated surface points representing the surface of the virtual receptor. The default grid spacing of 0.5 Å yields an average surface density of 6 points/Å². This gives an average distance between neighboring points (points in the same triangle) of about 0.47 Å.

(5) Each surface point from the virtual receptor contains information about the local properties of the receptor. These properties include electrostatic potential, partial charge, hydrophobicity, and hydrogen-bonding propensity. (6) With the virtual receptor model defined in steps (1)–(5), for each molecule in the data set, a number of molecular descriptors are derived by computing the ligand–receptor interaction energy between each surface point from the virtual receptor and the atoms in the molecule. (7) Finally, the molecular descriptors calculated for all the molecules are processed by PLS algorithm to generate the 3D-QSAR model.

- **CoSA** $\equiv$ Comparative Spectral Analysis $\rightarrow$ spectra descriptors
- **CoSASA** $\equiv$ Comparative Structurally Assigned Spectral Analysis $\rightarrow$ spectra descriptors
- **cosine similarity coefficient** $\rightarrow$ similarity/diversity (Table S9)
- **cospectral graphs** $\equiv$ *isospectral graphs* $\rightarrow$ graph
- **Coulomb potential energy function** $\rightarrow$ molecular interaction fields ($\odot$ electrostatic interaction fields)

■ count descriptors

These are simple molecular descriptors based on counting the defined elements of a compound. The most common chemical count descriptors are $\rightarrow$ *atom number* A , $\rightarrow$ *bond number* B , $\rightarrow$ *cyclomatic number* C , $\rightarrow$ *hydrogen-bond acceptor number* and $\rightarrow$ *hydrogen-bond donor number*, $\rightarrow$ *distance-counting descriptors*, $\rightarrow$ *path counts*, $\rightarrow$ *walk counts*, $\rightarrow$ *atom pairs*, and other related $\rightarrow$ *substructure descriptors*.

When the different chemical nature of atoms is considered, the **atom-type count** is defined as the number of atoms of the same chemical element. $A \rightarrow$ *molecular graph* G can be characterized by a vector of atom-type counts as

$$\{N_C; N_H; N_O; N_N; N_S; N_F; N_{Cl}; N_{Br}; N_I; \dots\}$$

whose entries represent the number of carbon, hydrogen, oxygen, nitrogen, sulfur, fluorine, chlorine, bromine, and iodine atoms, respectively. These descriptors are derived from the chemical formula, that is, they are $\rightarrow$ *OD-descriptors*. The **relative atom-type count** is the ratio between a given atom count and the total number A of atoms, therefore the following vector can be defined:

$$\{\bar{N}_C; \bar{N}_H; \bar{N}_O; \bar{N}_N; \bar{N}_S; \bar{N}_F; \bar{N}_{Cl}; \bar{N}_{Br}; \bar{N}_I; \dots\}$$

where

$$\bar{N}_X = \frac{N_X}{A}$$

The **atomistic topological indices** were proposed by Burden [Burden, 1996] as atom-type counts where each atom is classified by its element and the number of connections, thus also accounting for atom hybridization. In particular, N_p , N_s , N_t , and N_q are the number of primary, secondary, tertiary, and quaternary carbon atoms, respectively; N_{sp^3} , N_{sp^2} , and N_{sp} the numbers of sp^3 , sp^2 , and sp carbon atoms, respectively; N_{AR} the number of aromatic carbon atoms; and N_{Xsp^3} , N_{Xsp^2} , and N_{Xsp} the numbers of sp^3 , sp^2 , and sp heavy atoms, respectively.

Strictly related are the carbon-type counts called **STIMS indices (Simplest Topological Integers from Molecular Structures)** and proposed by Pal *et al.* [Pal, Sengupta *et al.*, 1988, 1989, 1990; Pal, Purkayastha *et al.*, 1992] as a subset of $\rightarrow$ *TAU indices*. These are number of methyl carbon (N_P), number of methylene carbons (N_I), number of tertiary carbons (N_Y), number of quaternary carbons (N_X), and number of branched carbons (N_B).

Count descriptors measuring the molecular unsaturation are within the $\rightarrow$ *multiple bond descriptors*, such as the number of double bonds (DB), the number of triple bonds (TB), the number of aromatic bonds (AB), the number of rings (NRG), which is the $\rightarrow$ *cyclomatic number* (denoted as C).

The **functional group count** can be defined considering the well-known *functional chemical groups*, which are groups of atoms having a characteristic and specific reactivity, such as

$$\{N_{OH}; N_{COOH}; N_{NH_2}; N_{C=O}; N_{OCH_3}; N_{SH}; N_{H_2C=}; N_{BENZ}; \dots\}$$

whose entries represent the number of oxydryl, carboxylic, aminic, carbonilic, methoxy, thyo, methylen, and phenyl functional groups, respectively.

Andrews descriptors are particular atom and functional group counts relative to those groups found to be statistically significant in receptor binding modeling [Andrews, Craik *et al.*, 1984]: CO_2^- , PO_4^- , N^+ , N, OH, C=O, ether and thioether groups, halogens, sp^3 and sp^2 carbon atoms, and the $\rightarrow$ *rotatable bond number*.

Even more general is the definition of **fragment count** as the number of a specific kind of *molecular fragments*, which are arbitrary-selected groups of adjacent atoms in a molecule. A general method for modeling $\rightarrow$ *physico-chemical properties* using fragment counts is the $\rightarrow$ *cluster expansion of chemical graphs*.

The **subgraph count set (SCS)** is a vectorial descriptor, where each entry is the number of times specific subgraphs are obtained by cutting one edge at a time in a $\rightarrow$ *H-depleted molecular graph* [Oberrauch and Mazzanti, 1990]:

$$\{N_{METHYL}; N_{ETHYL}; N_{PROPYL}; N_{ISOPROPYL}; \dots\}$$

The order of counts is not defined *a priori* and a subset of relevant subgraph counts can be used instead of the complete SCS. In chemical terms, these subgraphs are recognized as radicals.

Both the functional group count and the fragment count can be derived from recognized substructures within the molecule, that is, they are $\rightarrow$ *1D-descriptors*; in fact they are also considered specific $\rightarrow$ *substructure descriptors*.

Count descriptors give local chemical information, are insensitive to isomers, to conformational changes and show a high level of degeneracy. However, due to their immediate availability, they are among the most used descriptors.

Examples of count descriptors are reported by Feher and Schmidt [Feher and Schmidt, 2003]. Moreover, count descriptors are usually the basic molecular descriptors used to generate $\rightarrow$ *property filters*. Examples of count descriptors used for property filters are those proposed by [Zheng, Luo *et al.*, 2005] which are listed in Table C8.

Table C8 Some of the descriptors proposed in [Zheng, Luo *et al.*, 2005].

Symbol	Definition
A3	Number of sp^3 hybridized C, O, S, and N atoms
UNC	Number of sp , sp^2 and aromatic carbons
AUH	Number of atoms rather than H and halogens
BDUH	Number of the bonds that do not contain H and halogen atoms
C3p	N_{sp^3}/AUH
UNC_C3	UNC/N_{sp^3}
A3_C	$A3/N_C$
h_p	Ratio of the number of hydrogen atoms over the total number of nonhalogen heavy atoms
NO_C3	$(N_N + N_O)/N_{sp^3}$

📖 [Chiorboli, Piazza *et al.*, 1993a, 1993b, 1993c, 1996; Tosato, Piazza *et al.*, 1992; Okey and Stensel, 1996; Okey, Stensel *et al.*, 1996; Winkler, Burden *et al.*, 1998; Kaiser and Niculescu, 1999; Tan and Siebert, 2004]

➤ **counter-propagation neural network** → Self-Organizing Maps

■ counting polynomials

The counting polynomial is a description of a graph property in terms of a sequence of numbers, such as the distance degree sequence or the sequence of the number of k independent edge sets [Hosoya, 1988, 1990; Trinajstić, 1992; Diudea, Gutman *et al.*, 2001; Noy, 2003; Diudea, Vizitiu *et al.*, 2007]. The counting polynomial is defined as

$$P(G; x) = \sum_k m(G; k) \cdot x^k$$

where the exponent k represents the extent of the considered graph partitions and the coefficients $m(G; k)$ are related to the frequency of the occurrences of partitions of extent k . Polynomial coefficients are graph invariants and are thus related to the structure of a molecule graph.

Examples of counting polynomials are → *Z-counting polynomial*, → *Wiener polynomial*, → *Cluj polynomials*, → *matching polynomial*, and → *omega polynomial*.

The **Altenburg polynomial** is another example of counting polynomials defined for → *H-depleted molecular graph* G as

$$\alpha(G, a) = \sum_{k=1}^D {}^k f \cdot a_k$$

where the sum runs over all the distances in the graph, D being the → *topological diameter*, that is, the maximum distance in the graph, ${}^k f$ the → *graph distance count* of k th order, that is, the number of distances equal to k in the graph, and a_k the independent variables [Altenburg, 1961]. The Altenburg polynomial is closely related to the → *Wiener index* of a graph: graphs with the same Altenburg polynomials always have just the same Wiener numbers (the contrary does not always hold).

In general, coefficients, roots, and derivatives of counting polynomials can be used for characterization of molecular graphs and as molecular descriptors in QSAR/QSPR modeling.

- **covalent hydrogen-bond acidity** → Theoretical Linear Solvation Energy Relationships
- **covalent hydrogen-bond basicity** → Theoretical Linear Solvation Energy Relationships
- **covariance** → statistical indices (⊙ correlation measures)
- **covariance matrix** → statistical indices (⊙ correlation measures)
- **CPK volume** ≡ *Corey-Pauling-Koltun volume* → volume descriptors
- **CPSA descriptors** ≡ *charged partial surface area descriptors*
- **Craig plot** → Hansch analysis
- **Cramer coefficient** → statistical indices (⊙ concentration indices)
- **critical constants** → physico-chemical properties
- **critical micelle concentration** → technological properties
- **critical packing parameter** → GRID-based QSAR techniques (⊙ VolSurf descriptors)
- **critical pressure** → physico-chemical properties (⊙ critical constants)

- **critical temperature** → physico-chemical properties (⊙ critical constants)
- **critical volume** → physico-chemical properties (⊙ critical constants)
- **crosscorrelation descriptors** → autocorrelation descriptors
- **cross-validated R^2** → regression parameters
- **cross-validation** → validation techniques
- **CSA2_{Cl} index** → charged partial surface area descriptors (⊙ HDCA index)
- **CSA2_H index** → charged partial surface area descriptors (⊙ HDCA index)
- **CT vertex degree** → vertex degree
- **cubic root molecular weight** → physico-chemical properties (⊙ molecular weight)
- **CWLIMG** $\equiv$ *Correlation Weights of the Local Invariants of Molecular Graphs* → variable descriptors
- **cycle** $\equiv$ *cyclic path* → graph
- **cycle-edge incidence matrix** → incidence matrices (⊙ cycle matrices)
- **cycle matrices** → incidence matrices
- **cycle-vertex incidence matrix** → incidence matrices (⊙ cycle matrices)
- **cyclicity** → graph
- **cyclicity index** → detour matrix
- **cyclicity indices** → molecular complexity (⊙ molecular cyclicity)
- **cyclic graph** → graph
- **cyclic path** → graph
- **cyclomatic number** → ring descriptors
- **Czekanowski similarity coefficient** $\equiv$ *Dice similarity coefficient* → similarity/diversity (⊙ Table S9)

D

➤ **DAI indices** → atom-type-based topological indices

■ **DARC/PELCO analysis**

DARC/PELCO analysis, that is, a topological method dealing with structural environments, was originally proposed and, then, later refined and broadened by Dubois [Dubois, Laurent *et al.*, 1966, 1973a, 1973b, 1976; Dubois, 1976].

The method is based on a combination of the DARC system (**Description, Acquisition, Retrieval and Computer-aided design**) and the PELCO (**Perturbation of an Environment Limited Concentric and Ordered**) procedure. Moreover, it accounts for the simultaneous representation of all the data set compounds and the population of the compounds structurally contained in them; the data set compounds, which are those compounds for which a molecular property has been experimentally evaluated, generate an ordered multidimensional space that constitutes a → *hyperstructure*.

Each molecule is represented by an ordered → *chromatic graph*, which describes the topological and chemical nature of each site; vertex chromatism corresponds to the chemical nature of atoms, edge chromatism to the bond multiplicity.

The hyperstructure is built following the operations of focalization, organization, ordering, and chromatic evaluation of each data set molecule.

The superimposition of all the ordered graphs of the data set molecules provides the hyperstructure whose central topological vertex corresponds to the focus, and the environment is organized in concentric layers A–B where each vertex corresponds to an atom present in at least one compound.

The *focus* is the → *maximum common substructure* among the data set compounds. The *environment* is organized in concentric layers centered on the focus and is *limited* and *concentric* (ELC). The vertices at an odd distance from the focus belong to layer A; the vertices at an even distance belong to layer B. Each pair of successive A–B layers starting from the focus constitutes an *environment limited to B* (EB). The environment is *ordered* (ELCO) in the sense that each site (vertex) is located unambiguously by means of a topological coordinate (A_i or B_{ij}); each topological coordinate gives a development direction for the environment.

More specifically, from the first concentric environment of the hyperstructure, the different substitution sites give rise to the main development directions starting from the focus. Note that each hyperstructure vertex actually is a topochromatic site, which can contain more than one atom and can be labeled differently, in accordance with the chemical nature of the atoms in it.

The hyperstructure can be mathematically represented by the **DARC/PELCO matrix**, which has n rows, the number of data set compounds, and N_S columns, the topochromatic sites of the hyperstructure. Each row of the DARC/PELCO matrix is called **topochromatic vector**, denoted as I_i , and directly accounts for the overall topology of the one molecule. The **DARC/PELCO descriptors** of a molecule are the elements of its topochromatic vector and are binary variables I_{is} equal to 1 if the s th topochromatic site of the hyperstructure contains an atom of the i th molecule and zero otherwise.

The **DARC/PELCO model** is defined as

$$\hat{y}_i = y_0 + \sum_{s=1}^{N_S} b_s \cdot I_{is}$$

where y_0 is the response of the parent molecule, that is, the focus of the hyperstructure, and b_s are the regression coefficients called *perturbations*. In spite of the formal analogy with the $\rightarrow$ *Free-Wilson model*, where the contribution to the biological activity of each group in a substitution site is considered additive and independent of the structural variation in the rest of the molecule, the DARC/PELCO model considers the contribution of a substituent group to the biological activity as the sum of ordered perturbations given by all the vertices starting from the focus and characterizing that group (Figure D1).

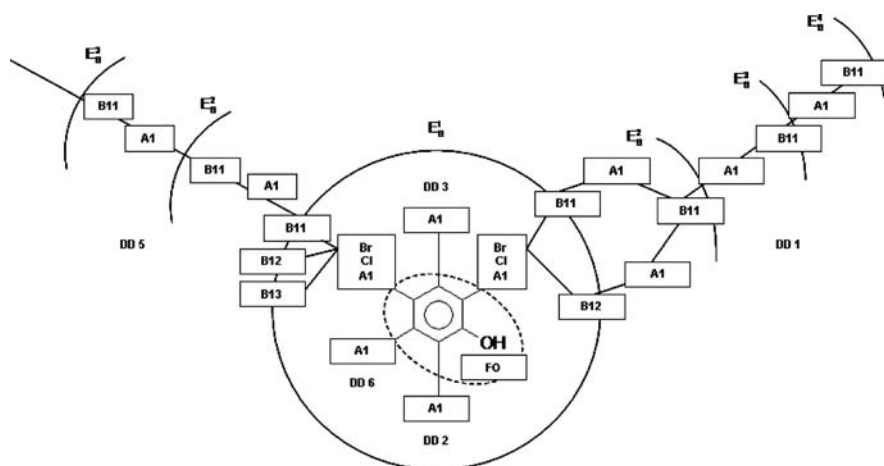


Figure D1 Example of DARC/PELCO hyperstructure. The focus is constituted by the phenol.

A particular advantage of this method is the determination of the structural area for reliable predictions. This area, called **proference**, consists of all the structures generated by the hyperstructure that do not belong to the population of data set compounds (**population trace**). In other words, predictable structures are localized in the hyperstructure and their activity is predicted by interpolation, using the corresponding topochromatic vector where each vertex present ($I_{is} = 1$) refers to a site existing in at least one data set compound. Different levels of reliability are determined depending on the extent to which the structure is surrounded by data set compounds.

📖 [Dubois, Laurent *et al.*, 1967, 1975a, 1975b; Duperray, Chastrette *et al.*, 1976a, 1976b; Dubois, Mercier *et al.*, 1979, 1986; Mercier and Dubois, 1979; Dubois, Chrétien *et al.*, 1980; Panaye, MacPhee *et al.*, 1980; Bawden, 1983; Doucet, Panaye *et al.*, 1983; Dubois, Sicouri *et al.*, 1984; Dubois and Sobel, 1985; Dubois, Panaye *et al.*, 1987; De La Guardia, Carrión *et al.*, 1988; Bonchev, 1989; Attias and Dubois, 1990; Mercier, Troullier *et al.*, 1990; Mercier, Mekenyan *et al.*, 1991; Dubois and Loukianoff, 1993; Mekenyan, Mercier *et al.*, 1993; Panaye, Doucet *et al.*, 1993; Carabédian and Dubois, 1998; Dubois, Doucet *et al.*, 1999]

- **DARC/PELCO descriptors** → DARC/PELCO analysis
- **DARC/PELCO matrix** → DARC/PELCO analysis
- **DARC/PELCO model** → DARC/PELCO analysis
- **Daren fitness function** → regression parameters
- **Dash–Behera steric density parameter** $\equiv$ *steric density parameter* → steric descriptors
- **data** → data set
- **data distance matrix** → similarity/diversity
- **data matrix** → data set

■ data set

This is a collection of *objects* described by one or more *variables*. An **object** is a basic unit in data analysis; for example, an individual, a molecule, an experiment, and a sample. Each object is described by one or more measurements, called **data**. A **variable** represents a characteristic of the objects that may take any value from a specified set, for example, a physico-chemical property, a molecular descriptor.

A data set is often considered as a sample from a population and the sample parameters calculated from the data set as estimates of the population parameters (→ *statistical indices*). Moreover, it is used to calculate statistical models such as quantitative → *structure/response correlations*. In this case, the data set is organized into a **data matrix X** with n rows and p columns where each row corresponds to an object of the data set and each column to a variable; therefore, each matrix element x_{ij} represents the value of the j th variable for the i th object ($i = 1, \dots, n$; $j = 1, \dots, p$).

Data set variables can be distinguished by their role in the models as independent and dependent variables. **Independent variables** (or **explanatory variables**, **predictor variables**) are those variables assumed to be capable of taking part of a function suitable to model the response variable. **Dependent variables** (or **response variables**) are variables (often obtained from experimental measures) for which the interest is to find a statistical dependence on one or more independent variables. Independent variables constitute the data matrix **X**, whereas dependent variables are collected into a matrix **Y** with n rows and r columns ($r = 1$ when only one response variable is defined) (Figure D2). Moreover, additional information about the belonging of objects to different classes can be stored in a class vector **c**, which consists of integers from 1 to G ; each integer indicates a class and G is the total number of classes.

→ *Regression analysis* is the methodology searching for mathematical models describing relationships between a set of independent variables and a response variable **y** (or a set of response variables **Y**), whereas → *classification* is the methodology searching for mathematical models describing relationships between a set of independent variables and the classification vector **c**, able to assign each object to its proper class.

$$\mathbf{X} = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & & \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & & x_{np} \end{vmatrix} \quad \mathbf{Y} = \begin{vmatrix} y_{11} & y_{12} & \dots & y_{1r} \\ y_{21} & y_{22} & & \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & & y_{nr} \end{vmatrix} \quad \mathbf{c} = \begin{vmatrix} c_1 \\ c_2 \\ \dots \\ c_n \end{vmatrix}$$

Figure D2 n objects described by an $\mathbf{X}$ data matrix of p independent variables, a $\mathbf{Y}$ matrix of r responses, and a $\mathbf{c}$ vector of G class assignments of the objects. $c_1, \dots, c_n$ are G integer numbers or labels representing the class assignment of the n objects.

In several cases, before applying classification methods, the class vector $\mathbf{c}$ is transformed into a set of G binary vectors by a procedure called **class unfolding**. This procedure consists in assigning each object a binary vector that is comprised of $G - 1$ values equal to 0 and one value equal to 1 corresponding to the class the object belongs to (Figure D3). In other words, class unfolding transforms the n -dimensional vector $\mathbf{c}$ into a binary matrix $\mathbf{C}$ with n rows (the objects) and G columns (the classes).

$$\mathbf{c} = \begin{vmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \end{vmatrix} \Rightarrow \mathbf{C} = \begin{vmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{vmatrix}$$

Figure D3 Example of class unfolding of 10 objects assigned to three different classes.

To estimate the predictive capabilities of a model by $\rightarrow$ *validation techniques*, the data set can be split into different parts: the **training set** (or **learning set**), that is, the set of objects used for model building, the **test set**, that is, the set of objects used to optimize the goodness of prediction of a model obtained from the training set, and the **external evaluation set** (or **evaluation set**), that is, a new data set used to perform further external validation of the model obtained from the training set.

The use of several variables in describing objects increases the complexity of the data and therefore the $\rightarrow$ *model complexity*: noise, variable correlation, redundancy of information provided by the variables, and unbalanced information and not useful information give the data an intrinsic complexity that must be resolved. This happens in the case of spectra, each constituted, for example, by 800–1000 digitalized signals, which are highly correlated variables. Usually, $\rightarrow$ *variable reduction* and $\rightarrow$ *variable selection* improve the quality of models (in particular, their predictive power) and information extracted from models. $\rightarrow$ *Chemometrics* provides several useful tools able to check the different kinds of information contained in the data [Frank and Todeschini, 1994].

Data sets can be analyzed by $\rightarrow$ *exploratory data analysis*, usually based on multivariate techniques, such as $\rightarrow$ *principal component analysis*; $\rightarrow$ *cluster analysis* allows the evaluation of similarity/diversity among the objects or, by transposing the **X** data matrix, among the variables. Similarity and diversity among the objects of a data set are encoded in the $\rightarrow$ *similarity matrix* and in the $\rightarrow$ *data distance matrix*, respectively.

- **Daylight fingerprints** $\rightarrow$ substructure descriptors ($\odot$ fingerprints)
- **Daylight-FingerPrint drug-like Score** $\rightarrow$ scoring functions ($\odot$ Property and Pharmacophore Features Score)
- **dBx descriptors** $\rightarrow$ shape descriptors
- **D/D index** $\rightarrow$ molecular geometry
- **decimal adjacency vector** $\rightarrow$ adjacency matrix
- **decomposition** $\rightarrow$ equivalence classes
- **degeneracy of molecular descriptors** $\rightarrow$ molecular descriptors
- **degradability** $\rightarrow$ environmental indices ($\odot$ persistence)
- **degree-adjacency matrix** $\equiv \chi$ *matrix* $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **degree centrality** $\rightarrow$ center of a graph
- **degree complexity** $\equiv$ *mean information content on the vertex degree magnitude* $\rightarrow$ topological information indices
- **degree distance of the graph** $\rightarrow$ Schultz molecular topological index
- **degree of unsaturation** $\rightarrow$ multiple bond descriptors
- **delocalization** $\rightarrow$ delocalization degree indices

■ delocalization degree indices

Delocalization degree indices are molecular descriptors accounting for the π -electron mobility in a molecule.

To indicate this peculiar behavior of the π -electrons in a molecule, several different terms were historically used. The term **conjugation** (or π -conjugation) referred to unsaturated hydrocarbons was introduced to indicate a molecule with alternating saturated and unsaturated bonds. Then, the term conjugation can be preferably referred to topological aspects of a molecule. The term **resonance** was derived in the framework of the *Valence Bond* (VB) theory, accordingly to which the reference structure corresponds to the most stable structure and resonance energies are obtained by taking into account the contributions of all other (less stable) resonance forms. The term **delocalization** was derived in the framework of the *Molecular Orbital* (MO) theory, according to which the wave function is composed by molecular orbitals already delocalized over the entire molecule and delocalization energies are calculated with respect to a reference system with completely localized orbitals, for example, the atomic orbitals [Bruschi, 2005].

The concept of delocalization is also closely related to the concept of **aromaticity**. In effect, the concept of aromaticity is one of the most important general concepts for an understanding of organic chemistry and physico-chemical properties [Lloyd, 1996]. The term aromaticity has a long history in the chemistry development and dates back to the first use by Kekulé, Erlenmeyer, and Körner in the 1860s [Kekulé, 1865, 1866a, 1866b; Erlenmeyer, 1866; Körner, 1869, 1874].

Aromaticity, as well as the terms of aromatic character and resonance, is associated with the ground-state properties of cyclic π -electron compounds which (a) are more stable than the chain

analogues or classical localized structures due to an energy called resonance energy, (b) have bond lengths between those typical of single and double bonds, (c) have a π -electron ring current that is induced when the molecule is exposed to external magnetic fields, leading to specific values of ^1H -NMR chemical shifts and increased values of the $\rightarrow$ *magnetic susceptibility*. Moreover, (d) from the chemical reactivity point of view, due to the tendency to maintain the π -electron structure, substitution is preferred to addition [Krygowski and Cyranski, 2001].

Several definitions of the molecule aromaticity were given, each based on one or more of the aforementioned properties of the so-called aromatic compounds.

However, for a general definition, *aromaticity* may be assumed to be a phenomenon that occurs, even though to different extents, when molecules show all the aforementioned properties.

In fact, aromaticity is an ‘excess property’, which means a deviation from a property additive scheme, and consequently quantitative measures of aromaticity need the assumption of some reference state.

Delocalization degree indices, commonly also called **resonance indices** and **aromaticity indices**, are molecular descriptors giving measures, in general, of the electron delocalization or, more specifically, of the aromatic character of compounds; these indices are of different kind, depending on different theoretical, physico-chemical, experimental, and geometrical aspects of the aromatic behavior. These can be distinguished into energy-based, geometry-based, and magnetic property-based indices.

The first theoretical explanation – based on quantum mechanical approach – of the aromatic character of a molecule was given by physical chemist E. Hückel in 1931. The **Hückel’s rule** estimates whether a planar ring molecule will have aromatic properties and was expressed as the $4n + 2$ rule formulated by von Doering in 1951. A cyclic ring molecule satisfies Hückel’s rule when the number of its π electrons equals $4n + 2$, n being zero or any positive integer.

Hückel’s rule is not valid for many compounds containing more than three fused aromatic nuclei in a cyclic fashion like in pyrene or coronene.

Strictly related to the concept of aromaticity is the resonance energy of a molecule.

The **resonance energy** RE is a theoretical quantity introduced to explain the stability of benzene and is used for predicting the electron delocalization degree of conjugated systems [Wheland, 1955; Salem, 1966; Randić, 1989; Trinajstić, 1991, 1992]. The general definition of resonance energy is

$$\text{RE} = E_{\pi}(\text{conjugated molecule}) - E_{\pi}(\text{reference structure})$$

where E_{π} is the π -electron energy.

The *resonance energy per electron* (REPE) is a size-independent quantity obtained by dividing the total resonance energy for the number of π -electrons.

Aromaticity indices defined in terms of resonance energy are commonly called **resonance indices**.

A number of resonance measures were proposed based on different theoretical quantum-chemistry approaches [Krygowski, Cyranski *et al.*, 2000; Randić, 2003a]. However, it was recognized that the main difference among the proposed approaches lies in the definition of the specific measure and the nonconjugated reference structure and not in the use of different MO theories.

The most common resonance indices are reported below.

- **Hückel resonance energy**

This is the classical definition of resonance energy obtained from a reference structure containing carbon–carbon isolated double bonds with π -electron energy of ethylene [Streitweiser, 1961]:

$$\text{HRE} = E_{\pi}(\text{conjugated molecule}) - 2 \cdot N_{\text{C}=\text{C}}$$

where $N_{\text{C}=\text{C}}$ is the number of double bonds in a Kekulé structure of the molecule.

This criterion to define the resonance energy fails in many cases, overestimating aromaticity of rather unstable compounds.

📖 [Gutman and Trinajstić, 1972]

- **Dewar resonance energy**

This is the resonance energy defined as the difference between the π -energy E_{π} of the compound and the reference energy estimated by the bond contributions of the corresponding nonconjugated structure, in the framework of SCF π -MO approximation:

$$\text{DRE} = E_{\pi}(\text{conjugated molecule}) - \sum_b n_b \cdot E_b$$

where n_b is the number of bonds having bond energy E_b [Dewar and Longuet-Higgins, 1952; Dewar and Gleicher, 1965; Dewar, 1969; Dewar and de Llano, 1969; Dewar, Kohn *et al.*, 1971].

This definition of resonance energy makes a clear distinction between aromatic (positive DRE), antiaromatic (negative DRE) and nonaromatic (near-zero DRE) conjugated molecules. Extensive tables of resonance energies were also obtained in the framework of the HMO approximation by Hess and Schaad [Hess Jr. and Schaad, 1971a, 1971b, 1973; Hess Jr., Schaad *et al.*, 1972]. Moreover, extensions and modifications of the calculation of the reference structure energy were proposed by other authors [Baird, 1969, 1971].

📖 [Dewar, Harget *et al.*, 1969; Dewar and Harget, 1970a; Dewar and Trinajstić, 1970; Schaad and Hess Jr., 1972; Hess Jr., Schaad *et al.*, 1975]

Other resonance energy indices (RE) are derived from the molecular structure and do not involve direct energy measures. Some of them were defined by fitting resonance energy values of a number of compounds, while others were derived from molecular topology.

The most popular of them are listed below.

- **Green resonance energy**

This is defined for benzenoid systems as

$$\text{GRE} = \frac{AB}{3} + \frac{AB^*}{10}$$

where AB is the total number of aromatic bonds and AB^* is the number of bonds contained in one benzene ring and linking two others [Green, 1956].

- **Bartell resonance energy**

This is a resonance energy index that relates π -energy to the Pauling $\rightarrow$ bond order P by the following expression:

$$\text{BRE} = \frac{4}{3} \times \beta \times \left(N_{\text{C}=\text{C}} - \sum_b P_b^2 \right)$$

where the summation runs over the π bonds, β is the resonance integral, and $N_{\text{C}=\text{C}}$ is the number of formal double bonds [Bartell, 1963, 1964]. The Pauling bond orders are obtained from the analysis of the Kekulé resonance structures.

- **Carter resonance energy**

This is defined for benzenoid systems as

$$\text{CRE} = 0.6 \times N_{\text{C}=\text{C}} + 1.5 \times \ln K - 1$$

where 0.6 and 1.5 are empirical parameters, $N_{\text{C}=\text{C}}$ the number of double bonds in one Kekulé structure, and K is the $\rightarrow$ *Kekulé number* of the molecule which is the number of Kekulé structures in a molecule [Carter, 1949].

- **Herndon resonance energy**

This is defined for benzenoid systems as

$$\text{HRE} = 1.185 \times \ln K$$

where 1.185 was obtained by fitting the Dewar-deLlano SCF π -MO resonance energy values and K is the $\rightarrow$ *Kekulé number* of the molecule [Herndon, 1973b, 1974b; Herndon and Ellzey Jr, 1974].

- **Wilcox resonance energy**

This is defined for general aromatic systems including alternant four-membered rings as

$$\text{WRE} = 0.445 \times \ln \text{CSC} - 0.17 \times N_4$$

where CSC is the $\rightarrow$ *corrected structure count* and N_4 is the number of four-membered rings [Wilcox Jr, 1968, 1969]. Parameter values were obtained by fitting the Hess-Schaad resonance energy values [Hess Jr., Schaad *et al.*, 1975].

- **McClelland resonance energy**

This is a measure of resonance energy defined in terms of the number of atoms A_π and bonds B_π involved in a π -system as [McClelland, 1971]:

$$\text{MCRE} = 0.92 \times \sqrt{2 \cdot A_\pi \cdot B_\pi}$$

- **Hosoya resonance energy**

This is a measure of aromatic stability of conjugated systems defined as [Hosoya, Hosoi *et al.*, 1975]

$$\text{HoRE} = \tilde{Z} - Z$$

where Z is the $\rightarrow$ *Hosoya Z index* and $\tilde{Z}$ is the $\rightarrow$ *stability index* defined as the sum of the absolute values of the coefficients c_{2i} appearing alternatively in the $\rightarrow$ *characteristic polynomial* of the adjacency matrix:

$$\tilde{Z} = \sum_{i=0}^{\lfloor A/2 \rfloor} |c_{2i}|$$

where the square brackets indicate the greatest integer not exceeding $A/2$ and A is the number of atoms.

In acyclic graphs, $\tilde{Z}$ and Z are equal, the $\rightarrow Z$ -counting polynomial being in this case coincident with the characteristic polynomial.

- **Aihara resonance energy**

This is a measure of aromatic stability of conjugated systems defined as

$$\text{ARE} = 6.0846 \cdot \log\left(\frac{Z^*}{Z}\right)$$

where Z is the $\rightarrow$ Hosoya Z index and Z^* is defined as

$$Z^* = \prod_{i=1}^A (1 + \lambda_i^2)^{1/2}$$

where λ_i denotes the eigenvalues of the $\rightarrow$ characteristic polynomial of the adjacency matrix of the molecular graph and A is the number of atoms [Aihara, 1976, 1977b, 1977a, 1978].

- **topological resonance energy**

This is a resonance energy index defined as [Gutman, Milun *et al.*, 1977; Trinajstić, 1992]:

$$\text{TRE} = E_{\pi} - E_{\text{REF}} = \sum_{i=1}^A g_i \cdot (\lambda_i - \lambda_i^{\text{REF}})$$

where E_{π} is the Hückel π -electron energy of the molecule and E_{REF} is the reference energy obtained from the $\rightarrow$ matching polynomial of the corresponding molecular graph; λ_i denotes the eigenvalues of the $\rightarrow$ characteristic polynomial of the adjacency matrix of the molecular graph and λ_i^{REF} the eigenvalues of the matching polynomial, g_i is the occupation number on the i th molecular orbital that can take values 0, 1, or 2, and A is the number of atoms.

The average TRE index is defined as TRE/N_{π} where N_{π} is the number of π electrons in the molecule.

📖 [Babic, Brinkmann *et al.*, 1997; Jurić, Nikolić *et al.*, 1997]

- **Krygowski bond energy**

Based on the concept of **Pauling's bond number** n [Pauling, 1947], n being the number of shared electron pairs in the bond, the Krygowski bond energy is defined as the sum over all the π bonds of a function of the differences of bond lengths between the reference and the actual bonds [Krygowski, Ciesielski *et al.*, 1995]:

$$\text{KBE} = \sum_{b=1}^{AB} E(n)_b = \sum_{b=1}^{AB} 87.99 \cdot \exp[2.255 \cdot (1.533 - R(n)_b)]$$

where AB is the number of π bonds, $E(n)$ is the energy of a bond with a Pauling's bond number n , 87.99 is the bond energy of a single C–C bond ($E(1)$), and $R(n)$ is the length of the bond with bond number n . This equation was derived from the empirical relationship that combines bond energy and bond number:

$$E(n) = E(1) \cdot n^k$$

where k is an empirical constant.

Table D1 Resonance indices for some benzenoid compounds.

Compound	K	Dewar	Green	Aihara	Herndon	TRE
Benzene	2	0.869	2.00	0.273	0.821	0.276
Naphthalene	3	1.323	3.67	0.389	1.302	0.390
Anthracene	4	1.600	5.33	0.475	1.643	0.476
Phenanthrene	5	1.933	5.43	0.546	1.907	0.546
Pyrene	6	2.098	6.43	0.562	2.123	0.592
Naphthacene	5	1.822	7.00	0.553	1.907	0.558
3,4-Benzophenanthrene	8	2.478	—	0.687	2.464	—
1,2-Benzanthracene	7	2.291	7.10	0.643	2.306	—
Chrysene	8	2.483	7.20	0.688	2.464	0.684
Triphenylene	9	2.654	7.30	0.739	2.604	—
Perylene	9	2.619	8.20	0.598	2.604	—

Data from [Swinborne-Sheldrake, Herndon *et al.*, 1975; Aihara 1977b; Trinajstić 1992]. K is the Kekulé number.

Several aromaticity indices were defined in terms of bond lengths r_b and bond orders π_b , exploiting the typical structural features of aromatic compounds [Krygowski, Ciesielski *et al.*, 1995]. The most known aromaticity indices of this kind are listed below and some values provided in Table D3.

• Julg–François index (A_j)

Based on the idea that bond alternation causes the aromatic character to decrease, the first aromaticity index based on the $\rightarrow$ *molecular geometry* was proposed by Julg and François [Julg and François, 1967] as

$$A_j = 1 - \frac{225}{AB} \times \sum_{b=1}^{AB} \left(1 - \frac{r_b}{\bar{r}_\pi} \right)^2$$

where AB is the number of π peripheral bonds involved in the aromatic system, r_b the geometric distance of the considered π bond, and $\bar{r}_\pi$ is the average π bond length. The constant 225 results from the normalization conditions to obtain a value of 0 for the Kekulé structure of benzene and 1 for any system with all bonds of equal length.

• bond alternation coefficient (BAC)

This is a purely geometric aromaticity index defined as

$$BAC = \sum_b (r_{b+1} - r_b)^2$$

where r_{b+1} and r_b are consecutive bond lengths in the rings; the summation runs over all π bonds of the molecule (or fragment); note that the sequence of bonds on which the sum runs is obviously well defined for monocyclic systems but less defined for polycyclic systems [Binsch and Heilbronner, 1968].

• **Bird aromaticity indices** ($\equiv I_R$ aromaticity indices)

These are general aromaticity indices based on the statistical degree of uniformity of the bond orders of the ring periphery and distinguished in I_5 , I_6 , and $I_{5,6}$ for five-, six-membered rings, and five-, six-fused rings, respectively [Bird, 1985, 1986]. These indices are defined as

$$I_R = 100 \cdot \left(1 - \frac{V}{V_R} \right)$$

where V_R is a constant depending on the considered ring (e.g., $V_R = 35$ for a five-membered heterocycle and $V_R = 33.3$ for a six-membered heterocycle; for systems consisting of a five-membered and a six-membered ring fused together, $V_R = 35$).

The term V is defined in terms of the bond order variance:

$$V = \frac{100}{\bar{\pi}} \cdot \sqrt{\frac{\sum_b (\pi_b - \bar{\pi})^2}{AB}}$$

where the sum runs on the π bonds, AB is the number of π bonds, $\bar{\pi}$ is the average bond order, and π_b is the $\rightarrow$ Gordy's bond order of the b bond defined as

$$\pi_b \equiv \pi_{ij} = \frac{a}{r_b^2} - b$$

where a and b are constants depending on the π bond type and r_b is the bond length.

• **RC index**

This is an aromaticity index based on the idea of *ring current* whose magnitude is determined by its weakest link in the ring [Jug, 1983, 1984]. The weakest link is considered as the bond with the minimum total bond order:

$$RC = \min_b (\pi_b)$$

where b runs over all the π bond system and π_b denotes the bond orders.

By analogy, a complementary aromatic index, called **maximum bond length** (LB) was defined as the longest bond length in the π -electron system under consideration [Krygowski and Ciesielski, 1995; Krygowski, Ciesielski *et al.*, 1995]:

$$LB = \max_b (r_b)$$

where b runs over all the π bonds and r_b denotes the π bond lengths.

• **HOMA index** ($\equiv$ Harmonic Oscillator Model of Aromaticity index)

This index is based on the degree of alternation of single/double bonds, measuring the bond length deviation from the optimal length attributed to the typical aromatic state [Kruszewski and Krygowski, 1972; Krygowski, 1993; Krygowski and Ciesielski, 1995; Krygowski, Ciesielski *et al.*, 1995; Krygowski, Cyranski *et al.*, 1996]. The *HOMA* index is defined as:

$$HOMA = \frac{1}{n_k} \cdot \sum_k \left[1 - \frac{\alpha_k}{AB_k} \cdot \sum_{b=1}^{AB_k} (r_k^{opt} - r_b)^2 \right]$$

where the first summation runs over n_k aromatic bond types, AB_k is the number of π bonds of the k -th aromatic bond type, r_b is the actual bond length, α_k and r_k^{opt} denote a numerical constant and the typical aromatic bond length referring to the k -th aromatic bond type (Table D2), respectively.

The *HOMA* index for the k th aromatic bond type can be decomposed in two terms describing two different contributions to a decrease in aromaticity; one contribution, due to the bond elongation, is called *EN* and the other one, due to the bond length alternation, is called *GEO*:

$$\begin{aligned} HOMA_k &= 1 - \frac{\alpha_k}{AB_k} \cdot \sum_{b=1}^{AB_k} (r_k^{opt} - r_b)^2 = 1 - \left[\alpha_k \cdot (r_k^{opt} - \bar{r}_k)^2 + \frac{\alpha_k}{AB_k} \cdot \sum_{b=1}^{AB_k} (\bar{r}_k - r_b)^2 \right] = \\ &= 1 - EN_k - GEO_k \end{aligned}$$

Table D2 Values of parameter α and optimal bond lengths r^{opt} for different aromatic bond types.

Bond	α	r^{opt}	Bond	α	r^{opt}
C $\approx$ C ^a	257.7	1.388	C $\approx$ P	118.91	1.698
C $\approx$ C	98.89	1.397	C $\approx$ S	94.09	1.677
C $\approx$ N	93.52	1.334	N $\approx$ N	130.33	1.309
C $\approx$ O	157.38	1.265	N $\approx$ O	57.21	1.248

^a1,3-butadiene.

• **HOSE index** ($\equiv$ *Harmonic Oscillator Stabilization Energy index*)

This is an aromaticity index based on the energy deformation derived from a simple harmonic oscillator potential and defined as [Krygowski and Wieckowski, 1981; Krygowski, Anulewicz *et al.*, 1983, 1995; Bird, 1997]:

$$HOSE = 301.15 \cdot \left[\sum_{b=1}^{n_1} k'_b \cdot (r'_b - r_0^s)^2 + \sum_{b=1}^{n_2} k''_b \cdot (r''_b - r_0^d)^2 \right]$$

where the first summation runs on single bonds and the second one on double bonds; r_0^s and r_0^d denote the reference bond lengths for single and double bonds, respectively; k'_b and k''_b are the corresponding force constants of the b bond; and r'_b and r''_b stand for the lengths of π bonds in the molecule. The force constants are calculated accordingly to the relation:

$$k_b = a + b \cdot r_b$$

where a and b are two constants and r_b is the actual π bond length.

HOSE index allows estimation of aromaticity also in cases of very small changes in aromatic character of the molecule in question.

• **Dewar index** (D_p)

This is a local aromaticity index [Dewar and Longuet-Higgins, 1952; Dewar, 1969], defined in the framework of the perturbation molecular orbital (PMO) theory as the energy difference

between a conjugated radical (R) and a conjugated hydrocarbon (H) obtained by joining a new carbon atom to the $a, b, c, \dots$ atoms of the radical:

$$D_P = E(H) - E(R) = 2 \cdot (C_a + C_b + \dots)$$

where P indicates the particular site of the molecule and $C_a, C_b, \dots$, are the coefficients of the normalized nonbonding molecular orbital of the radical, corresponding to the atom $a, b, \dots$

Approximate relationships between the Dewar index and topological descriptors were derived by Gutman [Gutman, 1977].

- **benzene-likeliness index (B_L)**

This is an aromaticity index calculated from molecular topology. It is defined in terms of the first-order $\rightarrow$ *valence connectivity index* ${}^1\chi^v$ divided by the number B of bonds of the molecule (hydrogen bonds excluded), and normalized on the benzene molecule [Kier and Hall, 1986]:

$$B_L = \frac{{}^1\chi^v/B}{2/6} = 3 \times \frac{{}^1\chi^v}{B}$$

where ${}^1\chi^v = 2$ and $B = 6$ for benzene. Some typical values of this index are $B_L(\text{benzene}) = 1$, $B_L(\text{thiophene}) = 0.97$, $B_L(\text{pyrrole}) = 0.95$, $B_L(\text{pyridine}) = 0.93$, and $B_L(\text{imidazole}) = 0.86$.

Other aromaticity indices were derived from the magnetic characteristics of aromatic compounds, such as exaltation of diamagnetic susceptibility and ${}^1\text{H}$ -NMR shielding.

The most important are discussed below.

- **diamagnetic susceptibility exaltation (Λ)**

It is a parameter that exploits the enhanced diamagnetic anisotropy (due to the “ring current”) of benzenoid aromatic systems with respect to the noncyclic delocalized counterpart and can be calculated both experimentally and theoretically. It is defined as the difference between the $\rightarrow$ *magnetic susceptibility* of the aromatic system M and that one of the reference structure R [Dauben, Wilson *et al.*, 1968]:

$$\Lambda_M = \chi_M - \chi_R$$

Positive values reveal the presence of ring currents and therefore of benzenoid aromatic systems. The diamagnetic susceptibility exaltation highly depends on the ring size.

Significant relationships between aromaticity measured by diamagnetic susceptibility exaltation and derivatives of the total molecular valence for molecule and HOMO orbital, calculated *ab initio* in the framework the $\rightarrow$ *Density Function Theory*, were also found [Balawender, Komorowski *et al.*, 1998].

- **NICS index ($\equiv$ Nucleus-Independent Chemical Shift index)**

The NICS index was defined as the negative value of the absolute shielding obtained by ${}^1\text{H}$ -NMR spectroscopy and computed at a ring center or at some other interesting point of the system [Schleyer, Maerker *et al.*, 1996; Krygowski and Cyranski, 2001]. Rings with negative NICS values are defined as aromatic, and the more negative NICS more aromatic the rings are. Consequently, antiaromatic systems show positive NICS values. NICS values are sensitive to the basis set used for calculation; it has been suggested to use NICS values obtained from the (6-31 + G*) basis set, whenever it is possible.

Some correlation between NICS index and *HOMA* index was found out [Szatyłowicz, Krygowski *et al.*, 2007]. However, criticism about the use of NICS as aromaticity descriptor is made by Lazzeretti [Lazzeretti, 2004], concluding that “... *a quantitative theory of aromaticity based on NICS is epistemologically inconsistent.*”

• Delocalization Index (DI)

The degree of π -delocalization between atoms A and B was quantified in the framework of the $\rightarrow$ AIM theory, by the delocalization index DI_{AB} , which is defined as [Poater, Fradera *et al.*, 2003a, 2003b; Poater, Garcia-Cruz *et al.*, 2004; Krygowski, Ejsmont *et al.*, 2004]

$$DI_{AB} = -2 \cdot \iint_A^B \Gamma_{\text{exc}}(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2$$

where Γ_{exc} is the exchange-correlation density over the basins of atoms A and B, which are defined from a condition of zero-flux gradient.

Moreover, the mean of all DI of *para*-related carbons in a given six-membered ring, called **Para-Delocalization Index (PDI)**, has been defined as an aromaticity criterion based on electron delocalization. This index is closely related to NICS and *HOMA* indices: the higher the PDI indices the higher the absolute value of NICS and the higher the *HOMA* values, thus reflecting greater aromaticity.

Table D3 Values of aromaticity indices for some compounds.

Compound	<i>HOMA</i>	BAC	A_j	I_6	<i>LB</i>	Δ	NICS ^a
Benzene	1.000	0.000	1.000	100.0	1.397	13.7	−9.7
Naphthalene	0.802	0.088	0.932	81.3	1.424	30.5	−9.9
Anthracene	0.696	0.098	0.889	79.2	1.446	48.6	−13.3/−8.2
Phenanthrene	0.727	0.101	0.878	77.1	1.465	46.2	−10.2/−6.5
Pyrene	0.728	0.094	0.916	80.1	1.438	57.3	—
3,4-Benzophenanthrene	0.671	0.107	—	73.8	1.460	—	—
Triphenylene	0.722	0.070	0.906	85.1	1.478	—	—

Data from [Dauben, Wilson *et al.* 1968; Krygowski, Ciesielski *et al.* 1995; Krygowski and Cyranski 2001; and Randić, 2003a].

^aWhen double, NICS data refer to values relative to the central ring and outer rings, respectively.

📖 Additional references are listed in the thematic bibliography (under *chemical compound classes: conjugated systems*).

- **Delocalization Index** $\rightarrow$ delocalization degree indices
- **delocalized effect** $\equiv$ *resonance effect* $\rightarrow$ electronic substituent constants
- **delta matrix** $\rightarrow$ distance-path matrix
- **delta number** $\rightarrow$ distance-path matrix
- **Dennis similarity coefficient** $\rightarrow$ similarity/diversity (☉ Table S9)

- **dense matrices** → algebraic operators ($\odot$ sparse matrices)
- **dense Wiener matrix** $\equiv$ *path-Wiener matrix* → Wiener matrix
- **density** → physico-chemical properties
- **Density Functional Theory** → quantum-chemical descriptors
- **density index** → adjacency matrix
- **Density Of States** → quantum-chemical descriptors ($\odot$ EIM descriptors)
- **Depczynski fitness function** → regression parameters
- **dependent variables** → data set
- **Description, Acquisition, Retrieval and Computer-aided design** → DARC/PELCO analysis
- **desolvation energy fields** → molecular interaction fields
- **$\det|A + D|$ index** → determinant-based descriptors
- **$\det|A|$ index** → determinant-based descriptors
- **$\det|D|$ index** → determinant-based descriptors
- **determinant** → algebraic operators

■ determinant-based descriptors

These are molecular descriptors defined in terms of the → *determinant* of a matrix representing a → *molecular graph*. Molecular descriptors similar to the determinant-based descriptors are also calculated by using → *permanent* and → *hafnian* of any matrix representing a molecular graph, such as → *per(D) index*, → *shaf(D) index*, and → *lhaf(D) index* [Schultz, Schultz *et al.*, 1992; Schultz and Schultz, 1992]. Moreover, still based on the determinant is the → *characteristic polynomial* of a molecular matrix, which plays a very important role in the calculation of molecular descriptors.

The most popular determinant-based descriptors are discussed below.

• $\det|A|$ index

The determinant of the → *adjacency matrix* **A**. It was observed that this determinant often equals zero and this is a necessary and sufficient condition for the presence of nonbonding molecular orbitals in Hückel theory. The actual numerical value of $\det|A|$ is correlated to the thermodynamic stability of the molecule [Graovac and Gutman, 1978, 1979; Trinajstić, 1992; Gutman and Vidović, 2002a].

• $\det|D|$ index

The determinant of the → *distance matrix* **D**. In an isomeric series, this index takes same values for all compounds, alternating from negative values for isomers with an even number of nonhydrogen atoms to positive values for compounds with an odd number of nonhydrogen atoms [Schultz, Schultz *et al.*, 1990, 1993]. For acyclic alkanes with a number *A* of atoms, the following relation holds [von Knop, Müller *et al.*, 1991]:

$$\det|D| = -(-2)^{A-2} \cdot (A-1)$$

• $\det|A + D|$ index

It is the determinant of the → *adjacency-plus-distance matrix*, which is the matrix resulting from the sum of the → *adjacency matrix* **A** and the → *distance matrix* **D** of a → *H-depleted molecular graph*; this matrix is also used to calculate the → *Schultz molecular topological index* [Schultz, Schultz *et al.*, 1990, 1993; von Knop, Müller *et al.*, 1991]. Demonstrated to be more discriminant

than previously described determinant-based descriptors, the absolute value of this index increases with the size of the molecules, negative for molecules with an even number of nonhydrogen atoms and positive for those with an odd number. Moreover, it decreases with increasing branching in an isomeric series of compounds, that is, the degree of substitution increases.

The logarithm of this index was used to model different $\rightarrow$ *physico-chemical properties* [Cash, 1995c], showing that it can suffer from the drawback that the determinant values are often equal to zero.

The determinant of the 3D adjacency-plus-geometry matrix (${}^b\mathbf{A} + \mathbf{G}$) was also proposed as a molecular topographic descriptor; ${}^b\mathbf{A}$ is the 3D adjacency matrix or $\rightarrow$ *bond length-weighted adjacency matrix* whose entries corresponding to bonded atoms are $\rightarrow$ *bond distances* instead of 1 and $\mathbf{G}$ the $\rightarrow$ *geometry matrix* [Mihalić, Nikolić *et al.*, 1992].

• **general a_N -index (GAI)**

A topological index defined as the absolute value of the determinant of the **orbital interaction matrix of linked atoms (OIMLA)** [Xu, Wang *et al.*, 1992a, 1992b; Xu, 1992]:

$$GAI = |\det(\mathbf{OIMLA})|$$

OIMLA is a symmetric $\rightarrow$ *weighted adjacency matrix* of dimension $2B \times 2B$ whose diagonal elements are the relative energies of the atomic hybrid orbitals (setting at zero the energy of the C_{sp3} orbital) and the off-diagonal elements represent the interaction type of hybrid orbitals, assumed to be proportional to the corresponding overlap integrals. This matrix is derived from a $\rightarrow$ *H-depleted molecular graph* called **orbital interaction graph of linked atoms (OIGLA)**, which is a $\rightarrow$ *directed graph* where arcs (ordered pairs of vertices) are used to describe interactions between hybrid orbitals. Entries equal to zero indicate that no interaction between hybrid orbitals is considered. Both atomic hybrid orbital energies and overlap integrals are obtained by methods of $\rightarrow$ *computational chemistry*.

GAI was found to be a useful index for the discrimination of *cis/trans* isomerism ($\rightarrow$ *cis/trans descriptors*) and to model the chromatographic behavior of phosphorus derivatives.

GAI is an extension to molecules containing heteroatoms and/or multiple bonds of the a_N -index, previously defined only for alkane derivatives. The a_N -index was calculated as the absolute value of the constant term of the characteristic polynomial of **OIMLA** where diagonal entries are zero and off-diagonal entries are calculated in a similar way as for GAI [Yang and Kiang, 1983].

Note that $\rightarrow$ *graph of atomic orbitals* (GAO) is another representation of molecules that accounts for atom orbitals.

📖 [McClelland, 1974; Kiang, 1980, 2008; Graovac, Juvan *et al.*, 1999; Morón, Campillo *et al.*, 2000]

- **detour complement index** $\rightarrow$ detour matrix
- **detour complement matrix** $\rightarrow$ detour matrix
- **detour-delta matrix** $\rightarrow$ detour matrix
- **detour distance** $\rightarrow$ detour matrix
- **detour-distance combined matrix** $\rightarrow$ detour matrix

- **detour distance–topological distance combined matrix** $\equiv$ *detour-distance combined matrix* $\rightarrow$ detour matrix
- **detour/distance quotient matrix** $\rightarrow$ detour matrix
- **detour distance–geometric distance combined matrix** $\rightarrow$ matrices of molecules ($\odot$ Table M3)
- **detour distance/geometric distance quotient matrix** $\rightarrow$ matrices of molecules ($\odot$ Table M2)
- **detour distance–resistance distance combined matrix** $\rightarrow$ matrices of molecules ($\odot$ Table M3)
- **detour distance/resistance distance quotient matrix** $\rightarrow$ matrices of molecules ($\odot$ Table M2)
- **detour distance–topographic distance combined matrix** $\rightarrow$ matrices of molecules ($\odot$ Table M3)
- **detour distance/topographic distance quotient matrix** $\rightarrow$ matrices of molecules ($\odot$ Table M2)
- **detour distance–topological distance combined matrix** $\rightarrow$ matrices of molecules ($\odot$ Table M3)
- **detour distance/topological distance quotient matrix** $\equiv$ *detour/distance quotient matrix* $\rightarrow$ detour matrix
- **detour index** $\rightarrow$ detour matrix

■ detour matrix (Δ)

The detour matrix Δ of a graph G (or **maximum path matrix**) is a square symmetric $A \times A$ matrix, A being the number of graph vertices, whose entry i – j is the length of the longest path from vertex v_i to vertex v_j ($^{\max}p_{ij}$), [Harary, 1969a; Buckley and Harary, 1990; Ivanciuc and Balaban, 1994b; Amić and Trinajstić, 1995; Trinajstić, Nikolić *et al.*, 1997; Randić, DeAlba *et al.*, 1998]:

$$[\Delta]_{ij} = \begin{cases} \Delta_{ij} = |^{\max}p_{ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The length of the longest path between the vertices v_i and v_j is the maximum number of edges that separate the two vertices and is called **detour distance** and denoted as Δ_{ij} .

This definition is exactly the “opposite” of the definition of the $\rightarrow$ *distance matrix* whose off-diagonal elements are the lengths of the shortest paths between the considered vertices. However, the distance and detour matrices coincide for acyclic graphs, there being only one path connecting any pair of vertices.

The maximum value entry in the i th row is called **atom detour eccentricity** $^{\Delta}\eta_i$ (also **vertex path eccentricity** or simply **path eccentricity**):

$$^{\Delta}\eta_i = \max_j([\Delta]_{ij})$$

From the distribution of the element values in the i th row of the detour matrix, the **maximum path degree sequence** of the i th vertex is derived as a local vector-descriptor defined as

$$\{n_{i0}, n_{i1}, \dots, n_{im}, \dots, n_{ik}\}$$

where n_{im} is the number of vertices in the molecular graph located at a detour distance equal to m from the vertex v_i , k is the maximum detour distance in the graph ($^{\max}\Delta$), and n_{i0} is equal to 1 by definition. Analogously, from the distribution of the element values in the upper or lower

triangle of the detour matrix, the **maximum path frequency sequence** is derived as a molecular vector-descriptor defined as

$$\{\Delta F_0, \Delta F_1, \dots, \Delta F_m, \dots, \Delta F_k\}$$

where ΔF_m is the number of detour distances equal to m in the molecular graph; obviously, ΔF_0 equals the number of vertices in the graph.

The **maximum path sum** of the i th vertex, denoted by $MPVS_i$, is a local vertex invariant defined as the sum of the lengths of the longest paths between vertex v_i and any other vertex in the molecular graph, that is,

$$MPVS_i = VS_i(\Delta) = \sum_{j=1}^A [\Delta]_{ij}$$

where VS is the $\rightarrow$ row sum operator.

$A \rightarrow$ Wiener-type index, originally called **MPS topological index** [Ivanciuc and Balaban, 1994b] but usually known as **detour index** and denoted by w [Amić and Trinajstić, 1995; Lukovits, 1996b; Lukovits and Razinger, 1997], was proposed as the sum of the detour distances between any two vertices in the molecular graph. It is calculated as

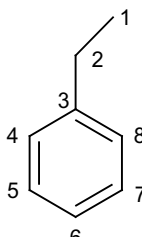
$$w \equiv Wi(\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A MPVS_i$$

where Wi is the $\rightarrow$ Wiener operator and $MPVS_i$ is the maximum path sum of the i th vertex.

Other molecular descriptors are derived from the $\rightarrow$ detour polynomial.

Example D1

Detour matrix Δ for the H-depleted molecular graph of ethylbenzene. $\Delta\eta_i$ is the atom detour eccentricity, $MPVS_i$ is the maximum path sum of the i th vertex, and w is the detour index.

		Atom	1	2	3	4	5	6	7	8	$MPVS_i$	Atom	$\Delta\eta_i$
	$\Delta =$	1	0	1	2	7	6	5	6	7	34	1	7
		2	1	0	1	6	5	4	5	6	28	2	6
		3	2	1	0	5	4	3	4	5	24	3	5
		4	7	6	5	0	5	4	3	4	34	4	7
		5	6	5	4	5	0	5	4	3	32	5	6
		6	5	4	3	4	5	0	5	4	30	6	5
		7	6	5	4	3	4	5	0	5	32	7	6
		8	7	6	5	4	3	4	5	0	34	8	7

$$w = \frac{1}{2} \times (34 + 28 + 24 + 34 + 32 + 30 + 32 + 34) = 124$$

For edge-weighted graphs, the **weighted detour matrix** (or **edge-weighted detour matrix**), denoted as ${}^w\Delta$, was proposed [Nikolić, Trinajstić *et al.*, 1996a]. The off-diagonal i - j entry is defined as the maximum path weight, that is, the maximum sum of edge weights along the

path between the vertices v_i and v_j , which is not necessarily the longest possible path between them.

A modified detour matrix was proposed by substituting diagonal zero elements with the length of the longest path from each vertex to itself (i.e., the size of the cycle containing the considered vertex). From this modified matrix, the same molecular descriptors defined above can be calculated [Rücker and Rücker, 1998].

The **detour-path matrix**, denoted as Δ_P , is a $\rightarrow$ *combinatorial matrix* analogously defined as the $\rightarrow$ *distance-path matrix* D_P ; it is a square symmetric matrix $A \times A$ whose off-diagonal entry $i-j$ is the count of all paths of any length m ($1 \leq m \leq \Delta_{ij}$) that are included within the longest path from vertex v_i to vertex v_j (Δ_{ij}) [Diudea, 1996a]. The diagonal entries are zero.

Each entry $i-j$ of the detour-path matrix is calculated from the detour matrix Δ as the following:

$$[\Delta_P]_{ij} = \binom{\Delta_{ij} + 1}{2} = \frac{\Delta_{ij}^2 + \Delta_{ij}}{2}$$

that is, as all the possible combinations of two elements taken from $\Delta_{ij} + 1$ elements (binomial coefficient).

The **hyperdetour index** ww can be obtained by applying the $\rightarrow$ *Wiener operator* Wi to the detour-path matrix as

$$ww \equiv Wi(\Delta_P) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta_P]_{ij}$$

or to the symmetric $\rightarrow$ *Cluj-detour matrix* as

$$ww \equiv Wi(\mathbf{SCJ}\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{SCJ}\Delta]_{ij}$$

For acyclic graphs, the hyperdetour index ww is equal to the $\rightarrow$ *hyper-distance-path index* D_P obtained from the distance-path matrix D_P and to the $\rightarrow$ *hyper-Wiener index* WW obtained from the $\rightarrow$ *Wiener matrix*.

The **detour-delta matrix**, denoted as Δ_Δ , is another $\rightarrow$ *combinatorial matrix* derived as the difference between the $\rightarrow$ *detour-path matrix* Δ_P and the $\rightarrow$ *detour matrix* Δ [Janežič, Miličević et al., 2007]:

$$\Delta_\Delta = \Delta_P - \Delta$$

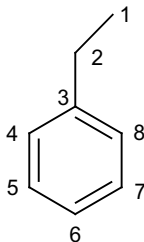
This matrix is a square symmetric matrix of dimension $A \times A$, A is the number of graph vertices, and enumerates the number of all longest paths larger than unity between vertices v_i and v_j in a graph: matrix entries are defined as the following binomial coefficients:

$$[\Delta_\Delta]_{ij} = \begin{cases} \binom{\Delta_{ij}}{2} = \frac{\Delta_{ij} \cdot (\Delta_{ij} - 1)}{2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where Δ_{ij} is the $\rightarrow$ *detour distance* between vertices v_i and v_j .

Example D2

Detour-path matrix and detour-delta matrix for the H-depleted molecular graph of ethylbenzene; ww is the hyperdetour index and $Wi(\Delta_{\Delta})$ is the $\rightarrow$ Wiener-type index derived from the detour-delta matrix.



detour-path matrix									detour-delta matrix								
Atom	1	2	3	4	5	6	7	8	Atom	1	2	3	4	5	6	7	8
1	0	1	3	28	21	15	21	28	1	0	0	1	21	15	10	15	21
2	1	0	1	21	15	10	15	21	2	0	0	0	15	10	6	10	15
3	3	1	0	15	10	6	10	15	3	1	0	0	10	6	3	6	10
4	28	21	15	0	15	10	6	10	4	21	15	10	0	10	6	3	6
5	21	15	10	15	0	15	10	6	5	15	10	6	10	0	10	6	3
6	15	10	6	10	15	0	15	10	6	10	6	3	6	10	0	10	6
7	21	15	10	6	10	15	0	15	7	15	10	6	3	6	10	0	10
8	28	21	15	10	6	10	15	0	8	21	15	10	6	3	6	10	0

$$ww = 2 \times 1 + 3 + 3 \times 6 + 7 \times 10 + 9 \times 15 + 4 \times 21 + 2 \times 28 = 368$$

$$Wi(\Delta_{\Delta}) = 1 + 3 \times 3 + 7 \times 6 + 9 \times 10 + 4 \times 15 + 2 \times 21 = 244$$

From the detour matrix and the distance matrix, a combined matrix, called **detour–distance combined matrix** $\Delta \wedge D$ (or **maximum–minimum path matrix** or **detour distance–topological distance combined matrix**), is defined as [Ivanciuc and Balaban, 1994b]

$$[\Delta \wedge D]_{ij} = \begin{cases} \Delta_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij} & \text{if } i > j \end{cases}$$

This is a square unsymmetrical $A \times A$ matrix, where the upper triangle of the matrix contains the elements of the detour matrix (information about the longest paths) and the lower triangle contains the elements of the topological $\rightarrow$ distance matrix (information about the shortest paths).

The **maximum–minimum path sum** of the i th vertex, denoted by $MmPVS_i$, is a local vertex invariant defined as the sum of the lengths of the longest and shortest paths between vertex v_i and any other vertex in the molecular graph. It is calculated as the sum of the elements in the $\Delta \wedge D$ matrix row and column corresponding to the i th vertex or, alternatively, as the sum of the $\rightarrow$ vertex distance degree σ_i calculated from the distance matrix D and the maximum path

sum $MPVS_i$ of the i th vertex calculated from the detour matrix Δ :

$$MmPVS_i = VS_i(\Delta \wedge D) + CS_{j=i}(\Delta \wedge D) = \sum_{j=1}^A [\Delta \wedge D]_{ij} + \sum_{i=1}^A [\Delta \wedge D]_{ij} = MPVS_i + \sigma_i$$

where VS_i and CS_j are the row sum and column sum operators, respectively.

A combined molecular index, called **detour–Wiener combined index** (or **MmPS topological index**) and denoted as $w \wedge W$, is defined as the sum of the lengths of the longest and shortest paths between any two vertices in the molecular graph and is calculated from the detour–distance combined matrix as

$$w \wedge W = \sum_{i=1}^A \sum_{j=1}^A [\Delta \wedge D]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A MmPVS_i = w + W$$

where w is the detour index, that is, the sum of the lengths of the longest paths in the graph, and W is the $\rightarrow$ *Wiener index*, that is, the sum of the lengths of the shortest paths.

It must be noted that for acyclic graphs the following relation holds:

$$w = W = (w \wedge W)/2$$

The transpose of the detour–distance combined matrix is the **distance–detour combined matrix** $D \wedge \Delta$ (or **minimum–maximum path matrix**, or **topological distance–detour distance combined matrix**), defined as [Janežič, Miličević *et al.*, 2007]

$$[D \wedge \Delta]_{ij} = \begin{cases} d_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ \Delta_{ij} & \text{if } i > j \end{cases}$$

Note that several molecular descriptors derived from this matrix, such as the $\rightarrow$ *spectral indices* and $\rightarrow$ *Wiener-type indices*, are the same as those from the detour–distance combined matrix, because eigenvalues of detour–distance matrix and distance–detour matrix and total sum of their matrix elements coincide.

The **distance/detour quotient matrix** (or **topological distance/detour distance quotient matrix**), denoted as D/Δ , is also derived from detour and distance matrices but it is a square symmetric matrix $A \times A$ whose off-diagonal entries are the ratio of the lengths of the shortest over the longest path between any pair of vertices [Randić, 1997c]. It is defined as

$$[D/\Delta]_{ij} = \begin{cases} \frac{d_{ij}}{\Delta_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} and Δ_{ij} are the topological and detour distances between vertices v_i and v_j , respectively. Some local and graph invariants can be calculated from this matrix. The row sums were proposed as local invariants showing a high discriminatory ability; branching vertices tend to have smaller row sums than bridging vertices. If the D/Δ matrix row sums of vertices belonging to single rings (or cycles) in the molecule are summed up, the **D/Δ ring indices**, which can be considered special substructure descriptors reflecting local geometrical environments in complex cyclic systems, are obtained. Moreover, the half sum of all row sums, which corresponds to the half sum of all entries of the D/Δ matrix, was proposed as an index

of $\rightarrow$ *molecular cyclicity*, showing regular variation with increase in cyclicity in graphs of the same size. It is called the **D/ Δ index** (or **Wiener sum index**) and is defined as [Randić, 1997c]:

$$D/\Delta \equiv \text{Wi}(D/\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [D/\Delta]_{ij}$$

where Wi is the $\rightarrow$ *Wiener operator*.

The D/Δ index decreases as the cyclicity of the molecule increases, so that it reaches the maximum value for the monocyclic graph C_A and the minimum for the $\rightarrow$ *complete graph* K_A , A being the number of vertices of the actual graph G_A . Therefore, a more suitable measure of molecular cyclicity was proposed as a standardized D/Δ index, called **cyclicity index** and denoted by γ [Randić, 1997c]

$$\gamma = \frac{D/\Delta(C_A) - D/\Delta(G_A)}{D/\Delta(C_A) - D/\Delta(K_A)}$$

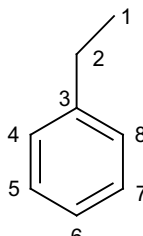
The **average cyclicity index** is calculated simply as

$$\bar{\gamma} = \frac{\gamma}{A}$$

where A is the number of graph vertices. Both the cyclicity and the average cyclicity index allow comparison of cyclic systems of different sizes; they represent the deviation of the cyclicity of the actual molecule from that of the size-corresponding monocyclic molecule C_A . Moreover, the leading eigenvalue of the distance/detour quotient matrix was proposed as another descriptor to account for cyclicity [Randić, 1997c; Pisanski, Plavšić *et al.*, 2000].

Example D3

Detour–distance combined matrix and distance/detour quotient matrix for the H-depleted molecular graph of ethylbenzene. $MmPVS_i$ is the maximum–minimum path sum of the i th vertex; VS_i and CS_j indicate the matrix row and column sums, respectively. w and W are the detour and Wiener index, respectively; $w \wedge W$ and D/Δ are the detour–Wiener combined index and the D/Δ index, respectively.

	Atom	1	2	3	4	5	6	7	8	VS_i
	1	0	1	2	7	6	5	6	7	34
	2	1	0	1	6	5	4	5	6	28
	3	2	1	0	5	4	3	4	5	24
	4	3	2	1	0	5	4	3	4	22
	5	4	3	2	1	0	5	4	3	22
	6	5	4	3	2	1	0	5	4	24
	7	4	3	2	3	2	1	0	5	20
	8	3	2	1	2	3	2	1	0	14
	CS_j	22	16	12	26	26	24	28	34	188

Atom	$MmPVS_i$		Atom	1	2	3	4	5	6	7	8	VS_i
1	56	$\mathbf{D}/\Delta =$	1	0	1	1	0.43	0.67	1	0.67	0.43	5.20
2	44		2	1	0	1	0.33	0.60	1	0.60	0.33	4.86
3	36		3	1	1	0	0.20	0.50	1	0.50	0.20	4.40
4	48		4	0.43	0.33	0.20	0	0.20	0.50	1	0.50	3.16
5	48		5	0.67	0.60	0.50	0.20	0	0.20	0.50	1	3.67
6	48		6	1	1	1	0.50	0.20	0	0.20	0.50	4.40
7	48		7	0.67	0.60	0.50	1	0.50	0.20	0	0.20	3.67
8	48		8	0.43	0.33	0.20	0.50	1	0.50	0.20	0	3.16
	376											

$w + W = 124 + 64 = 188$
 $w \wedge W = \frac{1}{2} \times (56 + 44 + 36 + 48 + 48 + 48 + 48 + 48) = 376/2 = 188$
 $D/\Delta = \frac{1}{2} \times (5.20 + 4.86 + 4.40 + 3.16 + 3.67 + 4.40 + 3.67 + 3.16) = 32.52/2 = 16.26$

A variant of the distance/detour quotient matrix is the **detour/distance quotient matrix** (or **detour distance/topological distance quotient matrix**), denoted by $\Delta/\mathbf{D}$, whose off-diagonal elements are the reciprocal of the distance/detour quotient matrix elements [Plavšić, Trinajstić *et al.*, 1998]:

$$[\Delta/\mathbf{D}]_{ij} = \begin{cases} \frac{\Delta_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

A Wiener type index derived from this matrix is the $\Delta/\mathbf{D}$ index defined as

$$\Delta/D \equiv Wi(\Delta/\mathbf{D}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta/\mathbf{D}]_{ij}$$

where Wi is the $\rightarrow$ Wiener operator.

Note that detour/distance and distance/detour quotient matrices do not have much sense for acyclic structures, all the elements being equal to 1, since detour and topological distances are the same.

For detour matrix, distance–detour combined matrix and detour–distance combined matrix there can also be defined the corresponding reciprocal matrices, which are the **reciprocal detour matrix** Δ^{-1} , **reciprocal distance–detour combined matrix** $\mathbf{D} \wedge \Delta^{-1}$, and the **reciprocal detour–distance combined matrix** $\Delta \wedge \mathbf{D}^{-1}$, as

$$[\Delta^{-1}]_{ij} = \begin{cases} \Delta_{ij}^{-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \quad [\mathbf{D} \wedge \Delta^{-1}]_{ij} = \begin{cases} d_{ij}^{-1} & \text{if } i < j \\ 0 & \text{if } i = j \\ \Delta_{ij}^{-1} & \text{if } i > j \end{cases} \quad [\Delta \wedge \mathbf{D}^{-1}]_{ij} = \begin{cases} \Delta_{ij}^{-1} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij}^{-1} & \text{if } i > j \end{cases}$$

All elements equal to zero are left unchanged in the reciprocal matrices. Moreover, $\rightarrow$ *Harary detour indices* are derived from the reciprocal detour matrix and $\rightarrow$ *Harary detour–distance indices* from the reciprocal detour–distance or distance–detour combined matrix.

The **detour complement matrix** ΔC for simple graphs is defined as [Janežič, Miličević *et al.*, 2007]

$$[\Delta C]_{ij} = \begin{cases} A - [\Delta]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where A is the number of atoms. The half sum of the detour complement matrix is the **detour complement index**:

$$Wi(\Delta C) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta C]_{ij}$$

where Wi is the $\rightarrow$ *Wiener operator*.

By analogy with the $\rightarrow$ *reverse Wiener matrix*, the **reverse detour matrix**, denoted as $R\Delta$, was defined as [Janežič, Miličević *et al.*, 2007]

$$[R\Delta]_{ij} = \begin{cases} {}^{\max}\Delta - [\Delta]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where ${}^{\max}\Delta$ is the length of the longest path in the graph. The half sum of the elements of this matrix is the **reverse detour index**:

$$Wi(R\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [R\Delta]_{ij}$$

where Wi is the $\rightarrow$ *Wiener operator*.

📖 [Harary, 1969a; Diudea, Pârv *et al.*, 1997a; Linert and Lukovits, 1997; Trinajstić, Nikolić *et al.*, 1997; Randić, DeAlba *et al.*, 1998]

- **detour-path matrix** $\rightarrow$ detour matrix
- **detour polynomial** $\rightarrow$ characteristic polynomial-based descriptors
- **detour–Wiener combined index** $\rightarrow$ detour matrix
- **Dewar–Grisdale approach** $\rightarrow$ electronic substituent constants ($\odot$ field/resonance effect separation)
- **Dewar–Golden–Harris approach** $\rightarrow$ electronic substituent constants ($\odot$ field/resonance effect separation)
- **Dewar index** $\rightarrow$ delocalization degree indices
- **Dewar resonance energy** $\rightarrow$ delocalization degree indices
- **DFT** $\equiv$ *Density Functional Theory* $\rightarrow$ quantum-chemical descriptors
- **DFT-based descriptors** $\rightarrow$ quantum-chemical descriptors
- **diagonal matrix** $\rightarrow$ algebraic operators
- **diagonal operator** $\rightarrow$ algebraic operators ($\odot$ diagonal matrix)
- **diamagnetic susceptibility exaltation** $\rightarrow$ delocalization degree indices
- **Dice similarity coefficient** $\rightarrow$ similarity/diversity ($\odot$ Table S9)

- **dielectric constant** → physico-chemical parameters
- **dielectric susceptibility** → physico-chemical parameters
- **difference in atomic charge-weighted surface area** → charged partial surface area descriptors
- **difference in charged partial surface area** → charged partial surface area descriptors
- **difference indices** → combined descriptors
- **difference in total charge-weighted surface area** → charged partial surface area descriptors
- **difference matrices** → matrices of molecules
- **differential connectivity indices** → combined descriptors
- **differential descriptors** → combined descriptors
- **differential Shannon's entropy** → information content
- **diffusivity** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **digraph** → graph
- **DiP descriptors** ≡ *Distance Profile descriptors* → substructure descriptors
- **dipolarity/polarizability term** → Linear Solvation Energy Relationships
- **dipole moment** → electric polarization descriptors
- **dipole moment components** → electric polarization descriptors
- **dipole polarization** → electric polarization descriptors
- **dipole term** ≡ *dipolarity/polarizability term* → Linear Solvation Energy Relationships
- **directed graph** ≡ *digraph* → graph
- **directional WHIM density** → WHIM descriptors (⊙ directional WHIM descriptors)
- **directional WHIM descriptors** → WHIM descriptors
- **directional WHIM shape** → WHIM descriptors (⊙ directional WHIM descriptors)
- **directional WHIM size** → WHIM descriptors (⊙ directional WHIM descriptors)
- **directional WHIM symmetry** → WHIM descriptors (⊙ directional WHIM descriptors)
- **disconnected graph** → graph (⊙ connected graph)
- **discrete wavelet transforms** → spectra descriptors
- **disjoint principal properties** → Principal Component Analysis
- **dispersion** → distance matrix
- **dissection of a graph** → graph
- **dissociation constant** → physico-chemical properties (⊙ equilibrium constants)
- **distance-adjacency map matrix** → biodescriptors (⊙ proteomics maps)
- **distance code centric index** → centric indices
- **distance complement/distance quotient matrix** → distance matrix
- **distance complement matrix** → distance matrix
- **distance connectivity index** ≡ *Balaban distance connectivity index*

■ distance-counting descriptors

Proposed by Clerc and Terkovich [Clerc and Terkovich, 1990], also called **start-end vectors** (or **SE-vectors**), are → *vectorial descriptors* collecting → *path counts* of different lengths relative to pairs of atom types in the → *H-depleted molecular graph*.

These descriptors are conceptually the same as the → *topological atom pairs*, the difference is that SE-vectors are based on simple atom types and between any two atom types all the paths are calculated instead of the shortest one.

To generate SE-vectors, first, all nonhydrogen atoms are assigned one or more atom types, which are defined by the chemical element of the atoms to account for heteroatoms. Moreover,

three additional atom types are considered: the generic atom type (T), which is assigned to any atom, the sp^2 -hybridized atom type ("2"), and the sp -hybridized atom type ("3"). Each atom is assigned at least one and at most three types.

Then, for each i th atom, the atomic path counts ${}^mP_i$ of length m ($m = 0, \dots, L$) are calculated; the path counts of the same m th order are summed up over all atoms to give the corresponding m th order molecular path count ${}^mP$, divided by 2 for lengths $m > 0$. L is the maximum path length considered and is usually set at a reasonable number depending on the $\rightarrow$ data set (typically set at 5).

Different vectors are obtained depending on the considered combination of atom types. For instance, the TT-vector encodes information on the paths between any two atoms in the graph, independent of their chemical type and hybridization state. This vector descriptor encodes information about branching, size, and the cyclicity of molecules. Moreover, the NT-vector consists of the number of paths of different lengths, starting from the nitrogen atoms and ending at all of the remaining atoms. Analogously, the NN-vector consists of the number of paths of different lengths, starting from a nitrogen atom and ending at the other nitrogen atoms; N2-vector and O2-vector represent the mutual position of double bonds and nitrogen or oxygen atoms in the molecule, respectively.

The first element of each SE-vector, that is, the zero-order molecular path count, corresponds to the number of occurrences of the considered graph elements: for example, the first entry in the TT-vector is the number of heavy atoms in the molecule, in the OO-vector it is the number of oxygen atoms, and in the 2T-vector the number of sp^2 -hybridized atoms. To calculate path counts of higher order, atomic path counts are summed up; however, if the atom types are the same, each path of nonzero length is counted twice and therefore the sum has to be divided by two.

The final distance-counting descriptor is obtained by chaining in an arbitrary but fixed way all of the calculated SE-vectors, such as

$$\{TT_0, TT_1, \dots, TT_L; NT_0, NT_1, \dots, NT_L; NN_0, NN_1, \dots, NN_L; \dots; 22_0, 22_1, \dots, 22_L; \dots\}$$

where each bin represent the occurrence number of paths of a given length for each combination of two atom types.

The number of bins in the final vector is

$$\text{number of bins} = \frac{n(n+1)}{2} \cdot (L+1)$$

where n is the number of different atom types and L the maximum path length. It follows that the dimension of SE-vectors increases linearly with the maximum path length and with the square of the number of atom types.

SE-vectors describe the global topology of a molecule, also taking into account the presence of heteroatoms and multiple bonds as well as their numbers and relative position. However, they do not encode information about stereochemistry of molecules.

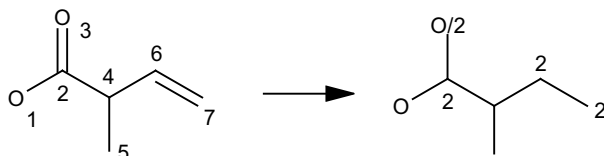
To obtain SE-vectors not depending on molecular size, a normalization scheme is introduced. The TT-vector is normalized with reference to the chain graph, by first subtracting the ${}^{CG}TT$ -vector of the chain graph and then dividing by it:

$$\left\{ \frac{TT_0 - {}^{CG}TT_0}{{}^{CG}TT_0}, \frac{TT_1 - {}^{CG}TT_1}{{}^{CG}TT_1}, \dots, \frac{TT_L - {}^{CG}TT_L}{{}^{CG}TT_L} \right\}$$

The SE-vectors other than TT-vector are normalized by dividing each entry by the total number of the corresponding atom types in the molecule, that is, the value of the first entry in the vector or the total sum of the vector entries.

Example D4

Some SE-vectors for the molecule shown below.



Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	1	1	2	2	1	0
2	1	3	2	1	0	0
3	1	1	2	2	1	0
4	1	3	3	0	0	0
5	1	1	2	3	0	0
6	1	2	2	2	0	0
7	1	1	1	2	2	0
sum	7	12	14	12	4	0
SE(TT)	7	6	7	6	2	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	1	1	2	2	1	0
3	1	1	2	2	1	0
sum	2	2	4	4	2	0
SE(OT)	2	2	4	4	2	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	1	0	1	0	0	0
3	1	0	1	0	0	0
sum	2	0	2	0	0	0
SE(OO)	2	0	1	0	0	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
2	1	3	2	1	0	0
3	1	1	2	2	1	0
6	1	2	2	2	0	0
7	1	1	1	2	2	0
sum	4	7	7	7	3	0
SE(2T)	4	7	7	7	3	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
2	1	1	1	1	0	0
3	1	1	0	1	1	0
6	1	1	1	1	0	0
7	1	1	0	1	1	0
sum	4	4	2	4	2	0
SE(22)	4	2	1	2	1	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	0	1	1	1	1	0
3	1	1	0	1	1	0
sum	1	2	1	2	2	0
SE(O2)	1	2	1	2	2	0

$$\{\text{SE(TT)}; \text{SE(2T)}; \text{SE(22)}; \dots\} \equiv \{7, 6, 7, 6, 2, 0; 4, 7, 7, 7, 3, 0; 4, 2, 1, 2, 1, 0; \dots\}$$

Modified SE-vectors were proposed simply using the shortest path between any pair of atom types in place of all the existing paths. These descriptors were called **SESP-Top vectors**, where “SP” stands for *shortest path* [Baumann, 2002a]. Moreover, to take into account not only the topological and hybridization aspects of the molecule but also the stereochemical and con-

formational information, a further development of SE-vectors was proposed exploiting the 3D spatial coordinates (x, y, z) of a molecule. For each pair of atom types, the shortest paths up to a maximum length are counted but rather than incrementing by one the corresponding bin value, each path gives a contribution equal to the ratio of the geometric distance to the topological distance. Of course, the algorithm is not applied to the first entry of each vector, the distance being equal to zero. Calculations are performed by using the $\rightarrow$ *geometric distance/topological distance quotient matrix* **G/D** instead of the $\rightarrow$ *distance matrix*. These 3D vectorial descriptors are called **SESP-Geo vectors** [Baumann, 2002a]; a different implementation is $\rightarrow$ *Distance Profile descriptors*, calculated by binning the geometric distances between pairs of atom types.

📖 [Baumann and Clerc, 1997; Baumann, Affolter *et al.*, 1997; Affolter, Baumann *et al.*, 1997; Baumann, 1999]

- **distance degree** $\rightarrow$ distance matrix
- **distance degree centric index** $\rightarrow$ centric indices

■ distance-degree matrices

Distance-degree matrices are a class of graph-theoretical matrices, which can be either vertex matrices, whose entries refer to atoms, or edge matrices, whose entries refer to bonds. In both cases, the matrix entries are defined by weighting topological distances between two graph elements (i.e., vertices or edges) by their connectivities [Ivanciuc, 1989, 1999c, 2000c].

Distance–valency matrices, denoted by **Dval**, are square ($A \times A$) matrices, A being the number of atoms, defined for vertex- and edge-weighted graphs as [Ivanciuc, 1999c, 2000c]

$$[\mathbf{Dval}(\alpha, \beta, \gamma; w)]_{ij} = \begin{cases} d_{ij}^{\alpha}(w) \cdot val_i^{\beta}(w) \cdot val_j^{\gamma}(w) & \text{if } i \neq j \\ w_i \cdot val_i^{(\beta+\gamma)}(w) & \text{if } i = j \end{cases}$$

where w is the $\rightarrow$ *weighting scheme* used to calculate vertex w_i and edge parameters w_{ij} , and val is the $\rightarrow$ *valency of the vertex* defined as the sum of the weights w_{ij} of the edges incident to the vertex v_i , $d_{ij}(w)$ is the weighted distance between vertices v_i and v_j (see $\rightarrow$ *weighted distance matrices*); α , β , and γ are exponential parameters; unsymmetrical matrices are obtained for $\beta \neq \gamma$.

For simple molecular graphs, where vertex and edge parameters are equal to 1, vertex valencies coincide with the $\rightarrow$ *vertex degrees* δ and the distance between pairs of vertices is the $\rightarrow$ *topological distance* d_{ij} ; in this case, distance–valency matrices are properly called **vertex–distance–vertex–degree matrices**, denoted by **D δ** [Janežič, Miličević *et al.*, 2007], and defined as

$$[\mathbf{D\delta}(\alpha, \beta, \gamma)]_{ij} = \begin{cases} d_{ij}^{\alpha} \cdot \delta_i^{\beta} \cdot \delta_j^{\gamma} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

Some distance–degree matrices, representing simple molecular graphs and derived from selected combinations of α , β , and γ parameters, result into other well-known graph-theoretical matrices defined in the literature. Examples are the $\rightarrow$ *distance matrix* ($\alpha = 1, \beta = 0, \gamma = 0$), the $\rightarrow$ *Harary matrix* ($\alpha = -1, \beta = 0, \gamma = 0$), and $\rightarrow$ *XI matrix* ($\alpha = 0, \beta = -1/2, \gamma = -1/2$).

The row sums of vertex–distance–vertex–degree matrices for different combinations of the three parameters are $\rightarrow$ *local vertex invariants*, proposed by Ivanciuc with the name $\rightarrow$ *VTI indices* [Ivanciuc, 1989] and extensively studied by Perdih [Perdih and Perdih, 2002a, 2002b, 2002c, 2002d, 2002e, 2003b, 2003e, 2003d].

Note. The vertex–distance–vertex–degree matrices with $\beta = 0$ were called **$v^m d^n$ matrices** by Perdih [Perdih and Perdih, 2002a] and the **general distance-degree matrix** was denoted by $G(a, b, c)$ [Perdih and Perdih, 2004], whose elements are $v_i^a v_j^b d_{ij}^c$ [Perdih and Perdih, 2003c], where v denotes the vertex degree δ and a , b , and c are the parameters corresponding to β , γ , and α , respectively. The diagonal elements of these matrices are equal to zero. Using the notations adopted in this book, the general distance-degree matrix elements are defined as the following:

$$[G(a, b, c)]_{ij} = \delta_i^a \cdot \delta_j^b \cdot d_{ij}^c$$

From $v^m d^n$ matrices, several $\rightarrow$ *branching indices* were proposed [Perdih, 2003].

$\rightarrow$ *Wiener-type indices* [Ivanciuc, 2000i] were calculated from symmetric distance–valency matrices, while $\rightarrow$ *matrix sum indices* were calculated from unsymmetrical distance–valency matrices [Ivanciuc, 1999c]. Moreover, $\rightarrow$ *characteristic polynomial-based descriptors*, $\rightarrow$ *Hosoya-type indices*, $\rightarrow$ *spectral indices*, $\rightarrow$ *hyper-Wiener-type indices*, and $\rightarrow$ *spectral moments* were derived from distance–valency matrices and tested in QSAR/QSPR modeling [Ivanciuc, 1999c, 2000c].

Edge-distance–edge-degree matrices, denoted by ${}^E D\epsilon$, are square ($B \times B$) matrices, B being the number of bonds, defined as [Janežič, Miličević *et al.*, 2007]

$$[{}^E D\epsilon(\alpha, \beta, \gamma)]_{ij} = \begin{cases} {}^e d_{ij}^\alpha \cdot \epsilon_i^\beta \cdot \epsilon_j^\gamma & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where ${}^e d_{ij}$ is the topological distance between any pair of bonds and ϵ_i and ϵ_j are their edge degrees, and α , β , and γ are exponential parameters. Similar to vertex–distance–vertex–degree matrices, unsymmetrical edge matrices are obtained for $\beta \neq \gamma$.

It must be noted that the edge-distance–edge-degree matrix of a molecular graph is the vertex–distance–vertex–degree matrix of the corresponding $\rightarrow$ *line graph*.

- **distance degree sequence** $\equiv$ *vertex distance code* $\rightarrow$ distance matrix
- **distance-delta matrix** $\equiv$ *delta matrix* $\rightarrow$ distance-path matrix
- **distance–detour combined matrix** $\rightarrow$ detour matrix
- **distance/detour quotient matrix** $\rightarrow$ detour matrix
- **distance–distance combined matrices** $\rightarrow$ molecular geometry
- **distance/distance complement quotient matrix** $\rightarrow$ distance matrix
- **distance/distance matrices** $\rightarrow$ molecular geometry
- **distance distribution moments** $\rightarrow$ distance matrix
- **distance-enhanced exponential sum connectivities** $\rightarrow$ exponential sum connectivities
- **distance exponent index** $\rightarrow$ biodescriptors ($\odot$ peptide sequences)
- **distance-extended matrices** $\equiv$ *expanded distance matrices*

■ Distance Geometry (DG)

A QSAR method proposed with the aim of automatically finding the simplest receptor binding site consistent with the binding data is based on the following assumptions: (1) binding is observed to occur on a single receptor site; (2) each ligand molecule has a well-determined chemical 3D structure and its flexibility is also taken into account; (3) no chemical modification of the molecules occurs during the binding, although their conformations may change; (4) the free energy of such a conformational change is small compared to the free energy of the binding; (5) the experimental free energy of binding is modeled by adding the interaction energies for all contact distances between parts of the ligand molecule and receptor site; and (6) the receptor site

is considered relatively rigid with respect to ligand conformational flexibility [Blumenthal, 1970; 1977, 1978, 1979, 1980, 1981, 1991; Ghose and Crippen, 1990].

In the framework of the DG method, each ligand molecule is represented as a collection of points in space, each corresponding to an atom or group of atoms, and the conformation of the molecule is described in terms of Euclidean distances between points. The matrix containing Euclidean distances between all possible pairs of points is the $\rightarrow$ *geometry matrix* of the molecule when each point corresponds to a single atom. To account for molecular flexibility, a matrix of lower bounds on the interpoint distances and a matrix of upper bounds are also defined; fixed spatial distances are represented by equal values in these matrices.

The binding site of the receptor is represented in an analogous way, that is, representing the interesting binding site regions by points and collecting their relative positions in a receptor geometry distance matrix. However, unlike molecule points, the site points may be called either “empty” or “filled”. An empty site point is a vacant place where a ligand point might be lying when binding takes place, whereas a filled site point indicates the position of receptor steric blocking groups, precluding the presence of any ligand point during binding.

The free energy of binding is calculated in a simplified all-or-nothing fashion by adding up contributions from each pair of ligand point and site point “contact.” The individual interaction energy contributions, taken from a reference list, are collected into an energy matrix where each row corresponds to a type of ligand point and each column to a type of receptor site point. If the fit between the calculated and the experimental binding free energy is not satisfactory, the interaction energy contributions or the number and/or geometry of the site points can be changed.

An extension of the distance geometry approach is given by the **Voronoi binding site models**, proposed with the aim of reducing excessive details in site model shape [Crippen, 1987; Srivastava, Richardson *et al.*, 1993]. In this approach, the receptor site is not represented by points but by nonoverlapping regions, called $\rightarrow$ *Voronoi polyhedra*, that cover the whole space. Each atom would always lie in one and only one region, that is, in a Voronoi polyhedron, and a binding mode would consist of a listing of the regions in which each atom is located [Boulu and Crippen, 1989; Boulu, Crippen *et al.*, 1990].

📖 [Gordon, 1980; Ghose and Crippen, 1982, 1983, 1984, 1985b, 1985a; Sheridan, Nilakantan *et al.*, 1986; Ghose, Logan *et al.*, 1995; Grdadolnik and Mierke, 1997; Wildman and Crippen, 2002; Imre, Veress *et al.*, 2003; Raymond and Willett, 2003; Ursu and Diudea, 2005]

➤ **distance index** $\equiv$ *distance degree* $\rightarrow$ distance matrix

■ distance matrix (D)

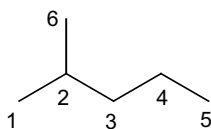
Derived from the $\rightarrow$ *H-depleted molecular graph* G , the distance matrix (viz., **vertex distance matrix** or **minimum path matrix**) summarizes in matrix form the topological distance information between all the pairs of nonhydrogen atoms in a molecule [Harary, 1964; Hakimi and Yau, 1965; Harary, 1969a; Patrinos and Hakimi, 1973; Gutman and Polansky, 1986b; Rouvray, 1986a; Buckley and Harary, 1990; Mihalić, Veljan *et al.*, 1992; Trinajstić, 1992; Hage and Harary, 1995]. The **topological distance** d_{ij} is the number of edges along the shortest $\rightarrow$ *path* $^{\min}p_{ij}$ between the vertices v_i and v_j , that is, the length of the $\rightarrow$ *geodesic* between v_i and v_j :

$$[\mathbf{D}]_{ij} \equiv d_{ij} = \begin{cases} |^{\min}p_{ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The off-diagonal entries of the distance matrix are equal to 1 if vertices v_i and v_j are adjacent (i.e., the atoms i and j are bonded and $d_{ij} = a_{ij} = 1$ where a_{ij} are elements of the $\rightarrow$ adjacency matrix A) and are greater than 1 otherwise. The diagonal elements are of course equal to zero. The distance matrix is symmetric with dimension $A \times A$, where A is the number of atoms.

Example D5

Distance matrix, vertex distance degrees σ_i , and atom eccentricities η_i of 2-methylpentane.



$D =$

Atom	1	2	3	4	5	6	σ_i	η_i
1	0	1	2	3	4	2	12	4
2	1	0	1	2	3	1	8	3
3	2	1	0	1	2	2	8	2
4	3	2	1	0	1	3	10	3
5	4	3	2	1	0	4	14	4
6	2	1	2	3	4	0	12	4

For vertex- and edge-weighted graphs, the distance matrix entry $i-j$ could be defined as the minimum sum of edge weights along the path between the vertices v_i and v_j , which is not necessarily the shortest possible path between them or otherwise as the sum of the weights of the edges along the shortest path between the considered vertices. Diagonal entries are the vertex weights. Different $\rightarrow$ weighting schemes were proposed from which a number of $\rightarrow$ weighted distance matrices were derived.

The **distance degree** (or **vertex distance degree**, **distance number**, **distance index**, **distance rank**, **distance sum**, **vertex distance sum**, **distance of a vertex**) is a local vertex invariant, denoted as σ_i , and defined as the distance matrix row sum:

$$\sigma_i \equiv VS_i(D) = \sum_{j=1}^A d_{ij}$$

where VS stands for vertex sum and is the $\rightarrow$ row sum operator, that is, the sum of the matrix elements in a row.

The distance degrees of the vertices of 2-methylpentane are shown in Example D5. For instance, the distance degree of vertex 2 in 2-methylpentane is $\sigma_2 = d_{21} + d_{23} + d_{24} + d_{25} + d_{26} = 1 + 1 + 2 + 3 + 1 = 8$.

Vertex distance degrees are $\rightarrow$ local vertex invariants: high values are observed for $\rightarrow$ terminal vertices (e.g., in 2-methylpentane, $\sigma = 12$ for terminal vertices 1 and 6, and $\sigma = 14$ for terminal vertex 5), while low values for $\rightarrow$ central vertices. Moreover, among the terminal vertices, vertex distance degrees are small if the vertex is next to a branching site (e.g., in 2-methylpentane, vertices 1 and 6 are directly bonded to vertex 2 that represents a branching site) and larger if the terminal vertex is far away (e.g., in 2-methylpentane, terminal vertex 5 is three bonds far away from the branching site 2).

The half-sum of all the elements d_{ij} of the distance matrix, which is equal to the half sum of the distance degrees σ_i of all the vertices [Harary, 1959], is the well-known $\rightarrow$ Wiener index W , which is one of the most popular topological indices used in QSAR modeling [Wiener, 1947c].

The total sum of the entries of the distance matrix is another topological index called **Rouvray index** (or **rank distance** or **total vertex distance**) and is denoted as I_{ROUV} , which is twice the Wiener index W :

$$I_{\text{ROUV}} = \sum_{i=1}^A \sum_{j=1}^A d_{ij} = \sum_{i=1}^A \sigma_i = 2W$$

For example, in 2-methylpentane, the Rouvray index, derived from distance values, is $I_{\text{ROUV}} = 10 \times 1 + 10 \times 2 + 6 \times 3 + 4 \times 4 = 64$ or, alternatively, derived from distance degrees, is $I_{\text{ROUV}} = 12 + 8 + 8 + 10 + 14 + 12 = 64$.

The average row sum of the distance matrix is a molecular invariant called **average graph distance degree**, which coincides with the average Rouvray index, defined as [Skorobogatov and Dobrynin, 1988]:

$$\bar{\sigma} = \frac{1}{A} \sum_{i=1}^A \sigma_i = \frac{2 \cdot W}{A} = \frac{I_{\text{ROUV}}}{A}$$

For example, in 2-methylpentane, the average distance degree is $\bar{\sigma} = 64/6 = 10.667$.

The **compactness**, denoted as $Comp$, was defined as another function of the Wiener index W [Doyle and Garver, 1977]:

$$\frac{1}{Comp} = \frac{4 \cdot W}{A(A-1)} = 2 \cdot \bar{W}$$

Note that the compactness is the reciprocal of twice the $\rightarrow$ mean Wiener index $\bar{W}$. The smaller the Wiener index the larger the compactness of the molecule. For example, in 2-methylpentane,

$$Comp = \frac{6 \cdot (6-1)}{4 \cdot 32} = 0.234$$

The **mean distance degree deviation**, denoted as $\Delta\sigma$, is derived from the average distance degree as [Skorobogatov and Dobrynin, 1988]:

$$\Delta\sigma = \frac{1}{A} \cdot \sum_{i=1}^A |\sigma_i - \bar{\sigma}|$$

For example, in 2-methylpentane

$$\Delta\sigma = \frac{|12-10.667| + |8-10.667| + |8-10.667| + |10-10.667| + |14-10.667| + |12-10.667|}{6} = 2$$

The minimum value of the distance degrees of the molecule atoms is another molecular invariant called **unipolarity** [Skorobogatov and Dobrynin, 1988]:

$$\sigma^* = \min_i(\sigma_i)$$

For example, in 2-methylpentane, $\sigma^* = \min\{12, 8, 8, 10, 14, 12\} = 8$.

Other molecular invariants immediately derived from distance degrees σ_i are **centralization** $\Delta\sigma^*$ [Skorobogatov and Dobrynin, 1988], **variation** $\Delta\sigma^+$, and **dispersion** σ_2^* , defined respectively as [Konstantinova and Skorobogatov, 1995]

$$\Delta\sigma^* = 2 \cdot W - A \cdot \sigma^* \quad \Delta\sigma^+ = \max_i(\sigma_i - \sigma^*) \quad \sigma_2^* = \min_i \left(\frac{1}{A} \cdot \sum_{j=1}^A d_{ij}^2 \right)$$

where W is the Wiener index, A is the number of graph vertices, σ^* is the unipolarity, and d_{ij} denotes the topological distances.

For example, in 2-methylpentane

$$\Delta\sigma^* = 64 - 6 \times 8 = 16 \quad \sigma_2^* = \min(5.667, 2.667, 2.333, 4, 7.667, 5.667) = 2.333$$

$$\Delta\sigma^+ = \max(12-8, 8-8, 8-8, 10-8, 14-8, 12-8) = \max(4, 0, 0, 2, 6, 4) = 6$$

The **PRS index** (or **Product of Row Sums index**) is defined as the product of the vertex distance degrees σ_i [Schultz, Schultz *et al.*, 1992]:

$$\text{PRS} = \prod_{i=1}^A \sigma_i \quad \text{or} \quad \log(\text{PRS}) = \log\left(\prod_{i=1}^A \sigma_i\right) = \sum_{i=1}^A \log(\sigma_i)$$

where the second expression is suggested in QSAR/QSPR modeling due to the large values that can be reached by the PRS index. This index is related to the $\rightarrow$ *permanent* of the distance matrix. For example, in 2-methylpentane

$$\begin{aligned} \log(\text{PRS}) &= 2 \times \log(12) + 2 \times \log(8) + \log(10) + \log(14) = \\ &= 2 \times 1.079 + 2 \times 0.903 + 1 + 1.146 = 6.1107 \end{aligned}$$

Still based on the vertex distance degree σ_i , **R^* indices** and **R^+ index** were defined as [Randić, Balaban *et al.*, 2001]:

$$R^* = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^\lambda \quad \lambda = \pm \frac{1}{2}; \pm 1 \quad R^+ = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot \frac{(\sigma_i + \sigma_j)}{2}$$

where a_{ij} denotes the elements of the $\rightarrow$ *adjacency matrix*, assuming value equal to 1 for pairs of adjacent vertices.

Table D4 Some molecular descriptors derived from the distance matrix for C8 data set (Appendix C – Set 1).

C8	I_{ROUV}	$\bar{\sigma}$	$\Delta\sigma$	σ^*	$\Delta\sigma^*$	$\Delta\sigma^+$	$\log(\text{PRS})$
n-octane	168	21.000	4.000	16	40	12	24.172
2M	158	19.750	3.750	15	38	12	23.694
3M	152	19.000	3.500	14	40	12	23.369
4M	150	18.750	3.313	13	46	12	23.252
3E	144	18.000	3.500	12	48	12	22.920
22MM	142	17.750	3.063	13	38	12	22.835
23MM	140	17.500	3.125	12	44	12	22.713
24MM	142	17.750	3.250	13	38	10	22.840
25MM	148	18.500	3.500	14	36	8	23.187
33MM	134	16.750	2.813	11	46	12	22.351
34MM	136	17.000	3.000	12	40	10	22.478
2M3E	134	16.750	3.250	11	46	10	22.357
3M3E	128	16.000	3.000	10	48	10	21.980
223MMM	126	15.750	2.563	11	38	10	21.881
224MMM	132	16.500	2.875	12	36	8	22.271
233MMM	124	15.500	2.625	10	44	10	21.748
234MMM	130	16.250	2.938	11	42	8	22.139
2233MMMM	116	14.500	2.250	10	36	6	21.241

I_{ROUV} , $\bar{\sigma}$, $\Delta\sigma$, σ^* , $\Delta\sigma^*$, and $\Delta\sigma^+$ are Rouvray index, average distance degree, mean distance degree deviation, unipolarity, centralization, and variation, respectively.

The maximum value entry in the i th row of the distance matrix **D** is called **atom eccentricity** (or **vertex eccentricity**) and denoted as η_i :

$$\eta_i = \max_j(d_{ij})$$

The atom eccentricity is a local vertex invariant representing the maximum distance from a vertex to any other vertex in the graph.

For example, the eccentricity of vertex 2 in 2-methylpentane is 3, which is the topological distance between vertices 2 and 5. Eccentricities of the nonhydrogen atoms of 2-methylpentane are listed in Example D5.

From the vertex eccentricity definition, a graph can be immediately characterized by two molecular descriptors known as **topological radius** R and **topological diameter** D . The topological radius of a molecule is defined as the minimum vertex eccentricity and the topological diameter is defined as the maximum vertex eccentricity, according to the following [Harary, 1969a]:

$$R = \min_i(\eta_i) \quad \text{and} \quad D = \max_i(\eta_i)$$

For example, the radius of 2-methylpentane is 2, while the diameter is 4.

Based on the combined use of topological radius and diameter is the $\rightarrow$ *graph-theoretical shape coefficient*. Moreover, simple molecular descriptors are calculated as some functions of vertex eccentricities [Konstantinova, 1996]. These are the **eccentricity** η , **average atom eccentricity** $\bar{\eta}$, and **eccentric** $\Delta\eta$ defined, respectively, as the following [Skorobogatov and Dobrynin, 1988]:

$$\eta = \sum_{i=1}^A \eta_i \quad \bar{\eta} = \frac{1}{A} \cdot \sum_{i=1}^A \eta_i \quad \Delta\eta = \frac{1}{A} \cdot \sum_{i=1}^A |\eta_i - \bar{\eta}|$$

where A is the number of graph vertices and η_i the eccentricity of the i th vertex.

Example D6

Eccentricity η , average atom eccentricity $\bar{\eta}$, and eccentric $\Delta\eta$ for 2-methylpentane.

$$\eta = 4 + 3 + 2 + 3 + 4 + 4 = 20 \quad \bar{\eta} = \frac{20}{6} = 3.333$$

$$\Delta\eta = \frac{|4-3.333| + |3-3.333| + |2-3.333| + |3-3.333| + |4-3.333| + |4-3.333|}{6} = 0.667$$

Other molecular descriptors based on atom eccentricity values combined with other $\rightarrow$ *local vertex invariants* are the $\rightarrow$ *eccentric connectivity index*, the $\rightarrow$ *eccentric distance sum*, the $\rightarrow$ *connective eccentricity index*, the $\rightarrow$ *eccentric adjacency topochemical indices*, the $\rightarrow$ *superadjacency index*, and the $\rightarrow$ *eccentricity-based Madan indices*.

From the frequencies of the row entries of the distance matrix, the **vertex distance code** (or **distance degree sequence** of a vertex, **DDS_i**) is defined as the ordered sequence of the numbers of occurrence of the different distance values for each i th vertex [Ivanciuc and

Balaban, 1999c]:

$$\mathbf{DDS}_i \equiv \{^1f_i, ^2f_i, ^3f_i, \dots, \eta_i f_i\}$$

where $^1f_i, ^2f_i, ^3f_i, \dots$, called **vertex distance counts**, indicate the frequencies of distances equal to 1, 2, 3, ..., respectively, from vertex v_i to any other vertex and η_i is the i th atom eccentricity. The vertex distance count of first-order ^1f_i coincides with the $\rightarrow$ *vertex degree* δ_i , that is, the number of first neighbors, while ^2f_i and ^3f_i correspond to the $\rightarrow$ *connection number* (i.e., number of second neighbors), and *polarity number* (i.e., number of third neighbors) for the i th vertex, respectively. Moreover, the distance degree can also be expressed in terms of vertex distance counts as

$$\sigma_i = \sum_{k=1}^{\eta_i} {}^k f_i \cdot k$$

where the sum runs over the different distance values k and η_i is the maximum distance from the i th vertex.

The total number of distances in the graph equal to k is called **graph distance count** of k th order ${}^k f$; it is obtained as

$${}^k f = \frac{1}{2} \cdot \sum_{i=1}^A {}^k f_i$$

where A is the number of graph vertices and ${}^k f_i$ is the number of distances equal to k .

The **graph distance code** is the ordered sequence of graph distance counts:

$$\mathbf{GDC} \equiv \{^1 f, ^2 f, ^3 f, \dots, {}^D f\}$$

where D is the topological diameter.

The **graph distance index** is defined as the square sum of all graph distance counts [Rouvray, 1983]:

$$GDI = \sum_{k=1}^D ({}^k f)^2$$

Example D7

Distance degree sequences of the atoms of 2-methylpentane.

Atom	$^1 f_i$	$^2 f_i$	$^3 f_i$	$^4 f_i$
1	1	2	1	1
2	3	1	1	—
3	2	3	—	—
4	2	1	2	—
5	1	1	1	2
6	1	2	1	1

For instance, the distance degree sequence of vertex 2 is $\mathbf{DDS}_2 = \{3, 1, 1\}$, which means that there are three vertices one bond away from v_2 , one vertex located at distance two from v_2 , and one vertex at distance three. The graph distance code is $\mathbf{GDC} = \{5, 5, 3, 2\}$ and the graph distance index is $GDI = 25 + 25 + 9 + 4 = 63$.

The **polarity number** p (or **Wiener polarity number**) was defined by Wiener in 1947 [Wiener, 1947c] as the number of pairs of graph vertices, which are separated by three edges. It is usually assumed that the polarity number accounts for the flexibility of acyclic structures, p being equal to the number of bonds around which free rotations can take place. Moreover, it relates to the steric properties of molecules.

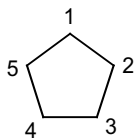
The polarity number is usually calculated from the distance matrix as the number of pairs of vertices at a topological distance equal to 3, that is,

$$p_2 = {}^3f$$

where ${}^3f$ is the graph distance count of third order, that is, count of entries equal to 3 in the upper or lower triangular submatrix of $\mathbf{D}$ [Platt, 1947, 1952]. In this case, the polarity number is denoted by p_2 to distinguish it from the original p . For acyclic graphs $p = p_2$, while in cycle-containing graphs the two numbers are generally different.

Example D8

H-depleted molecular graph, distance matrix, and polarity numbers p and p_2 for cyclopentane.



Atom	1	2	3	4	5
1	0	1	2	2	1
2	1	0	1	2	2
3	2	1	0	1	2
4	2	2	1	0	1
5	1	2	2	1	0

The polarity number p , calculated from the molecular graph, is equal to 5, which corresponds to the count of the following paths: $p_{14} = \{1, 2, 3, 4\}$, $p_{25} = \{2, 3, 4, 5\}$, $p_{31} = \{3, 4, 5, 1\}$, $p_{42} = \{4, 5, 1, 2\}$, and $p_{53} = \{5, 1, 2, 3\}$. The polarity number p_2 , calculated on the distance matrix is zero, because there are no entries equal to 3 in the distance matrix of cyclopentane.

Two other definitions of polarity number, based on the same original concept, were also introduced to have a greater discriminating ability among cyclic structures suitable for QSAR modeling purposes [Lukovits and Linert, 1998]. The polarity number p_3 was defined as the number of ways a path of length three can be laid upon the hydrogen-depleted graph. Moreover, the polarity number p_4 was defined as the number of ways a path of length 3 can be laid upon the acyclic edges of the graph (including those cases in which the second edge of the path considered coincides with a cyclic edge) $+ 1.8 \times N$, where N is the number of ways the path of length 3 can be laid upon the cyclic part of the graph; all edges not belonging to a cycle are acyclic edges and the product $1.8 \times N$ has to be rounded to yield an integer. For acyclic structures, $p = p_2 = p_3 = p_4$ (Table D5).

Table D5 Values of polarity numbers p , p_2 , p_3 , p_4 , and Wiener index W for some molecules.

Compound	p	p_2	p_3	p_4	W
Cyclopropane	0	0	3	5	3
Cyclobutane	4	0	4	7	8
Methyl-cyclopropane	2	0	3	5	8
<i>n</i> -Pentane	2	2	2	2	28
Cyclopentane	5	0	5	9	15
<i>i</i> -Propyl-cyclopentane	11	6	11	9	62
<i>n</i> -Propyl-cyclopentane	11	5	10	10	67

For example, in cyclopropane, the polarity number p is equal to zero because there are no pairs of vertices that are separated by three edges, while the polarity number p_3 is equal to 3 because there are three paths of length 3: $p_{11} = \{1, 2, 3, 1\}$, $p_{22} = \{2, 3, 1, 2\}$, $p_{33} = \{3, 1, 2, 3\}$. The polarity number p_4 is $1.8 \times 3 = 5.4$, which is rounded to 5, because the contribution of the acyclic part is zero and there are three ways the path of length 3 can be laid upon the cyclic part of the graph (in this case this number coincides with p_3).

The $\rightarrow$ distance polynomial is the characteristic polynomial of the distance matrix used to derive the $\rightarrow$ Hosoya Z' index; moreover, $\rightarrow$ eigenvalues of the distance matrix, $\rightarrow$ spectral moments, $\rightarrow$ determinant, and $\rightarrow$ permanent were proposed as molecular descriptors, even if their use in QSAR modeling is limited since they tend to reach very high values.

Topological distances, distance degrees, eccentricities, distance frequencies, topological radius, and diameter are used to calculate several $\rightarrow$ topological information indices and other molecular descriptors such as $\rightarrow$ Balaban distance connectivity index, $\rightarrow$ hyper-Wiener index, $\rightarrow$ expanded Wiener number, $\rightarrow$ expanded distance indices, $\rightarrow$ Schultz molecular topological index, $\rightarrow$ eccentricity-based Madan indices, $\rightarrow$ Petitjean shape indices, $\rightarrow$ Molecular Distance-Edge vector, $\rightarrow$ Sh indices, $\rightarrow$ Xu index, $\rightarrow$ global flexibility index, several $\rightarrow$ ID numbers, $\rightarrow$ delta number, $\rightarrow$ hyperdistance path index, some among the $\rightarrow$ triplet topological indices, $\rightarrow$ second-grade structural parameters, $\rightarrow$ steric vertex topological index, $\rightarrow$ superpendentic index, and $\rightarrow$ topological charge indices.

Moreover, the most common $\rightarrow$ autocorrelation descriptors are calculated from a molecular graph by using the topological distance as the lag.

Table D6 Some molecular descriptors derived from the distance matrix for C8 data set (Appendix C – Set 1).

C8	R	D	η	$\bar{\eta}$	$\Delta\eta$	p_2	MSD
<i>n</i> -Octane	4	7	44	5.500	1.000	5	0.463
2M	3	6	39	4.875	0.906	5	0.431
3M	3	6	38	4.750	0.813	6	0.410
4M	3	6	37	4.625	0.875	6	0.403
3E	3	5	33	4.125	0.656	7	0.381
22MM	3	5	34	4.250	0.750	5	0.380
23MM	3	5	33	4.125	0.656	7	0.371
24MM	3	5	33	4.125	0.656	6	0.377
25MM	3	5	34	4.250	0.750	5	0.398

(Continued)

Table D6 (Continued)

C8	R	D	η	$\bar{\eta}$	$\Delta\eta$	p_2	MSD
33MM	3	5	32	4.000	0.500	7	0.353
34MM	3	5	32	4.000	0.500	8	0.357
2M3E	2	4	27	3.375	0.625	8	0.349
3M3E	2	4	26	3.250	0.563	9	0.331
223MMM	2	4	27	3.375	0.625	8	0.326
224MMM	2	4	28	3.500	0.625	5	0.346
233MMM	2	4	26	3.250	0.563	9	0.319
234MMM	2	4	27	3.375	0.625	8	0.338
2233MMMM	2	3	22	2.750	0.375	9	0.295

R is the topological radius, D the topological diameter, η the eccentricity, $\bar{\eta}$ the average atom eccentricity, $\Delta\eta$ the eccentric, p_2 the polarity number, and MSD the mean square distance index.

Some graph-theoretical matrices derived from the distance matrix are reported below.

Generalized distance matrices, denoted as $\mathbf{D}^\lambda$, are derived from the distance matrix by raising its elements to an exponent λ :

$$[\mathbf{D}^\lambda]_{ij} = \begin{cases} d_{ij}^\lambda & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where λ is any real exponent.

The **distance distribution moments**, denoted as D_λ , are the moments of the distribution of topological distances d_{ij} in a molecular graph, derived from generalized distance matrices defined for positive integer λ values [Klein and Gutman, 1999]:

$$D_\lambda \equiv Wi(\mathbf{D}^\lambda) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij}^\lambda \quad \lambda = 1, 2, 3, \dots$$

where λ is a positive integer and denotes the power of the distance matrix elements and Wi is the $\rightarrow$ *Wiener operator*. Note that distance distribution moments are a subclass of the generalized $\rightarrow$ W_λ indices proposed by Gutman [Gutman, 1997; Gutman, Vidović *et al.*, 1998] and D_1 is the $\rightarrow$ *Wiener index* W .

Normalized distance distribution moments were used to define $\rightarrow$ *molecular profiles* and the second moment D_2 takes part in defining the $\rightarrow$ *hyper-Wiener index*; moreover, the index D_2 was demonstrated to be equal to half the trace of the distance matrix $\mathbf{D}$ raised to the second power [Diudea, 1996a; Diudea and Gutman, 1998].

The **mean square distance index**, denoted as MSD , is calculated from the second-order distance distribution moment as [Balaban, 1983a]:

$$MSD = \left(\frac{\sum_{i=1}^A \sum_{j=1}^A d_{ij}^2}{A \cdot (A-1)} \right)^{1/2} = \left(\frac{\sum_{k=1}^D k f \cdot k^2}{\sum_{k=1}^D k f} \right)^{1/2}$$

where $k f$ is the graph distance count of order k , D is the topological diameter, and A is the number of vertices. The same index restricted to the $\rightarrow$ *terminal vertices* (vertices of degree

one) is called **end point mean square distance index** $D1$ and is applicable only to acyclic graphs.

Both the MSD and the $D1$ indices decrease with increasing $\rightarrow$ *molecular branching* in an isomeric series.

Example D9

The square distance matrix $\mathbf{D}^2$, second-order distance distribution moment D_2 , and the mean square distance (MSD) index for 2-methylpentane.

	Atom	1	2	3	4	5	6	
$\mathbf{D}^2 =$	1	0	1	4	9	16	4	$D_2 = 5 \times 1 + 5 \times 4 + 3 \times 9 + 2 \times 16 = 84$
	2	1	0	1	4	9	1	
	3	4	1	0	1	4	4	From square distances:
	4	9	4	1	0	1	9	
	5	16	9	4	1	0	16	$MSD = \left(\frac{10 \times 1 + 10 \times 4 + 6 \times 9 + 4 \times 16}{6 \times (6-1)} \right)^{1/2}$
	6	4	1	4	9	16	0	
								$= \left(\frac{168}{30} \right)^{1/2} = 2.366$
								From distance counts:
								$MSD = \left(\frac{5 \times 1^2 + 5 \times 2^2 + 3 \times 3^2 + 2 \times 4^2}{5 + 5 + 3 + 2} \right)^{1/2}$
								$= \left(\frac{84}{15} \right)^{1/2} = 2.366$

The **reciprocal distance matrix**, denoted as $\mathbf{D}^{-1}$ (or **Harary matrix**, $\mathbf{H}$, or **vertex Harary matrix**) is a square symmetric $A \times A$ matrix derived from the distance matrix as

$$[\mathbf{D}^{-1}]_{ij} \equiv [\mathbf{H}]_{ij} = \begin{cases} 1/d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where each off-diagonal element is the reciprocal of the topological distance d between the vertices considered [Ivanciuc, Balaban *et al.*, 1993b; Plavšić, Nikolić *et al.*, 1993b; Lučić, Miličević *et al.*, 2002]; the diagonal elements are zero by definition for simple molecular graphs. The original $\rightarrow$ *Harary index* H is calculated from this matrix as an analogue of the $\rightarrow$ *Wiener index* (see below).

For vertex- and edge-weighted molecular graphs, the reciprocal distance matrix was defined as [Ivanciuc, 2000i]:

$$[\mathbf{D}^{-1}(w)]_{ij} = \begin{cases} 1/[\mathbf{D}(w)]_{ij} & \text{if } i \neq j \\ [\mathbf{D}(w)]_{ii} & \text{if } i = j \end{cases}$$

where $\mathbf{D}(w)$ is a $\rightarrow$ *weighted distance matrix* and w denotes a $\rightarrow$ *weighting scheme*.

The **reciprocal distance sum** RDS_i of the i th vertex is a local invariant defined as the sum of the reciprocal distance matrix elements in the i th row [Ivanciuc, Balaban *et al.*, 1993b]:

$$RDS_i \equiv VS_i(\mathbf{D}^{-1}) = \sum_{j=1}^A d_{ij}^{-1} \quad j \neq i$$

where the symbol VS stands for the $\rightarrow$ row sum operator. From this local vertex invariant, the **Harary index** or **RDSUM index** is derived as the half sum of RDS_i over all the vertices [Plavšić, Nikolić *et al.*, 1993b]:

$$H \equiv RDSUM \equiv Wi(D^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [D^{-1}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A RDS_i$$

where the symbol Wi refers to $\rightarrow$ Wiener operator.

Moreover, two other molecular descriptors, called **RDSQ index** and **RDCHI index**, respectively, were defined, based on a Randić-like formula [Ivanciuc, Balaban *et al.*, 1993b]:

$$RDSQ = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{1/2} \quad RDCHI = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{-1/2}$$

where A is the number of vertices and a_{ij} is equal to 1 only for pairs of adjacent vertices and zero otherwise. While the **RDSQ** index increases with both molecular size and $\rightarrow$ molecular branching, the **RDCHI** index increases with molecular size and decreases with molecular branching.

By analogy with Kier–Hall connectivity indices, **Topological Distance Connectivity Indices** (TDCIs), also called **Topological Distance Measure Connectivity Indices** (TDMCIs), were also defined by using the reciprocal distance sum RDS in place of the $\rightarrow$ vertex degree δ [Balaban, Ciubotariu *et al.*, 1990; Ciubotariu, Medeleanu *et al.*, 2004]:

$$\begin{aligned} {}^1\tau &= \sum_{i=1}^A (RDS_i)^{-1/2} \\ {}^2\tau &\equiv RDCHI = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{-1/2} \\ {}^3\tau &= \sum_{k=1}^{2P} (RDS_i \cdot RDS_l \cdot RDS_j)^{-1/2}_k \end{aligned}$$

where a_{ij} is equal to 1 only for pairs of adjacent vertices and zero otherwise. The third index ${}^3\tau$ is derived from paths of length 2 in the graph by weighting each path by the product of the reciprocal distance sums of the vertices involved in the path; ${}^2P$ is the total number of paths of length 2.

A local vertex invariant, called **generalized reciprocal distance sum**, was proposed as

$${}^\lambda RDS_i = \sum_{j=1}^A d_{ij}^{-\lambda} \quad j \neq i$$

where λ is a positive integer; the **Generalized Topological Distance Indices** (GTDIs), denoted as ${}^1\tau^\lambda$, ${}^2\tau^\lambda$, and ${}^3\tau^\lambda$, were derived from the generalized reciprocal distance sum and defined by using the same formulas as for the TDCIs [Ciubotariu, Medeleanu *et al.*, 2004]. An additional descriptor, the generalized topological distance index of order zero, was also defined as

$${}^0\tau^\lambda = \sum_{i=1}^A {}^\lambda RDS_i = \sum_{i=1}^A \sum_{j=1}^A d_{ij}^{-\lambda} \quad \lambda = 1, 2, 3, 4, \dots \quad i \neq j$$

For $\lambda = 1$, this index coincides with twice the Harary index and for $\lambda = 2$ with twice the Harary number (see below).

From the Harary matrix, a $\rightarrow$ *Balaban-like index*, called **Harary–Balaban index**^{RJ} [Nikolić, Plavšić *et al.*, 2001] or **Harary-connectivity index** [Randić and Pompe, 2001b] was also proposed as

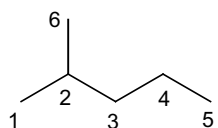
$${}^R J \equiv J(\mathbf{D}^{-1}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{-1/2}$$

where RDS_i and RDS_j are the reciprocal distance sum of vertex v_i and vertex v_j , respectively; a_{ij} are elements of the adjacency matrix equal to 1 only for pairs of bonded vertices, and zero otherwise, B is the number of edges, C the $\rightarrow$ *cyclomatic number*, and $J(\mathbf{D}^{-1})$ stands for $\rightarrow$ *Balaban-like indices*. Note that in the formula proposed by Randić–Pompe, the coefficient $B/(C+1)$ is not considered.

Moreover, the reciprocal distance sums of the graph vertices were proposed as an alternative to vertex distance degrees to detect the $\rightarrow$ *graph center* [Plavšić, Nikolić *et al.*, 1993b].

Example D10

Reciprocal distance matrix $\mathbf{D}^{-1}$ and related molecular descriptors of 2-methylpentane. RDS_i is the reciprocal distance sum of the i th vertex.

	Atom	1	2	3	4	5	6	RDS_i
	1	0	1	0.50	0.33	0.25	0.50	2.58
	2	1	0	1	0.50	0.33	1	3.83
	3	0.50	1	0	1	0.50	0.50	3.50
	4	0.33	0.50	1	0	1	0.33	3.16
	5	0.25	0.33	0.50	1	0	0.25	2.33
	6	0.50	1	0.50	0.33	0.25	0	2.58

$$H \equiv RDSUM \equiv Wi(\mathbf{D}^{-1}) = (2.58 + 3.83 + 3.50 + 3.16 + 2.33 + 2.58)/2 = 17.98/2 = 8.99$$

$$RDSQ = (2.58 \times 3.83)^{1/2} + (3.83 \times 3.50)^{1/2} + (3.83 \times 2.58)^{1/2} + (3.50 \times 3.16)^{1/2} + (3.16 \times 2.33)^{1/2} + \sqrt{9.88} + \sqrt{13.40} + \sqrt{9.88} + \sqrt{11.06} + \sqrt{7.36} = 15.99$$

$$RDCHI = (2.58 \times 3.83)^{-1/2} + (3.83 \times 3.50)^{-1/2} + (3.83 \times 2.58)^{-1/2} + (3.50 \times 3.16)^{-1/2} + (3.16 \times 2.33)^{-1/2} = 1.58$$

$${}^R J \equiv J(\mathbf{D}^{-1}) = \frac{5}{0+1} \times [(2.58 \times 3.83)^{-1/2} + (3.83 \times 3.50)^{-1/2} + (3.83 \times 2.58)^{-1/2} + (3.50 \times 3.16)^{-1/2} + (3.16 \times 2.33)^{-1/2}] = 7.89$$

Analogously, the **reciprocal square distance matrix** $\mathbf{D}^{-2}$ is derived from the distance matrix by the following:

$$[\mathbf{D}^{-2}]_{ij} = \begin{cases} 1/d_{ij}^2 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where each off-diagonal element is the square reciprocal of the corresponding element of the distance matrix. This matrix can be useful to take into account the fact that atom interactions

decrease as the square distance between atoms increases, such as in the $\rightarrow$ *topological charge indices*. A variant of the Harary index H (called **Harary number** and denoted as H') was derived from this matrix and proposed as molecular descriptor [Mihalić and Trinajstić, 1992]:

$$H' \equiv Wi(\mathbf{D}^{-2}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}^{-2}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A VS_i(\mathbf{D}^{-2})$$

where Wi refers to the $\rightarrow$ *Wiener operator* and VS_i denotes the i th row sum of the reciprocal square distance matrix:

$$VS_i(\mathbf{D}^{-2}) = \sum_{j=1}^A d_{ij}^{-2} \quad j \neq i$$

Example D11

Reciprocal square distance matrix $\mathbf{D}^{-2}$ of 2-methylpentane. VS_i indicates the matrix row sum.

Atom	1	2	3	4	5	6	VS_i
1	0	1	0.25	0.111	0.063	0.25	1.674
2	1	0	1	0.25	0.111	1	3.361
3	0.25	1	0	1	0.25	0.25	2.750
4	0.111	0.25	1	0	1	0.111	2.472
5	0.063	0.111	0.25	1	0	0.063	1.486
6	0.25	1	0.25	0.111	0.063	0	1.674

$$H' \equiv Wi(\mathbf{D}^{-2}) = \frac{1}{2} \times (1.674 + 3.361 + 2.750 + 2.472 + 1.486 + 1.674) = \frac{13.416}{2} = 6.708$$

The **distance complement matrix** (also called **reversed distance matrix** or **vertex distance complement matrix**), denoted as **DC**, for simple graphs is defined as [Randić, 1997a; Randić and Pompe, 2001b]:

$$[\mathbf{DC}]_{ij} = \begin{cases} A - d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where A is the number of vertices and d_{ij} the distance between vertices v_i and v_j . The half-sum of the matrix elements is the **complement Wiener index** (or **reversed Wiener index**).

For vertex- and edge-weighted molecular graphs, the distance complement matrix was defined as [Ivanciuc, 2000i; Ivanciuc, 2002c]

$$[\mathbf{DC}(w)]_{ij} = \begin{cases} A - [\mathbf{D}(w)]_{ij} & \text{if } i \neq j \\ w_i & \text{if } i = j \end{cases}$$

where $\mathbf{D}(w)$ is a $\rightarrow$ *weighted distance matrix* and w denotes a $\rightarrow$ *weighting scheme*.

Reciprocal and complement distance matrices were defined because, unlike the distance matrix, the value of the matrix elements corresponding to pairs of vertices decreases when the distance between vertices increases; therefore, molecular descriptors calculated from reciprocal or complement matrices have numerical behavior and meaning opposite to molecular

Table D7 Some molecular descriptors derived from distance, reciprocal distance, and reciprocal square distance matrices for C8 data set (Appendix C – Set 1).

C8	W	$\bar{W}$	H	H'	RDCHI	RDSQ
n-Octane	84	3.000	9.502	13.743	1.997	24.823
2M	79	2.821	9.731	14.100	1.909	25.922
3M	76	2.714	9.814	14.267	1.885	26.379
4M	75	2.679	9.837	14.317	1.879	26.510
3E	72	2.571	9.920	14.483	1.851	26.966
22MM	71	2.536	10.176	14.767	1.774	28.000
23MM	70	2.500	10.108	14.733	1.788	27.791
24MM	71	2.536	10.059	14.650	1.798	27.552
25MM	74	2.643	9.966	14.467	1.823	27.047
33MM	67	2.393	10.318	15.033	1.737	28.744
34MM	68	2.429	10.179	14.867	1.768	28.164
2M3E	67	2.393	10.201	14.917	1.760	28.294
3M3E	64	2.286	10.438	15.250	1.703	29.355
223MMM	63	2.250	10.576	15.417	1.658	29.940
224MMM	66	2.357	10.431	15.167	1.689	29.222
233MMM	62	2.214	10.625	15.500	1.646	30.180
234MMM	65	2.321	10.389	15.167	1.700	29.120
2233MMMM	58	2.071	11.000	16.000	1.549	31.825

W is the Wiener index, $\bar{W}$ the average Wiener index, H the Harary index, H' the Harary number, RDCHI the RDCHI index, and RDSQ the RDSQ index.

descriptors derived from the distance matrix. The row sums of the distance matrix **D** (i.e., distance degrees) are greater for outer vertices than for the core vertices; thus, for instance, the value of the $\rightarrow$ *Balaban distance connectivity index*, which is based on the inverse of the distance degrees, is much more determined by the core vertices than the outer ones. On the contrary, the row sums of the reciprocal or complement distance matrix are greater for the core vertices and, therefore, $\rightarrow$ *Balaban-like indices* derived from these matrices are much more determined by the outer vertices.

From the weighted distance complement matrix **DC**, the *complement Wiener indices* were derived as $\rightarrow$ *Wiener-type indices* [Ivanciuc, 2000i]:

$$Wi(\mathbf{DC}, w) = \sum_{i=1}^A \sum_{j=i}^A [\mathbf{DC}(w)]_{ij}$$

where *w* is a weighting scheme for molecular graphs. Different complement Wiener indices are obtained depending on the weighting scheme *w*.

Moreover, the **complement Balaban index** ${}^C J$ [Nikolić, Plavšić *et al.*, 2001] or **reversed Balaban index** $1/J$ [Randić and Pompe, 2001b] was derived as a Balaban-like index as

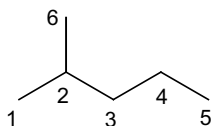
$$\begin{aligned} {}^C J \equiv 1/J \equiv J(\mathbf{DC}) &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot ({}^C \sigma_i \cdot {}^C \sigma_j)^{-1/2} \\ &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot [(A \cdot (A-1) - \sigma_i) \cdot (A \cdot (A-1) - \sigma_j)]^{-1/2} \end{aligned}$$

where ${}^c\sigma$ is the row sum of the distance complement matrix, A the number of vertices, B and C are the number of edges and the $\rightarrow$ *cyclomatic number*, respectively, and a_{ij} denotes the elements of the adjacency matrix equal to 1 for pairs of adjacent vertices and zero otherwise.

Example D12

Distance complement matrix and related molecular descriptors for 2-methylpentane.

${}^c\sigma$ indicates the matrix row sums.



DC =

Atom	1	2	3	4	5	6	${}^c\sigma$
1	0	5	4	3	2	4	18
2	5	0	5	4	3	5	22
3	4	5	0	5	4	4	22
4	3	4	5	0	5	3	20
5	2	3	4	5	0	2	16
6	4	5	4	3	2	0	18

$$Wi(\mathbf{DC}) = \frac{1}{2} \times (18 \times 2 + 22 \times 2 + 20 + 16) = 58$$

$${}^cJ \equiv 1/J \equiv J(\mathbf{DC}) =$$

$$= 5 \times \left(\frac{1}{\sqrt{18 \times 22}} + \frac{1}{\sqrt{22 \times 22}} + \frac{1}{\sqrt{22 \times 18}} + \frac{1}{\sqrt{22 \times 20}} + \frac{1}{\sqrt{20 \times 16}} \right) = 1.250$$

A variant of the distance complement matrix is the **complementary distance matrix CD**, which for simple graphs was defined as [Balaban, Mills *et al.*, 2000; Ivanciuc, Ivanciuc *et al.*, 2000a; Ivanciuc, 2000i]:

$$[\mathbf{CD}]_{ij} = \begin{cases} 1 + D - d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where D is the molecule diameter, which is the maximum distance in the graph, and 1 represents the minimum distance in the graph.

For vertex- and edge-weighted molecular graphs, the complementary distance matrix was defined as [Ivanciuc, 2000i]

$$[\mathbf{CD}(w)]_{ij} = \begin{cases} d_{\min}(w) + d_{\max}(w) - [\mathbf{D}(w)]_{ij} & \text{if } i \neq j \\ w_i & \text{if } i = j \end{cases}$$

where $\mathbf{D}(w)$ is a $\rightarrow$ *weighted distance matrix*, w denotes a $\rightarrow$ *weighting scheme*, $d_{\min}$ and $d_{\max}$ are the minimum and maximum distances in the weighted molecular graph, respectively. Obviously, for simple graphs, $d_{\min} = 1$ and $d_{\max} = D$.

A number of topological indices were derived from the complementary distance matrix **CD** and tested in QSPR models [Ivanciuc, Ivanciuc *et al.*, 2000a]. Among these are the **complementary Wiener indices** derived from the complementary distance matrix **CD** as $\rightarrow$ *Wiener-type indices* [Ivanciuc, 2000i]:

$$Wi(\mathbf{CD}, w) = \sum_{i=1}^A \sum_{j=i}^A [\mathbf{CD}(w)]_{ij}$$

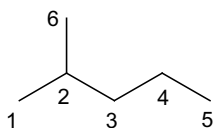
where w is a $\rightarrow$ *weighting scheme*. Different complementary Wiener indices are obtained depending on the weighting scheme w for molecular graphs.

A simple relationship between the complementary Wiener index and the $\rightarrow$ *Wiener index* W holds [Ivanciuc, Ivanciuc *et al.*, 2002b]:

$$Wi(\mathbf{CD}) = \frac{1}{2} \cdot (d_{\max} + d_{\min}) \cdot A \cdot (A-1) - W$$

Example D13

Complementary distance matrix $\mathbf{CD}$ and related Wiener-type index $Wi(\mathbf{CD})$ for 2-methylpentane. VS_i indicates the matrix row sums.



$$\mathbf{CD} = \begin{array}{c|cccccc|c} \text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 & VS_i \\ \hline 1 & 0 & 4 & 3 & 2 & 1 & 3 & 13 \\ 2 & 4 & 0 & 4 & 3 & 2 & 4 & 17 \\ 3 & 3 & 4 & 0 & 4 & 3 & 3 & 17 \\ 4 & 2 & 3 & 4 & 0 & 4 & 2 & 15 \\ 5 & 1 & 2 & 3 & 4 & 0 & 1 & 11 \\ 6 & 3 & 4 & 3 & 2 & 1 & 0 & 13 \end{array}$$

$$Wi(\mathbf{CD}) = \frac{1}{2} \cdot (13 + 17 + 17 + 15 + 11 + 13) = 43$$

Very similar to the complementary distance matrix $\mathbf{CD}$, the **reverse Wiener matrix**, denoted by $\mathbf{RW}$, was defined as [Balaban, Mills *et al.*, 2000]

$$[\mathbf{RW}]_{ij} = \begin{cases} D - d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where D is the molecule diameter, which is the maximum distance in the graph. Note that all the entries in the reverse Wiener matrix are lower by one than those in the complementary distance matrix for simple graphs. Another property of the reversed Wiener matrix is that matrix elements corresponding to the diameter in the distance matrix have zero values in the reverse Wiener matrix.

The i th row sum of the reverse Wiener matrix was called **reverse-distance sum** and defined as

$$VS_i(\mathbf{RW}) = \sum_{j=1}^A [\mathbf{RW}]_{ij}$$

where VS_i is the row sum operator.

The half sum of the matrix elements was proposed as molecular descriptor with the name of **reverse Wiener index**, denoted by Λ [Balaban, Mills *et al.*, 2000]:

$$\Lambda \equiv Wi(\mathbf{RW}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{RW}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A VS_i(\mathbf{RW})$$

where the symbol Wi denotes the $\rightarrow$ *Wiener operator*. The following relationship between the reverse Wiener index and the $\rightarrow$ *Wiener index* W holds [Ivanciuc, Ivanciuc *et al.*, 2002b]:

$$Wi(\mathbf{RW}) = \frac{1}{2} \cdot D \cdot A \cdot (A-1) - W$$

Moreover, the $\rightarrow$ *spectral indices* $MinSp(\mathbf{RW})$ and $MaxSp(\mathbf{RW})$, the $\rightarrow$ *Hosoya-type index* $Ho(\mathbf{RW})$, the $\rightarrow$ *hyper-Wiener-type index* $HyWi(\mathbf{RW})$, the $\rightarrow$ *Balaban-like index* $J(\mathbf{RW})$, and $\rightarrow$ *Balaban-like information indices* $U(\mathbf{RW})$, $V(\mathbf{RW})$, $X(\mathbf{RW})$, and $Y(\mathbf{RW})$ were tested in property modeling of some alkanes [Ivanciuc, Ivanciuc *et al.*, 2002b].

A **standardized complementary distance matrix** was also proposed to obtain **constant interval reciprocal indices** (CIR indices); the idea was that standardized distances should be uniformly spaced as topological distances [Schultz and Schultz, 1998, 2000]. The elements of the standardized complementary distance matrix are defined as

$$[\mathbf{CIR}]_{ij} = \begin{cases} \frac{1 + d_{\max} - d_{ij}}{d_{\max}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} denotes the elements of the distance matrix, $d_{\max}$ is the largest integer in the distance matrix, that is, the graph diameter D , or in the i th row, that is, the $\rightarrow$ *atom eccentricity* η_i , or arbitrarily user-defined. If $d_{\max}$ is chosen as the largest value in each row, the resulting matrix is unsymmetric. Moreover, $d_{\max}$ can also be equal to the number of graph vertices A .

Using these modified standardized distance matrices instead of the classical distance matrix, **CIRD indices** were defined as $\rightarrow$ *Wiener-type indices* by analogy with the $\rightarrow$ *Wiener index*:

$$CIRD \equiv Wi(\mathbf{CIR}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{CIR}]_{ij}$$

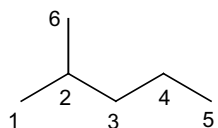
Moreover, **CIRS indices** and **CIRS' indices** were defined by analogy with the $\rightarrow$ *reciprocal Schultz indices*:

$$CIRS = \sum_{i=1}^A [(\mathbf{A} + \mathbf{CIR}) \cdot \mathbf{v}]_i \quad CIRS' = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{CIR} \cdot \mathbf{v}]_{ij} = \sum_{i=1}^A \delta_i \cdot {}^{CIR}\sigma_i$$

where $\mathbf{A}$ is the $\rightarrow$ *adjacency matrix*, $\mathbf{v}$ is the A -dimensional vector of $\rightarrow$ *vertex degrees* δ and ${}^{CIR}\sigma_i$ is the i th row sum of the **CIR** matrix. Different $CIRD$, $CIRS$, and $CIRS'$ indices can be calculated depending on the value chosen for the $d_{\max}$ parameter.

Example D14

Standardized complementary distance matrix **CIR** and related molecular descriptors for 2-methylpentane. Calculation is based on $d_{\max} = D = 4$.

	Atom	1	2	3	4	5	6	$^{CIR}\sigma_i$
	1	0	1	0.75	0.50	0.25	0.75	3.25
	2	1	0	1	0.75	0.50	1	4.25
	3	0.75	1	0	1	0.75	0.75	4.25
	4	0.25	0.75	1	0	1	0.50	3.50
	5	0.25	0.50	0.75	1	0	0.25	2.75
	6	0.75	1	0.75	0.50	0.25	0	3.25

$$CIRD = 5 \times 1 + 2 \times 0.25 + 3 \times 0.50 + 5 \times 0.75 = 10.75$$

$$CIRS' = 1 \times 3.25 + 3 \times 4.25 + 2 \times 4.25 + 2 \times 3.50 + 1 \times 2.75 + 1 \times 3.25 = 37.50$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0.75 & 0.50 & 0.25 & 0.75 \\ 1 & 0 & 1 & 0.75 & 0.50 & 1 \\ 0.75 & 1 & 0 & 1 & 0.75 & 0.75 \\ 0.25 & 0.75 & 1 & 0 & 1 & 0.50 \\ 0.25 & 0.50 & 0.75 & 1 & 0 & 0.25 \\ 0.75 & 1 & 0.75 & 0.50 & 0.25 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \\ 2 \\ 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 9.50 \\ 10.00 \\ 12.25 \\ 9.00 \\ 7.50 \\ 9.50 \end{bmatrix}$$

$$CIRS = 9.50 + 10.00 + 12.25 + 9.00 + 7.50 + 9.50 = 57.75$$

Two quotient matrices were derived from the distance complement matrix **DC**.

The **distance/distance complement quotient matrix**, denoted by **D/DC**, was defined as [Nikolić, Plavšić *et al.*, 2001]

$$[D/DC]_{ij} = \begin{cases} \frac{d_{ij}}{A-d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

and its reciprocal, the **distance complement/distance quotient matrix**, denoted by **DC/D**, as [Randić and Pompe, 2001b; Nikolić, Plavšić *et al.*, 2001]

$$[DC/D]_{ij} = \begin{cases} \frac{A-d_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where A is the number of graph vertices and d_{ij} the topological distance between vertices v_i and v_j .

From the matrix $\mathbf{D}/\mathbf{DC}$ and the matrix $\mathbf{DC}/\mathbf{D}$, two $\rightarrow$ *Balaban-like indices* were derived and called **quotient Balaban index of the first kind** and **quotient Balaban index of the second kind**, respectively [Nikolić, Plavšić *et al.*, 2001].

From distance complement matrix $\mathbf{DC}$, complementary distance matrix $\mathbf{CD}$, and reverse Wiener matrix $\mathbf{RW}$, the corresponding reciprocal matrices, called **reciprocal distance complement matrix** $\mathbf{DC}^{-1}$ [Ivanciuc, 2000i], **reciprocal complementary distance matrix** $\mathbf{CD}^{-1}$ [Ivanciuc, 2000i], and **reciprocal reverse Wiener matrix** $\mathbf{RRW}$ [Balaban, Mills *et al.*, 2000], were also derived by substituting the off-diagonal matrix elements by the corresponding reciprocal and leaving the diagonal elements unchanged. Moreover, $\rightarrow$ *Wiener-type indices* were derived from these reciprocal matrices to be used in QSAR modeling.

Opposite to the distance matrix is the $\rightarrow$ *detour matrix*, where the entries correspond to the length of the longest path between two vertices. Other matrices related to the distance matrix are $\rightarrow$ *geometric distance/topological distance quotient matrix*, $\rightarrow$ *detour-distance combined matrix*, $\rightarrow$ *distance/detour quotient matrix*, $\rightarrow$ *distance-degree matrices*, $\rightarrow$ *expanded distance matrices*, $\rightarrow$ *distance-path matrix*, $\rightarrow$ *delta matrix*, $\rightarrow$ *distance sum layer matrix*, $\rightarrow$ *distance-sequence matrix*.

Note. The distance matrix derived from a graph must not be confused with the distance matrix derived from a $\rightarrow$ *data set*, called here $\rightarrow$ *data distance matrix*.

📖 [Entiger, Jackson *et al.*, 1976; Doyle and Garver, 1977; Bersohn, 1983; Plesnik, 1984; Rouvray, 1986a; Müller, Szymanski *et al.*, 1987; Senn, 1988; Mihalić, Nikolić *et al.*, 1992; Kunz, 1993, 1994; Thangavel and Venuvanalingam, 1993; Dobrynin and Gutman, 1994; Dobrynin, 1995; Chepoi, 1996; Ivanciuc, Laidboeur *et al.*, 1997; Ivanciuc and Ivanciuc, 1999]

- **distance measures** $\rightarrow$ similarity/diversity
- **distance measure connectivity indices** $\rightarrow$ combined descriptors
- **distance-normalized exponential sum connectivities** $\rightarrow$ exponential sum connectivities
- **distance number** $\equiv$ *distance degree* $\rightarrow$ distance matrix

■ distance-path matrix ($\mathbf{D_P}$)

The distance-path matrix or **vertex distance-path matrix**, denoted as $\mathbf{D_P}$, is a $\rightarrow$ *combinatorial matrix* derived from the $\rightarrow$ *distance matrix* $\mathbf{D}$. It is a square symmetric matrix $A \times A$ whose off-diagonal entries are the number of all paths of any length m ($1 \leq m \leq d_{ij}$) that are included in the shortest path from vertex v_i to vertex v_j (whose $\rightarrow$ *topological distance* is d_{ij}) [Diudea, 1996a]. The diagonal entries are zero.

Each entry i - j of the matrix $\mathbf{D_P}$ is calculated as the following:

$$[\mathbf{D_P}]_{ij} = \binom{d_{ij} + 1}{2} = \frac{d_{ij}^2 + d_{ij}}{2}$$

that is, as all the possible combinations of two elements taken from $d_{ij} + 1$ elements (binomial coefficient).

For vertex- and edge-weighted molecular graphs, the distance-path matrix was defined as [Ivanciuc and Ivanciuc, 1999; Ivanciuc, 2000i]

$$[\mathbf{D_P}(w)]_{ij} = \binom{[\mathbf{D}(w)]_{ij} + 1}{2} = \frac{[\mathbf{D}(w)]_{ij}^2 + [\mathbf{D}(w)]_{ij}}{2}$$

where $\mathbf{D}(w)$ is a $\rightarrow$ *weighted distance matrix* calculated on the basis of the $\rightarrow$ *weighting scheme* w . Unlike the distance-path matrix of simple graphs, whose diagonal elements are equal to zero, the distance-path matrix of weighted graphs has diagonal elements different from zero since depending on the vertex weights.

The **hyper-distance-path index** D_P can be obtained by applying the $\rightarrow$ *Wiener operator* Wi to the distance-path matrix as

$$D_P \equiv Wi(\mathbf{D}_P) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}_P]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij} + \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \binom{d_{ij}}{2} = W + D_\Delta$$

where in the right expression the first term is just the $\rightarrow$ *Wiener index* W and the second the $\rightarrow$ *delta number* D_Δ , which can be considered the “non-Wiener part” of the index [Diudea, Katona *et al.*, 1997].

For acyclic graphs, the hyperdistance-path index D_P coincides with the $\rightarrow$ *hyper-Wiener index* WW derived from the $\rightarrow$ *Wiener matrix*, and with the $\rightarrow$ *hyper-detour index* derived from the $\rightarrow$ *detour-path matrix*. Moreover, it was proposed as extension of the hyper-Wiener index for any graph [Klein, Lukovits *et al.*, 1995].

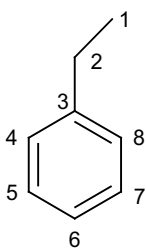
The **reciprocal distance-path matrix** $\mathbf{D}_P^{-1}$ is a matrix whose elements are the reciprocal of the corresponding distance-path matrix elements. The general definition for weighted molecular graphs is [Ivanciuc, 2000i]

$$[\mathbf{D}_P^{-1}(w)]_{ij} = \begin{cases} 1/[\mathbf{D}_P(w)]_{ij} & \text{if } i \neq j \\ [\mathbf{D}_P(w)]_{ii} & \text{if } i = j \end{cases}$$

The $\rightarrow$ *hyper-Harary distance index* is the $\rightarrow$ *Wiener-type index* derived from this matrix.

Example D15

The distance-path matrix $\mathbf{D}_P$ for ethylbenzene. VS_i indicates the matrix row sums. D_P is the hyper-distance-path index.



$\mathbf{D}_P =$

Atom	1	2	3	4	5	6	7	8	VS_i
1	0	1	3	6	10	15	10	6	51
2	1	0	1	3	6	10	6	3	30
3	3	1	0	1	3	6	3	1	18
4	6	3	1	0	1	3	6	3	23
5	10	6	3	1	0	1	3	6	30
6	15	10	6	3	1	0	1	3	39
7	10	6	3	6	3	1	0	1	30
8	6	3	1	3	6	3	1	0	23

$$D_P \equiv Wi(\mathbf{D}_P) = \frac{1}{2} \cdot (51 + 3 \times 30 + 18 + 2 \times 23 + 39) = 122$$

$\rightarrow$ *Characteristic polynomial* of combinatorial matrices as well as $\rightarrow$ *spectral indices* were proposed as molecular descriptors for QSAR problems [Ivanciuc, Diudea *et al.*, 1998].

The **delta matrix** or **distance–delta matrix** or **vertex distance–delta matrix**, denoted as $\mathbf{D}_\Delta$, is another combinatorial matrix defined as the difference between the $\rightarrow$ distance-path matrix $\mathbf{D}_P$ and the $\rightarrow$ distance matrix $\mathbf{D}$ [Diudea, 1996a; Diudea, Katona *et al.*, 1997; Ivanciuc, Diudea *et al.*, 1998]:

$$\mathbf{D}_\Delta = \mathbf{D}_P - \mathbf{D}$$

The delta matrix is a square symmetric matrix of dimension $A \times A$, A being the number of graph vertices, whose entries represent the number of all paths larger than unity included in the shortest path between the considered vertices and are formally defined as

$$[\mathbf{D}_\Delta]_{ij} \equiv \begin{cases} \binom{d_{ij}}{2} = \frac{d_{ij} \cdot (d_{ij} - 1)}{2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} is the $\rightarrow$ topological distance between vertices v_i and v_j .

Applying the $\rightarrow$ Wiener operator Wi to the delta matrix a molecular descriptor called **delta number** D_Δ is obtained as

$$D_\Delta \equiv Wi(\mathbf{D}_\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}_\Delta]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \binom{d_{ij}}{2}$$

The delta number can also be derived by

$$D_\Delta = D_P - W$$

where W is the $\rightarrow$ Wiener index and D_P the $\rightarrow$ hyper-distance-path index which coincides with the $\rightarrow$ hyper-Wiener index WW for acyclic graphs. It follows that the delta number is the “non-Wiener” part of the hyper-Wiener index.

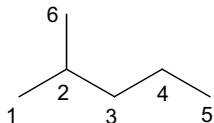
Moreover, the delta number D_Δ can be related to the distance matrix and the Wiener index W by

$$D_\Delta = \frac{\text{tr}(\mathbf{D}^2) - 2 \cdot W}{4} = \frac{D_2 - W}{2}$$

where $\mathbf{D}^2$ is the distance matrix raised to the second power, and D_2 is the second-order $\rightarrow$ distance distribution moment, that is, the sum of the square distances in the graph.

Example D16

Delta matrix $\mathbf{D}_\Delta$ and delta number for the 2-methylpentane.



$$\mathbf{D}_\Delta =$$

Atoms	1	2	3	4	5	6
1	0	0	1	3	6	1
2	0	0	0	1	3	0
3	1	0	0	0	1	1
4	3	1	0	0	0	3
5	6	3	1	0	0	6
6	1	0	1	3	6	0

$$D_\Delta = \frac{1}{2} \cdot \sum_{i=1}^6 \sum_{j=1}^6 [\mathbf{D}_\Delta]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^6 \sum_{j=1}^6 \frac{d_{ij}^2 - d_{ij}}{2} = 26$$

- **distance of a vertex** $\equiv$ *distance degree* $\rightarrow$ distance matrix
- **distance polynomial** $\rightarrow$ characteristic polynomial-based descriptors
- **Distance Profile descriptors** $\rightarrow$ substructure descriptors
- **distance rank** $\equiv$ *distance degree* $\rightarrow$ distance matrix
- **distance/resistance quotient matrix** $\rightarrow$ resistance matrix
- **distance-sequence matrix** $\rightarrow$ sequence matrices
- **distance sum** $\equiv$ *distance degree* $\rightarrow$ distance matrix
- **distance-sum-connectivity matrix** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **distance sum layer matrix** $\rightarrow$ layer matrices
- **distance-valency matrices** $\rightarrow$ distance-degree matrices
- **Diverse Property-Derived method** $\rightarrow$ cell-based methods
- **diversity** $\rightarrow$ similarity/diversity
- **diversity matrix** $\equiv$ *data distance matrix* $\rightarrow$ similarity/diversity
- **DLI score** $\rightarrow$ scoring functions
- **DNA sequences** $\rightarrow$ biodescriptors
- **docking** $\rightarrow$ drug design
- **donor superdelocalizability** $\equiv$ *nucleophilic superdelocalizability* $\rightarrow$ quantum-chemical descriptors
- **Dosmrov complexity index** $\rightarrow$ molecular complexity
- **double-bond count** $\rightarrow$ multiple bond descriptors
- **double bond equivalenst** $\equiv$ *index of hydrogen deficiency* $\rightarrow$ multiple bond descriptors
- **DOS** $\equiv$ *Density Of States* $\rightarrow$ quantum-chemical descriptors ($\odot$ EIM descriptors)

■ double invariants

Double invariants include various molecular descriptors calculated by a general approach for the derivation of graph-invariants, which generates a graph-theoretical matrix whose elements are molecular subgraphs rather than numbers [Randić, Plavšić *et al.*, 1997]. Matrices of this kind are called **graphical matrices** denoted as **GG**; they are square symmetric matrices of dimension $A \times A$, A being the number of graph vertices, and the element $i-j$ is defined as a certain subgraph in some way related to vertices v_i and v_j . There are different approaches to derive graphical matrices. In the first graphical matrix, called **path matrix**, the element $i-j$ was defined as the subgraph made up of all the shortest paths joining vertices v_i and v_j ; this subgraph may be a path or a cyclic fragment, including the molecule as a whole. The diagonal elements are isolated vertices.

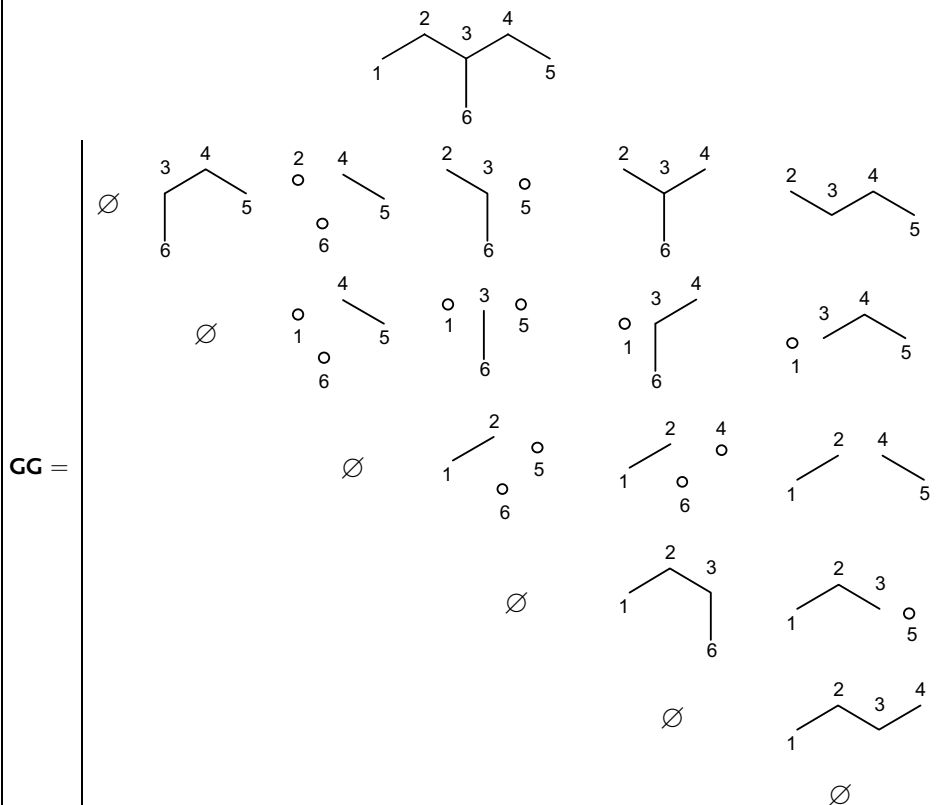
Another graphical matrix was defined so that the off-diagonal element $i-j$ is the subgraph obtained by deleting vertices v_i and v_j and incident edges from the molecular graph [Randić, Basak *et al.*, 2004]. By definition, all the diagonal elements are an empty subgraph. In this vertex-graphical matrix, the entries may be disconnected fragments and all the fragments have the same number of vertices. *Vertex-graphical matrices* can be further distinguished into *dense*, if all the matrix elements except for the diagonal elements are nonzero, and *sparse*, if only off-diagonal elements corresponding to pairs of adjacent vertices are different from the empty subgraph [Nikolić, Miličević *et al.*, 2005].

Another way of constructing graphical matrices is to define their elements $i-j$ as the subgraph obtained by removal of the paths joining vertices v_i and v_j from the molecular graph [Nikolić, Miličević *et al.*, 2005]. These matrices include the *path-graphical matrices*, which are dense since

they are obtained by considering all the pairs of vertices, and the *edge-graphical matrices*, which are sparse since they collect only subgraphs derived by removal of paths connecting pairs of adjacent vertices.

Example D17

Vertex-graphical matrix of 3-methylpentane, obtained by the criterion proposed in [Randić, Basak *et al.*, 2004].



To calculate molecular descriptors, graphical matrices need to be transformed into numerical matrices; this transformation is carried out by means of a graph invariant (e.g., the $\rightarrow$ *Wiener index*). Then, in the resulting numerical matrix the element $i-j$ is defined as the value of the selected graph invariant derived from the subgraph in position $i-j$ in the graphical matrix. By definition, all the elements corresponding to the entries of the graphical matrix being the empty subgraph are equal to zero. The graph invariant of a disconnected subgraph is calculated by adding the values of the graph invariant derived from all the components of the subgraph.

Finally, a second graph invariant, which can be different from the first invariant, is applied to the graphical matrix in the numerical form to obtain the double invariant.

Therefore, double invariants are functions of the type:

$$\mathcal{D}_2\{\mathcal{D}_1(\mathbf{GG})\}$$

It must be noted that the two invariants have to be compatible, that is, the domain of $\mathcal{D}_2$ is in the range of $\mathcal{D}_1$; moreover, the two invariant operators do not commute, that is,

$$\mathcal{D}_2\{\mathcal{D}_1(\mathbf{GG})\} \neq \mathcal{D}_1\{\mathcal{D}_2(\mathbf{GG})\}$$

Given the large number of available graph-invariants, there are many possible combinations of compatible invariants, resulting in an explosion of new descriptors.

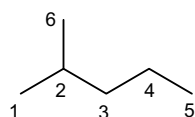
A molecular branching index was proposed for acyclic graphs, based on the double invariant approach and called $\rightarrow \lambda\lambda_1$ *branching index* [Randić, 1997d, 1998a]. A calculation example for 2-methylpentane is given below.

Other proposed double invariants are the **Wiener–Wiener number** [Randić, Basak *et al.*, 2004], denoted as $W(W)$, and the **Hosoya–Wiener index** [Nikolić, Miličević *et al.*, 2005], denoted as ZW .

The double invariant approach can be extended to 3D molecular geometry, defining graphical geometry matrices and the corresponding **3D double invariants**.

Example D18

Calculation of the double invariant $\lambda\lambda_1$ for 2-methylpentane: the graphical matrix $\mathbf{GG}$ consists of the paths joining two vertices v_i and v_j ; the first invariant is the leading eigenvalue of the $\rightarrow$ *adjacency matrix* $\mathbf{A}$ of each subgraph and the double invariant is the leading eigenvalue of the graphical matrix in the numerical form. The symbol ${}^m p_{ij}$ represents a path of length m between vertices v_i and v_j .



$$\lambda\lambda_1 = 6.8313$$

$$\mathbf{GG} =$$

Atom	1	2	3	4	5	6
1	0	${}^1 p_{12}$	${}^2 p_{13}$	${}^3 p_{14}$	${}^4 p_{15}$	${}^2 p_{16}$
2		0	${}^1 p_{23}$	${}^2 p_{24}$	${}^3 p_{25}$	${}^1 p_{26}$
3			0	${}^1 p_{34}$	${}^2 p_{35}$	${}^2 p_{36}$
4				0	${}^1 p_{45}$	${}^3 p_{46}$
5					0	${}^4 p_{56}$
6						0

Leading eigenvalues of $\mathbf{GG}$ matrix subgraphs

Atom	1	2	3	4	5	6
1	0	1.0000	1.4142	1.6180	1.7321	1.4142
2	1.0000	0	1.0000	1.4142	1.6180	1.0000
3	1.4142	1.0000	0	1.0000	1.4142	1.4142
4	1.6180	1.4142	1.0000	0	1.0000	1.6180
5	1.7321	1.6180	1.4142	1.0000	0	1.7321
6	1.4142	1.0000	1.4142	1.6180	1.7321	0

- **DP descriptor** → charge descriptors (⊖ submolecular polarity parameter)
- **DPD method** ≡ *Diverse Property-Derived method* → cell-based methods
- **DP indices** ≡ *indices of differences of path lengths*

■ DRAGON descriptors

These are various molecular descriptors ranging from → *count descriptors* to more complex → *geometrical descriptors* calculated by the software DRAGON [DRAGON – Talete s.r.l., 2007].

DRAGON software was conceived for the calculation of molecular descriptors by Milano Chemometrics and QSAR Research Group of Todeschini. DRAGON dates back to 1996, with the name of WHIM-3D/QSAR, while the first DRAGON version was released in 2000. Some details about DRAGON are given in Refs [Tetko, 2003; Mauri, Consonni *et al.*, 2006].

The number of DRAGON descriptors is 1664 in the version 5.4; however, this number increased up to 3224 after inclusion of 2D binary atom pairs and 2D frequency atom pairs, both consisting of 780 descriptors in version 5.5.

DRAGON descriptors are distinguished into 22 categories, which are listed in Table D8.

Table D8 List of descriptor categories in DRAGON software.

ID	Description	No.	ID	Description	No.
1	Constitutional descriptors	48	12	Geometrical descriptors	74
2	Topological descriptors	119	13	RDF descriptors	150
3	Walk and path counts	47	14	3D MoRSE descriptors	160
4	Connectivity indices	33	15	WHIM descriptors	99
5	Information indices	47	16	GETAWAY descriptors	197
6	2D autocorrelations	96	17	Functional group counts	154
7	Edge adjacency indices	107	18	Atom-centered fragments	120
8	Burden eigenvalues	64	19	Charge descriptors	14
9	Topological charge indices	21	20	Molecular properties	29
10	Eigenvalue-based indices	44	21	2D binary atom pairs	780
11	Randić molecular profiles	41	22	2D frequency atom pairs	780

Different molecular formats can be used for the descriptor calculation, and principal components can also be calculated, separately for each descriptor category.

In addition to descriptors' calculation, some explorative tools are also available that allow one to calculate descriptor values and their univariate statistics, to project and visualize molecules in the descriptor/response space, to calculate descriptor pair correlations, and to identify the most and the least correlated descriptors with a selected one.

The DRAGON software can be used in combination with the **MobyDigs software** [MobyDigs – Talete s.r.l., 2003; Todeschini, Consonni *et al.*, 2003], which allows QSAR/QSPR analysis by → *variable selection* based on genetic algorithms. To manage the huge number of molecular descriptors provided by DRAGON, several tools are available. Different fitness functions can be selected to evaluate the quality of QSAR models, several populations of models can be contemporarily developed, different validation techniques can be used for model validation, and predictions of the property of new molecules can be performed by means of selected models; → *consensus analysis* is also available.

📖 Additional references are provided in the thematic bibliography (see Introduction).

➤ **drug** → drug design

■ drug design

Drug design is a research field where several disciplines are involved, including not only the design but also the pharmacokinetics and the toxicity of drugs. Appropriate chemometric tools, such as experimental design, multivariate analysis, artificial neural networks, are usually used in the planning and evaluation of pharmacokinetics and toxicological experiments, as well as in modeling of the biological activity of drugs. Moreover, → *similarity searching* and → *substructural analysis* are actually very useful for screening a large database of chemical compounds in the search for new drugs and pharmacophores [Purcell, Bass *et al.*, 1973; Martin, 1978; Ariëns, 1979; Franke, 1984; Dean, 1987; Hadzi and Jerman-Blazic, 1987; Ramsden, 1990; Tute, 1990; Kubinyi, 1993a; van de Waterbeemd, 1995; van de Waterbeemd, Testa *et al.*, 1997; Kubinyi, Folkers *et al.*, 1998a, 1998b].

Drug design techniques are all implemented in software packages that run on powerful workstations and can be distinguished as three main groups: **Computer-Aided Drug Design** (CADD) involves all computer-assisted techniques used to discover, design, and optimize biologically active compounds with a possible use as drugs; **Computer-Aided Molecular Design** (CAMD) involves all computer-assisted techniques used to discover, design, and optimize biologically active compounds with desired structure and properties; **Computer-Aided Molecular Modeling** (CAMM, or simply **Molecular Modeling**) is the investigation of molecular structures and properties using → *computational chemistry* and graphical visualization techniques.

Some fundamental terms and concepts in drug design are briefly reviewed below [van de Waterbeemd, Carter *et al.*, 1997; IUPAC Recommendations, 1997, 1998].

A **drug** is any substance for treating, curing, or preventing a disease in human beings or animals. A drug may also be used to make a medical diagnosis or to restore, correct, or modify physiological functions. A **lead compound** is a compound that, because of its biological properties, is taken as a reference structure that is the starting point for the process of identifying new active compounds; moreover, a lead compound should require the presence of at least one marketed drug, derived from that particular lead structure [Oprea, Davis *et al.*, 2001].

A **reference compound** (or **reference structure**) is a compound assumed as the reference for some considered aspect, such as a physico-chemical or biological property, or a molecular skeleton common to a set of compounds (→ *maximum common substructure*).

Lead discovery, generation, and optimization are basic activities in drug design, which are devoted to identifying active new chemical entities, developing new active compounds, and optimizing those able to be transformed into clinically useful drugs.

ADME properties (*Absorption, Distribution, Metabolism, and Elimination properties*) are properties of compounds that are of fundamental pharmaceutical importance; these are all those properties a drug needs to exert its pharmacological activity [van de Waterbeemd, Smith *et al.*, 2001a; Butina, Segall *et al.*, 2002; Ekins and Rose, 2002].

ADME properties have been recognized as a major reason for the failure of drug candidates; thus, several efforts are performed to develop molecular descriptors and models able to predict ADME properties of drug candidates before their synthesis. Bibliographic references to theoretical aspects and QSAR studies on ADME properties are reported in the thematic bibliography.

Bioisosterism is a strategy of medicinal chemistry for rational design of new drugs, applied to a lead compound as a special process of molecular modification consisting in the exchange of one bioisostere for another with the aim of enhancing biological activity without a significant change in the chemical structure [Burger, 1991; Cramer III, Clark *et al.*, 1996; Patani and LaVoie, 1996; Lima and Barreiro, 2005]. Bioisosteres are functional groups or substituents or molecular fragments with similar → *physico-chemical properties* that give similar biological properties to a chemical compound. To apply this approach, not only the chemical structure but also the → *mechanism of action*, at the level of interaction with the receptor, of the lead compound should be completely known, including the knowledge of all of its pharmacophoric centers.

Binding affinity is the tendency of a molecule to associate with another. In particular, the affinity of a drug is its ability to bind to its biological target, and a **ligand** is a compound that can bind to a biological target, thus giving origin to **docking**. Docking studies are computational techniques for the exploration of the possible binding modes of a ligand to a given receptor, enzyme, or other binding site.

A **receptor** is a molecule or a polymeric structure in a cell that acts as the biological target, recognizing and binding a compound. **Receptor mapping** is the technique used to describe geometric, electronic, and other physico-chemical features of a binding site. The active site cavity of a receptor is called the **binding site cavity** (or **receptor cavity**).

A **pharmacophore** is the ensemble of steric, electronic, and other physico-chemical properties that is necessary to ensure optimal supramolecular interactions with a specific biological target structure. In other words, the pharmacophore concept is based on the kinds of interactions observed in molecular recognition: hydrogen bonding, charge–charge, and hydrophobic interactions characterized by defined spatial arrangements. A pharmacophore does not represent a real molecule or a real association of functional groups, it is purely an abstract concept that accounts for the common molecular interaction capacity of a set of compounds with the target structure. A pharmacophore can be considered to be the largest common denominator shared by a set of active molecules.

The concepts of receptor and pharmacophore play a basic role in the alignment of molecules in → *grid-based QSAR techniques*.

A more detailed definition of pharmacophore is given by Bersuker who introduces three additional characteristics that are important in molecule-receptor interaction [Bersuker, 2003].

- 1 The pharmacophore should be defined by not just atoms from the periodic table, but by appropriate atom-in-molecule electronic characteristics.
- 2 Both electronic characteristics and geometry parameters of the pharmacophore vary from one active molecule to another within certain tolerances and the activity may be a function of these variations (i.e., pharmacophore flexibility).
- 3 The pharmacophore is a necessary but not sufficient condition of activity. Even if the pharmacophore is present, the activity of the molecule may be reduced by groups that hinder its proper docking with the receptor or may be enhanced by other groups responsible for properties useful to activity such as hydrophobicity.

The term pharmacophore is used in drug design while the term **toxiphore** is used in toxicology to refer to the set of unique molecular features in a given spatial

arrangement common to all the toxic molecules and deemed to represent the toxicity under consideration.

📖 Additional references are provided in the thematic bibliography (see Introduction).

- **drug-like indices** → property filters
- **drug-like scores** $\equiv$ *scoring functions*
- **DSI index** → electronegativity-based connectivity indices
- **dual degree** → vertex degree
- **dual electronic constants σ^+ and σ^-** → electronic substituent constants ($\odot$ resonance electronic constants)
- **Dubois steric constant** → steric descriptors ($\odot$ Taft steric constant)
- **Duchowitz–Castro $\log P$** → lipophilicity descriptors
- **dummy variables** $\equiv$ *indicator variables*
- **Dunn model based on surface area** → lipophilicity descriptors
- **DV index** → multiple bond descriptors
- **d-WDEN indices** $\equiv$ *directional WHIM density* → WHIM descriptors ($\odot$ directional WHIM descriptors)
- **d-WSHA indices** $\equiv$ *directional WHIM shape* → WHIM descriptors ($\odot$ directional WHIM descriptors)
- **d-WSIZ indices** $\equiv$ *directional WHIM size* → WHIM descriptors ($\odot$ directional WHIM descriptors)
- **d-WSYM indices** $\equiv$ *directional WHIM symmetry* → WHIM descriptors ($\odot$ directional WHIM descriptors)
- **dynamic QSAR** → Structure/Response Correlations
- **dynamic reactivity indices** → reactivity indices
- **DZ^K descriptors** → autocorrelation descriptors
- **D/ Δ index** → detour matrix
- **D/ Δ ring indices** → detour matrix
- **D/ Ω index** → resistance matrix

E

- **EA indices** $\equiv$ *Extended Adjacency matrix indices* $\rightarrow$ spectral indices
- **EAm_{ax} index** $\rightarrow$ spectral indices ($\odot$ extended adjacency matrix indices)
- **EA Σ index** $\rightarrow$ spectral indices ($\odot$ extended adjacency matrix indices)
- **eccentric** $\rightarrow$ distance matrix
- **eccentric adjacency index** $\rightarrow$ eccentricity-based Madan indices ($\odot$ Table E1)
- **eccentric adjacency topochemical indices** $\rightarrow$ eccentricity-based Madan indices ($\odot$ Table E1)

■ eccentricity-based Madan indices

These are a series of $\rightarrow$ *graph invariants* calculated on the $\rightarrow$ *H-depleted molecular graph* and defined as different combinations of $\rightarrow$ *atom eccentricity* η , which is the maximum topological distance from an atom, and other $\rightarrow$ *Local Vertex Invariants*. Eccentricity-based Madan indices are listed in Table E1 where index names, formulas, and bibliographic references are given.

These consist of two sets of indices: one is topological (ID 1-7) as it accounts only for features of a simple graph; the other set (ID 8-12) was defined topochemical as it was derived from vertex-weighted graphs to account for heteroatoms.

Together with atom eccentricity η , local invariants used to derive these molecular descriptors are $\rightarrow$ *vertex degree* δ_i , which is the number of vertices bonded to the i th vertex; $\rightarrow$ *distance degree* σ_i , which is the sum of the topological distances in the graph from vertex v_i ; Morgan's $\rightarrow$ *extended connectivity* of first order, EC_i^1 , obtained by summing up the vertex degrees of all vertices bonded to vertex v_i ; $\rightarrow$ *Madan chemical degree*, δ_i^c , calculated by summing up the relative atomic weights of the adjacent vertices; $\rightarrow$ *chemical atom eccentricity*, η_i^c , which is the maximum distance weighted by relative atomic weights from the considered vertex; $\rightarrow$ *chemical extended connectivity*, EC_i^{1c} , obtained by summing up the Madan chemical degrees of all first neighbor vertices. M_i , used for the calculation of the augmented eccentric connectivity index (ID-7), is the product of the degrees of all the vertices adjacent to the vertex v_i [Baja, Sami *et al.*, 2006a]:

$$M_i = \prod_{j=1}^A (\delta_j)^{a_{ij}}$$

where the exponent a_{ij} takes value 1 only for pairs of adjacent vertices, zero otherwise, thus giving a contribution of 1 in the product. This local invariant is a modification of the extended connectivity and thus may be called **augmented connectivity**.

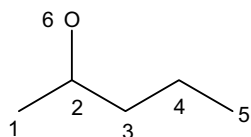
Table E1 Definitions and bibliographic references of the eccentricity-based Madan indices.

ID	Descriptor name	Formula	Reference
1	Eccentric connectivity index	$\xi^c = \sum_{i=1}^A \eta_i \cdot \delta_i$	[Sharma, Goswami <i>et al.</i> , 1997; Sardana and Madan, 2002c]
2	Eccentric distance sum	$\xi^{DS} = \sum_{i=1}^A \eta_i \cdot \sigma_i$	[Gupta, Singh <i>et al.</i> , 2002]
3	Adjacent eccentric distance sum index	$\xi^{SV} = \sum_{i=1}^A \frac{\eta_i \cdot \sigma_i}{\delta_i}$	[Sardana and Madan, 2002b]
4	Connective eccentricity index	$C^\xi = \sum_{i=1}^A \frac{\delta_i}{\eta_i}$	[Gupta, Singh <i>et al.</i> , 2000]
5	Eccentric adjacency index	$\xi^A = \sum_{i=1}^A \frac{EC_i^1}{\eta_i}$	[Gupta, Singh <i>et al.</i> , 2001b]
6	Superadjacency index	$\int^A = \sum_{i=1}^A \frac{\delta_i \cdot EC_i^1}{\eta_i}$	[Bajaj, Sambi <i>et al.</i> , 2004b]
7	Augmented eccentric connectivity index	$^A\xi^c = \sum_{i=1}^A \frac{M_i}{\eta_i}$	[Bajaj, Sambi <i>et al.</i> , 2006a]
8	Superadjacency topochemical index	$\int^{Ac} = \sum_{i=1}^A \frac{\delta_i^c \cdot EC_i^{1c}}{\eta_i^c}$	[Bajaj, Sambi <i>et al.</i> , 2004b]
9	Eccentric connectivity topochemical index	$\xi_c^c = \sum_{i=1}^A \eta_i^c \cdot \delta_i^c$	[Kumar, Sardana <i>et al.</i> , 2004]
10	Eccentric adjacency topochemical index (1)	$\xi_{1C}^A = \sum_{i=1}^A \frac{EC_i^{1c}}{\eta_i}$	[Gupta, Singh <i>et al.</i> , 2003]
11	Eccentric adjacency topochemical index (2)	$\xi_{2C}^A = \sum_{i=1}^A \frac{EC_i^1}{\eta_i^c}$	[Gupta, Singh <i>et al.</i> , 2003]
12	Eccentric adjacency topochemical index (3)	$\xi_{3C}^A = \sum_{i=1}^A \frac{EC_i^{1c}}{\eta_i^c}$	[Gupta, Singh <i>et al.</i> , 2003]

Topochemical indices are calculated from three graph theoretical matrices: (1) a $\rightarrow$ *chemical adjacency matrix* obtained by substituting row elements of the adjacency matrix, corresponding to bonded atoms, with relative atomic weights of the bonded atoms; (2) an $\rightarrow$ *atomic weight-weighted distance matrix* obtained by substituting topological distances with the sum of relative atomic weights of all the vertices involved in the shortest path; and (3) an $\rightarrow$ *additive chemical adjacency matrix* obtained by modifying the $\rightarrow$ *additive adjacency matrix*, replacing the vertex degree by the $\rightarrow$ *Madan chemical degree*.

Example E1

Eccentricity-based Madan indices for 2-pentanol.



Vertex degrees δ

Atom	1	2	3	4	5	6
δ	1	3	2	2	1	1

Distance matrix

Atom	1	2	3	4	5	6	σ_i
1	0	1	2	3	4	2	12
2	1	0	1	2	3	1	8
3	2	1	0	1	2	2	8
4	3	2	1	0	1	3	10
5	4	3	2	1	0	4	14
6	2	1	2	3	4	0	12

Atom	η_i
1	4
2	3
3	2
4	3
5	4
6	4

Additive adjacency matrix

Atom	1	2	3	4	5	6	EC_i^1
1	0	3	0	0	0	0	3
2	1	0	2	0	0	1	4
3	0	3	0	2	0	0	5
4	0	0	2	0	1	0	3
5	0	0	0	2	0	0	2
6	0	3	0	0	0	0	3

Eccentric connectivity index:

$$\xi^c = \sum_{i=1}^6 \eta_i \cdot \delta_i = 4 \cdot 1 + 3 \cdot 3 + 2 \cdot 2 + 3 \cdot 2 + 4 \cdot 1 + 4 \cdot 1 = 31$$

Eccentric distance sum:

$$\xi^{DS} = \sum_{i=1}^6 \eta_i \cdot \sigma_i = 4 \cdot 12 + 3 \cdot 8 + 2 \cdot 8 + 3 \cdot 10 + 4 \cdot 14 + 4 \cdot 12 = 222$$

Adjacency eccentric distance sum index:

$$\xi^{sv} = \sum_{i=1}^6 \frac{\eta_i \cdot \sigma_i}{\delta_i} = \frac{4 \cdot 12}{1} + \frac{3 \cdot 8}{3} + \frac{2 \cdot 8}{2} + \frac{3 \cdot 10}{2} + \frac{4 \cdot 14}{1} + \frac{4 \cdot 12}{1} = 183$$

Connective eccentricity index:

$$C^{\xi} = \sum_{i=1}^6 \frac{\delta_i}{\eta_i} = \frac{1}{4} + \frac{3}{3} + \frac{2}{2} + \frac{2}{3} + \frac{1}{4} + \frac{1}{4} = 3.417$$

Eccentric adjacency index:

$$\xi^A = \sum_{i=1}^6 \frac{EC_i^1}{\eta_i} = \frac{3}{4} + \frac{4}{3} + \frac{5}{2} + \frac{3}{3} + \frac{2}{4} + \frac{3}{4} = 6.833$$

Superadjacency index:

$$\int^A = \sum_{i=1}^6 \frac{\delta_i \cdot EC_i^1}{\eta_i} = \frac{1 \cdot 3}{4} + \frac{3 \cdot 4}{3} + \frac{2 \cdot 5}{2} + \frac{2 \cdot 3}{3} + \frac{1 \cdot 2}{4} + \frac{1 \cdot 3}{4} = 13$$

The multiplicative connectivities of the vertices are

$$M_1 = \delta_2 = 3, \quad M_2 = \delta_1 \cdot \delta_3 \cdot \delta_6 = 2, \quad M_3 = \delta_2 \cdot \delta_4 = 6, \\ M_4 = \delta_3 \cdot \delta_5 = 2, \quad M_5 = \delta_4 = 2, \quad M_6 = \delta_2 = 3$$

Augmented eccentric connectivity index:

$$A_{\xi^c} = \sum_{i=1}^6 \frac{M_i}{\eta_i} = \frac{3}{4} + \frac{2}{3} + \frac{6}{2} + \frac{2}{3} + \frac{2}{4} + \frac{3}{4} = 6.333$$

For the calculation of the topochemical indices the relative atomic weights ($m_1 = m_2 = m_3 = m_4 = m_5 = 1$, and $m_6 = 1.332$) and the following graph-theoretical matrices were used:

Chemical adjacency matrix

Atom	1	2	3	4	5	6	δ_i^c
1	0	1	0	0	0	0	1
2	1	0	1	0	0	1.332	3.332
3	0	1	0	1	0	0	2
4	0	0	1	0	1	0	2
5	0	0	0	1	0	0	1
6	0	1	0	0	0	0	1

Additive chemical adjacency matrix

Atom	1	2	3	4	5	6	EC_i^{1c}
1	0	3.332	0	0	0	0	3.332
2	1	0	2	0	0	1	4
3	0	3.332	0	2	0	0	5.332
4	0	0	2	0	1	0	3
5	0	0	0	2	0	0	2
6	0	3.332	0	0	0	0	3.332

Atomic weight-weighted distance matrix

Atom	1	2	3	4	5	6	η_i^c
1	0	1	2	3	4	2.332	4
2	1	0	1	2	3	1.332	3
3	2	1	0	1	2	2.332	2.332
4	3	2	1	0	1	3.332	3.332
5	4	3	2	1	0	4.332	4.332
6	2	1	2	3	4	0	4

Superadjacency topochemical index:

$$\int^{A_c} = \sum_{i=1}^6 \frac{\delta_i^c \times EC_i^{1c}}{\eta_i^c} = \frac{1 \times 3.332}{4} + \frac{3.332 \times 4}{3} + \frac{2 \times 5.332}{2.332} + \frac{2 \times 3}{3.332} \\ + \frac{1 \times 2}{4.332} + \frac{1 \times 3.332}{4} = 12.944$$

Eccentric connectivity topochemical index:

$$\xi_{\xi^c}^c = \sum_{i=1}^6 \eta_i^c \cdot \delta_i^c = 1 \times 4 + 3.332 \times 3 + 2 \times 2.332 + 2 \times 3.332 + 1 \times 4.332 + 1 \times 4 = 33.656$$

Eccentric adjacency topochemical indices:

$$\xi_{1C}^A = \sum_{i=1}^6 \frac{EC_i^{1c}}{\eta_i} = \frac{3.332}{4} + \frac{4}{3} + \frac{5.332}{2} + \frac{3}{3} + \frac{2}{4} + \frac{3.332}{4} = 7.165$$

$$\xi_{2C}^A = \sum_{i=1}^6 \frac{EC_i^1}{\eta_i^c} = \frac{3}{4} + \frac{4}{3} + \frac{5}{2.332} + \frac{3}{3.332} + \frac{2}{4.332} + \frac{3}{4} = 6.339$$

$$\xi_{3C}^A = \sum_{i=1}^6 \frac{EC_i^{1c}}{\eta_i^c} = \frac{3.332}{4} + \frac{4}{3} + \frac{5.332}{2.332} + \frac{3}{3.332} + \frac{2}{4.332} + \frac{3.332}{4} = 6.648$$

📖 [Sardana and Madan, 2003; Kumar and Madan, 2004]

- **eccentric connectivity index** → eccentricity-based Madan indices (⊙ Table E1)
- **eccentric connectivity topochemical index** → eccentricity-based Madan indices (⊙ Table E1)
- **eccentric distance sum** → eccentricity-based Madan indices (⊙ Table E1)
- **eccentricity** → distance matrix
- **eccentricity** ≡ *geometrical eccentricity*
- **ECFC fingerprints** → substructure descriptors (⊙ fingerprints)
- **ECFP fingerprints** ≡ *Extended Connectivity FingerPrints* → substructure descriptors (⊙ fingerprints)
- **EC method** ≡ *Electron-Conformational method* → Electronic-Topological method
- **ECN index** → electronegativity-based connectivity indices
- **ECP index** → electronegativity-based connectivity indices

■ edge adjacency matrix (E)

Derived from the → *molecular graph* G , the edge adjacency matrix, denoted by E , or more formally as ${}^E A$, also called **bond matrix**, encodes information about the connectivity between graph edges:

$$[{}^E A]_{ij} \equiv [E]_{ij} = \begin{cases} 1 & \text{if } (i, j) \text{ are adjacent bonds} \\ 0 & \text{otherwise} \end{cases}$$

It is a square symmetric matrix of dimension $B \times B$, where B is the number of bonds, and is usually derived from a → *H-depleted molecular graph* [Bonchev, 1983]. It is to be noted that the edge adjacency matrix of a graph G is equal to the → *adjacency matrix* of the → *line graph* of G [Gutman and Estrada, 1996].

The entries $[E]_{ij}$ of the matrix are equal to one if edges e_i and e_j are adjacent (the two edges thus forming a → *path* of length two) and zero otherwise. For multigraphs, the edge adjacency matrix can be augmented by a row and a column for each multiple edge.

The **edge degree** ε_i provides the simplest information related to the bond considered and is calculated from the edge adjacency matrix as follows [Bonchev, 1983]:

$$\varepsilon_i \equiv VS_i(E) = \sum_{j=1}^B [E]_{ij}$$

where VS_i is the → *row sum operator*.

The edge degree ε is related to the $\rightarrow$ vertex degree δ by the following relationship:

$$\varepsilon_i = \delta_{i(1)} + \delta_{i(2)} - 2$$

where $i(1)$ and $i(2)$ refer to the two vertices incident to the i th edge.

The number of edges whose same edge degree is equal to g is called the **edge degree count** ${}^gF_E$; therefore, the vector

$$\{{}^1F_E, {}^2F_E, \dots, {}^6F_E\}$$

can be associated with each graph G , provided the maximum edge degree is equal to 6.

Related to the previous definition is the **edge type count** $ne_{gg'}$, defined as the number of edges with the same vertex degree of the incident vertices, where g and g' are the degree values of the incident vertices.

The **total edge adjacency index** A_E , also known as **Platt number**, F [Platt, 1947, 1952], is the sum over all entries of the edge adjacency matrix:

$$A_E \equiv F = \sum_{i=1}^B \sum_{j=1}^B [E]_{ij} = \sum_{i=1}^B \varepsilon_i = 2 \cdot N_2$$

where N_2 is the number of graph connections.

Two adjacent edges constitute a second-order path and this subgraph is called **connection** (or **link**). The **connection number** N_2 , also known as **Gordon–Scantlebury index** (N_{GS}) [Gordon and Scantlebury, 1964], is the simplest graph invariant obtained from the edge adjacency matrix that considers both vertices and edges and is calculated as

$$N_2 \equiv N_{GS} \equiv {}^2P = A_E/2$$

where A_E is the total edge adjacency index and ${}^2P$ is the second-order path count. Note that the connection number also coincides with the $\rightarrow$ Bertz branching index.

The number of connections N_2 of a molecular graph is related to the $\rightarrow$ first Zagreb index M_1 and the $\rightarrow$ quadratic index Q by the following relationships:

$$N_2 = M_1/2 - A + 1 = Q + A - 2$$

where A is the number of atoms.

When multiple bonds are considered, the connection number can be calculated on multi-graphs by using more general approaches. Given an H-depleted molecular graph $\bar{G}$, the number of connections is equal to the number of edges in the line graph of $\bar{G}$. Moreover, the connection number can also be calculated as the sum of vertex contributions, taking into account that each connection intersects at a particular vertex and is therefore simply a function of the number of hydrogens on that atom:

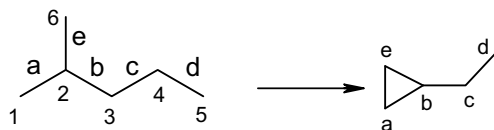
$$N_2 = \frac{1}{2} \cdot \sum_{i=1}^A (4 - h_i) \cdot (3 - h_i) - DB - 3 \cdot TB$$

where h_i is the number of hydrogen atoms attached to the i th atom and DB and TB denote the number of double and triple bonds in the molecule, respectively; the sum runs over all nonhydrogen atoms in the molecule [Hendrickson, Huang *et al.*, 1987].

As the number of connections is sensitive to different features of molecular structure such as size, branching, cyclicity, and multiple bonds, it was used by Bertz to define its $\rightarrow$ *molecular complexity index*.

Example E2

Line graph of 2-methylpentane and its edge adjacency matrix. ϵ_i is the edge degree and ϵ the edge connectivity index.



$$A_E = \sum_{i=1}^5 \sum_{j=1}^5 [E]_{ij} = \sum_{i=1}^5 \epsilon_i = 2 \cdot N_2 = 10$$

$$E =$$

edge	a	b	c	d	e	ϵ_i
a	0	1	0	0	1	2
b	1	0	1	0	1	3
c	0	1	0	1	0	2
d	0	0	1	0	0	1
e	1	1	0	0	0	2

$$\epsilon = (2 \cdot 3)^{-1/2} + (3 \cdot 2)^{-1/2} + (2 \cdot 1)^{-1/2} + (2 \cdot 2)^{-1/2} + (3 \cdot 2)^{-1/2} = 2.432$$

From the edge adjacency matrix, a graph-theoretical invariant analogous to the $\rightarrow$ *Randić connectivity index* was derived by Estrada [Estrada, 1995a; Cash, 1995b]; it is called **edge connectivity index** (or **bond connectivity index**), denoted by ϵ , and defined as

$$\epsilon = \sum_k (\epsilon_i \cdot \epsilon_j)^{-1/2}_k$$

where k runs over all connections N_2 , and ϵ_i and ϵ_j are the edge degrees of the two edges in the connection. It coincides with the Randić connectivity index χ of the line graph of G .

The **extended edge connectivity indices** (or **bond connectivity indices**) were defined as a generalization of the edge connectivity index in analogy to the $\rightarrow$ *Kier–Hall connectivity indices*:

$${}^m\epsilon_t = \sum_{k=1}^K \left(\prod_i \epsilon_i \right)_k^{-1/2}$$

where k runs over all of the m th order subgraphs, m is the number of edges in the subgraph, and K is the total number of m th order subgraphs; the edge degrees of all the edges in the subgraph are considered. The subscript “ t ” for the connectivity indices refers to the type of $\rightarrow$ *molecular subgraph* and is “ ch ” for chain or ring, “ pc ” for path-cluster, “ c ” for cluster, and “ p ” for path (this can also be omitted) [Estrada, Guevara *et al.*, 1998a; Estrada and Rodriguez, 1999; Estrada, 1999b]. Some mathematical relationships between the extended edge connectivity indices and $\rightarrow$ *line graph connectivity indices* were found.

The **spectral moments of the edge adjacency matrix** E were defined [Estrada, 1996, 1997, 1998b] as

$$\mu^k = \text{tr}(E^k)$$

where k is the power of the edge adjacency matrix and tr the $\rightarrow$ trace, that is, the sum of the diagonal elements. The zero-order spectral moment μ^0 corresponds to the number of edges in the graph, that is, the trace of the resulting B -dimensional identity matrix.

Since the k th-order spectral moment μ^k corresponds to the sum of all self-returning walks of length k in the line graph of the molecular graph G , it can be expressed as the linear combination of the $\rightarrow$ embedding frequencies of the molecular graph, that is, the counts of different structural fragments (subgraphs) in the graph, each fragment corresponding to a specific self-returning walk. Several relations between spectral moments and embedding frequencies were derived by Estrada [Estrada, 1998b; Marković and Gutman, 1999]. A **local spectral moment** of the edge adjacency matrix is defined as the sum of diagonal entries of the different powers of the edge adjacency matrix corresponding to a molecular fragment [Estrada and Molina, 2001b, 2001c; Estrada and Gonzalez, 2003]. It is defined as

$$\mu^k(F) = \sum_{b=1}^{B_F} [\mathbf{E}^k]_{bb}$$

where F indicates the molecular fragment considered, and the summation goes over all the bonds forming the fragment. **Bond spectral moments** are obtained when the fragment corresponds to a single bond; therefore, they simply are the single diagonal entries of the bond matrix raised to the k th power. Consequently, the total spectral moments of k th order can be expressed as the sum of the bond spectral moments of the same order.

The **weighted edge adjacency matrix** ${}^w\mathbf{E}$ is derived from an edge-weighted molecular graph obtained by applying any edge $\rightarrow$ weighting scheme w , which encodes information about each bond in the molecule. The weighting scheme can be based on quantities directly characterizing bonds, such as bond distances or bond dipoles, or quantities derived from weights of those atoms involved in each bond, such as atomic mass, atomic electronegativity, surface area contribution of polar atoms, atomic charges, and so on.

A weighted edge adjacency matrix ${}^w\mathbf{E}$ can be calculated as

$$[{}^w\mathbf{E}]_{ij} = \begin{cases} w_j & \text{if } (i,j) \text{ are adjacent bonds} \\ 0 & \text{otherwise} \end{cases}$$

where the off-diagonal elements of the matrix for a weighted graph are zero for nonadjacent edges, while, if two edges e_i and e_j are adjacent, the entry $i-j$ is defined by the weight w_j of the j th edge and the symmetric entry $j-i$ is defined by the weight w_i of the i th edge, thus resulting in an unsymmetrical matrix.

The **weighted edge degree** ${}^w\epsilon$ is calculated by applying the $\rightarrow$ vertex sum operator VS to the weighted edge adjacency matrix ${}^w\mathbf{E}$ as

$${}^w\epsilon_i = VS_i({}^w\mathbf{E}) = \sum_{j=1}^B [{}^w\mathbf{E}]_{ij} = \sum_{j=1}^B [\mathbf{E}]_{ij} \cdot w_j$$

where B is the number of edges in the molecular graph and w the edge-weighting scheme. The weighted edge degree of each i th edge is then the summation of the weights of all the edges adjacent to i th edge, that is, $[\mathbf{E}]_{ij} = 1$.

The **bond order-weighted edge adjacency matrix** ${}^\pi\mathbf{E}$ is obtained for weighted graphs whose edges are weighted by the $\rightarrow$ bond order π calculated by $\rightarrow$ computational chemistry methods on

selected molecular geometries [Estrada and Montero, 1993]:

$$[\pi \mathbf{E}]_{ij} = \begin{cases} \pi_j & \text{if } (i,j) \text{ are adjacent bonds} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the i th edge degree calculated on the weighted edge adjacency matrix is the sum of the bond orders associated with all ϵ_i bonds adjacent to the i th edge:

$$\pi \epsilon_i = \sum_{j=1}^B [\pi \mathbf{E}]_{ij} = \sum_{j=1}^B [\mathbf{E}]_{ij} \cdot \pi_j$$

The **bond order-weighted edge connectivity index** ${}^\pi \epsilon$ is then defined as [Estrada and Ramirez, 1996]

$${}^\pi \epsilon = \sum_k (\pi \epsilon_i \cdot \pi \epsilon_j)^{-1/2}_k$$

where k runs over all connections N_2 .

Characteristics of this index are its sensitivity to the presence of heteroatoms and heteroatom position in the molecule (greater values referring to central positions), and the discriminating power of conformational isomers.

By analogy with the bond order-weighted edge adjacency matrix, a **resonance-weighted edge adjacency matrix** ${}^k \mathbf{E}$ was also proposed, replacing the bond orders with parameters k_{C-X} used in the Hückel matrix and related to the resonance integral β_{C-X} of the bond between the heteroatom X and the carbon atom by the relationship:

$$\beta_{C-X} = k_{C-X} \cdot \beta_{C-C}$$

where β_{C-C} is the resonance integral of the carbon-carbon bond [Estrada, 1995b]. The literature reports several values for k_{C-X} parameters and some proposed by Estrada are listed in Table E2.

Table E2 Values of the k_{C-X} constants proposed by Estrada.

C–X bond	k_{C-X}	C–X bond	k_{C-X}
C–C	1.0	C–S	0.7
C–N	1.0	C–F	0.7
C–O	0.8	C–Cl	0.4
C=O	1.6	C–Br	0.3

The **resonance-weighted edge connectivity index** ${}^k \epsilon$ is derived from the row sums of the above-defined matrix in the same way as the edge connectivity index ϵ and it is sensitive to the presence of heteroatoms and multiple bonds in the molecule.

The **electronegativity-weighted edge connectivity index**, denoted by ${}^m F$, was calculated from an electronegativity-weighted edge adjacency matrix ${}^x \mathbf{E}$ and by the formula of the extended edge connectivity indices [Mu and Feng, 2004; Mu, Feng *et al.*, 2006]:

$${}^m F = \sum_{k=1}^K \left(\prod_{i=1}^m {}^x \epsilon_i \right)^{-1/2}_k$$

where k runs over all of the m th-order path subgraphs, m being the number of edges in the path subgraph, K is the total number of m th-order paths. For each path, the electronegativity-weighted edge degrees χ_{ϵ_i} of all edges involved in the path are multiplied.

To calculate this weighted edge connectivity index, the following weighting scheme is adopted:

$$w_i = \frac{\chi_{i(1)}^h + \chi_{i(2)}^h}{\chi_C^h + \chi_C^h} = \frac{\chi_{i(1)}^h + \chi_{i(2)}^h}{4.96}$$

where χ^h denotes the hybridized-dependent electronegativities (Table E3) and $i(1)$ and $i(2)$ refer to the two atoms forming the i th bond.

Table E3 Hybridized-dependent electronegativities for carbon and oxygen atoms.

Atom hybrid	C _{sp³}	C _{sp²}	C _{sp}	O _{sp³}	O _{sp²}
χ^h	2.48	2.75	3.29	4.93	5.54

By analogy with the $\rightarrow$ *augmented adjacency matrix*, the **augmented edge adjacency matrix** ${}^a\mathbf{E}(w)$ can be derived from an edge-weighted molecular graph, for any $\rightarrow$ *weighting scheme* w as

$$[{}^a\mathbf{E}(w)]_{ij} = \begin{cases} 1 & \text{if } (i,j) \text{ are adjacent bonds} \\ w_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where the diagonal elements are different from zero in order to encode information about the different bonds in the molecule and edge adjacencies are codified as in the standard edge adjacency matrix.

The **bond distance-weighted edge adjacency matrix** ${}^a\mathbf{E}(r)$ is an augmented edge adjacency matrix based on bond lengths calculated by computational chemistry methods [Estrada, 1997]:

$$[{}^a\mathbf{E}(r)]_{ij} = \begin{cases} 1 & \text{if } (i,j) \text{ are adjacent bonds} \\ r_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where r_i is the bond length associated to the i th edge. Bond lengths are used as weights for edges of the molecular graph, thus allowing discrimination among the different isomers, heteroatoms, conformations, and so on.

The **spectral moments of the bond distance-weighted edge adjacency matrix** were defined [Estrada, 1997; Estrada, 1998a] as

$$\mu^k(r) = \text{tr}[{}^a\mathbf{E}^k(r)]$$

where k is the power of the bond distance-weighted edge adjacency matrix and tr the trace of this matrix. In a similar way, spectral moments of any weighted edge adjacency matrix can be calculated by using different weighting schemes for the bonds.

A graph-theoretical approach called **TOPological SubStructural MOlecular DEsign (TOPS-MODE)**, previously called **TOPological SubStructural MOlecular DEsign (TOSS-MODE)**, was

proposed to express physical and biological properties in terms of substructural features of molecules by the spectral moments of the edge adjacency matrix [Estrada, 1998b, 2008; Estrada, Peña *et al.*, 1998; Estrada and Uriarte, 2001a, 2003; Estrada, Molina *et al.*, 2001b]. The main steps to be conducted for the application of this approach to QSAR/QSPR modeling are (a) compute the spectral moments of the weighted edge adjacency matrix with the selected weighting scheme for each molecule in the data set; (b) find a QSAR/QSPR model by using any appropriate linear or nonlinear multivariate statistical technique; and (c) replace the spectral moments in the QSAR/QSPR model with their expressions in terms of the contributions of the different structural fragments of the molecule, obtaining an equation that relates the property directly with the molecular structure and thus allows $\rightarrow$ *reversible decoding*.

Applications of the TOPS-MODE approach reported in literature are [Estrada, Gutierrez *et al.*, 2000; Estrada, Uriarte *et al.*, 2000; Estrada and Peña, 2000; Estrada, Molina *et al.*, 2001a; Estrada, Vilar *et al.*, 2002; Estrada, Patlewicz *et al.*, 2003; Estrada and Gonzalez, 2003; Pérez González and Helguera, 2003; Pérez González, Gonzalez *et al.*, 2003, 2003; Estrada, Patlewicz *et al.*, 2004; Estrada, Quincoces *et al.*, 2004; Pérez González and Moldes Teran, 2004; Pérez González, Helguera Morales *et al.*, 2004; Pérez González, Helguera *et al.*, 2004; Vilar, Estrada *et al.*, 2005; Amić, Davidović-Amić *et al.*, 2007].

📖 [Estrada, Rodriguez *et al.*, 1997; Nikolić and Trinajstić, 1998, 1998; Marković, 1999; Estrada and Peña, 2000; Estrada, Uriarte *et al.*, 2000, 2003; Estrada, Gutierrez *et al.*, 2000; Estrada, Molina *et al.*, 2001a, 2001b; Estrada and Uriarte, 2001a; Marković, Marković *et al.*, 2001, 2001b; Estrada, Vilar *et al.*, 2002; Estrada and Gonzalez, 2003; Estrada, Patlewicz *et al.*, 2003; Marković, 2003; Pérez González and Helguera, 2003; Pérez González, Gonzalez *et al.*, 2003, 2004; Estrada, Quincoces *et al.*, 2004; Pérez González, Helguera Morales *et al.*, 2004, 2006; Pérez González and Moldes Teran, 2004; Pérez González, Helguera *et al.*, 2004; Vilar, Estrada *et al.*, 2005; Helguera Morales, Cabrera Pérez *et al.*, 2006]

- **edge-adjacency-plus-edge-distance matrix** $\rightarrow$ Schultz molecular topological index
- **edge centric indices** $\rightarrow$ centric indices
- **edge centric indices for multigraphs** $\rightarrow$ centric indices
- **edge chromatic decomposition** $\rightarrow$ chromatic decomposition
- **edge chromatic information index** $\rightarrow$ chromatic decomposition
- **edge chromatic number** $\rightarrow$ chromatic decomposition
- **edge-Cluj matrices** $\rightarrow$ Cluj matrices
- **edge complete centric index** $\rightarrow$ centric indices
- **edge connectivity** $\rightarrow$ connectivity indices
- **edge connectivity index** $\rightarrow$ edge adjacency matrix
- **edge-connectivity matrix** $\equiv \chi^E$ *matrix* $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **edge counting** $\equiv$ *bond number*
- **edge-cycle incidence matrix** $\rightarrow$ incidence matrices ($\odot$ cycle matrices)
- **edge cyclic degree** $\rightarrow$ incidence matrices ($\odot$ cycle matrices)
- **edge degree** $\rightarrow$ edge adjacency matrix
- **edge degree count** $\rightarrow$ edge adjacency matrix
- **edge degree-distance index** $\rightarrow$ Cao-Yuan indices
- **edge-degree Zagreb indices** $\rightarrow$ Zagreb indices
- **edge distance code** $\rightarrow$ edge distance matrix

- **edge distance code centric index** → centric indices
- **edge distance counts** → edge distance matrix
- **edge distance degree** → edge distance matrix
- **edge distance degree centric index** → centric indices
- **edge-distance-edge-degree matrix** → distance-degree matrices
- **edge distance index** $\equiv$ *edge distance degree* → edge distance matrix

■ **edge distance matrix** (${}^E\mathbf{D}$)

Usually derived from the → *H-depleted molecular graph* G , the edge distance matrix is the edge analogue of the vertex → *distance matrix*, and summarizes in matrix form the topological distance information among all the pairs of bonds [Bonchev, 1983; Estrada and Gutman, 1996]. It is simply the distance matrix of the → *line graph* of G .

The **topological edge distance** ${}^E\mathbf{D}_{ij}$ between the edges e_i and e_j is defined as the length of the shortest → *path* between them, that is, number of vertices in the shortest path connecting edges e_i and e_j not counting the terminal vertices of the path.

As each i th edge is characterized by two vertices $v_{i(1)}$ and $v_{i(2)}$ incident to the edge, the topological edge distance between the edges e_i and e_j can also be obtained from the minimum → *topological distance* d between two pairs of vertices as

$${}^E\mathbf{D}_{ij} = \min\{d_{i(1)j(1)}, d_{i(1)j(2)}, d_{i(2)j(1)}, d_{i(2)j(2)}\} + 1$$

For acyclic graphs, the topological edge distance can be calculated by the following formula:

$${}^E\mathbf{D}_{ij} = \frac{1}{4} \cdot (d_{i(1)j(1)} + d_{i(1)j(2)} + d_{i(2)j(1)} + d_{i(2)j(2)})$$

The off-diagonal entries ${}^E\mathbf{D}_{ij}$ of the edge distance matrix are equal to 1 if edges e_i and e_j are adjacent (i.e., the edges e_i and e_j are connected and ${}^E\mathbf{D}_{ij} = [\mathbf{E}]_{ij} = 1$, where $[\mathbf{E}]_{ij}$ denotes the elements of the → *edge adjacency matrix* $\mathbf{E}$), otherwise they are more than 1. The diagonal elements are equal to zero. The edge distance matrix, usually derived from a → *H-depleted molecular graph*, is square symmetric with dimension $B \times B$, where B is the number of bonds.

The maximum value entry in the i th row is called **bond eccentricity** ${}^b\eta_i$ (or **edge eccentricity**):

$${}^b\eta_i = \max_j ({}^E\mathbf{D}_{ij})$$

From the eccentricity definition, a graph G can be immediately characterized by two molecular descriptors known as **topological radius from edge eccentricity** ${}^ER$ and **topological diameter from edge eccentricity** ${}^ED$. The radius is defined as the minimum bond eccentricity and the diameter as the maximum bond eccentricity, according to:

$${}^ER = \min_i ({}^b\eta_i) \quad \text{and} \quad {}^ED = \max_i ({}^b\eta_i)$$

From the edge distance matrix several → *topological information indices* are calculated. Moreover, the atomic and molecular descriptors already defined for the vertex distance matrix are analogously defined for the edge distance matrix.

From the frequencies of the matrix row entries, the **edge distance code** is defined as the ordered sequence of the occurrence of increasing edge distance values for the i th considered edge,

$$\{^1f_i, ^2f_i, ^3f_i, \dots, {}^{b_{\eta_i}}f_i\}$$

where $^1f_i, ^2f_i, ^3f_i, \dots$, called **edge distance counts**, indicate the frequencies of the edge distances equal to 1, 2, 3, ..., respectively, from edge e_i to any other edge and b_{η_i} is the i th bond eccentricity.

The **edge distance degree** (or **edge distance index**, **edge distance sum**) is the row sum ${}^E\sigma_i$ obtained by summing the i th row entries of the edge distance matrix:

$${}^E\sigma_i \equiv VS_i({}^E\mathbf{D}) = \sum_{j=1}^B [{}^E\mathbf{D}]_{ij} = \sum_{k=1}^{b_{\eta_i}} {}^kf_i \cdot k$$

where VS_i is the $\rightarrow$ row sum operator and ${}^kf_i$ is the edge distance count of k th-order, which runs over the different edge distance values.

The sum of the edge distance degrees, that is, the sum of all matrix elements, is called **total edge distance** D_E and defined as

$$D_E = \sum_{i=1}^B \sum_{j=1}^B [{}^E\mathbf{D}]_{ij} = \sum_{i=1}^B {}^E\sigma_i$$

where B is the number of graph edges.

$A \rightarrow$ Wiener-type index, called **edge Wiener index** and denoted as ${}^EW$, can be obtained from the edge distance matrix ${}^E\mathbf{D}$ as

$${}^EW \equiv Wi({}^E\mathbf{D}) \equiv \frac{1}{2} \cdot \sum_{i=1}^B \sum_{j=1}^B [{}^E\mathbf{D}]_{ij}$$

where Wi is the $\rightarrow$ Wiener operator and B the number of graph edges [Gutman and Estrada, 1996]. It was demonstrated that ${}^EW$ differs from the standard $\rightarrow$ Wiener index only by a constant and, consequently, reflects the same molecular structure features, that is,

$${}^EW = W - \frac{A \cdot (A-1)}{2}$$

where A is the number of vertices in G , that is, the number of heavy atoms in the molecule considered.

More interesting is the $\rightarrow$ edge-type Schultz index derived from both the edge distance matrix and the edge adjacency matrix.

Bond multiplicity is taken into account by augmenting the edge distance matrix with a supplementary column and row for each multiedge, therefore obtaining an **edge distance matrix for multigraphs** [Bonchev, 1983]. All the local vertex invariants and molecular descriptors defined above can also be calculated on this matrix.

Moreover, if the chemical bond distances are calculated in real molecules, a **geometric edge distance matrix**, denoted as ${}^EG$, can be obtained from which analogous geometric descriptors are derived. Each entry of this matrix is calculated from the interatomic distances between the end point vertices of the pair of edges considered:

$$[{}^EG]_{ij} = \frac{1}{4} \cdot (r_{i(1),j(1)} + r_{i(1),j(2)} + r_{i(2),j(1)} + r_{i(2),j(2)})$$

where r is the Euclidean distance; $i(1)$ and $i(2)$ are the vertices incident to the i th edge, while $j(1)$ and $j(2)$ are the vertices incident to the j th edge.

The **reciprocal edge distance matrix** (or **edge Harary matrix**), denoted as ${}^E\mathbf{D}^{-1}$, is a square symmetric $B \times B$ matrix whose off-diagonal entries are the reciprocal distances between edges considered:

$$[{}^E\mathbf{D}^{-1}]_{ij} = \begin{cases} 1/[{}^E\mathbf{D}]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The diagonal entries are zero by definition. This matrix is the $\rightarrow$ *reciprocal distance matrix* of the line graph of the molecular graph G [Ivanciuc, Ivanciuc *et al.*, 1997].

📖 [Estrada, Rodriguez *et al.*, 1997; Estrada and Rodriguez, 1997]

- **edge distance matrix for multigraphs** $\rightarrow$ edge distance matrix
- **edge distance sum** $\equiv$ *edge distance degree* $\rightarrow$ edge distance matrix
- **edge eccentricity** $\equiv$ *bond eccentricity* $\rightarrow$ edge distance matrix
- **edge-Gutman index** $\rightarrow$ Schultz molecular topological index
- **edge-Harary index** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **edge Harary matrix** $\equiv$ *reciprocal edge distance matrix* $\rightarrow$ edge distance matrix
- **edge-Hosoya matrix** $\rightarrow$ Hosoya Z matrix
- **edge layer matrix** $\rightarrow$ layer matrices
- **edge matrices** $\rightarrow$ matrices of molecules
- **edge orbital information content** $\rightarrow$ orbital information indices
- **edge radial centric information index** $\rightarrow$ centric indices
- **edges** $\rightarrow$ graph
- **edge-Schultz index** $\rightarrow$ Schultz molecular topological index
- **edge-Szeged matrices** $\rightarrow$ Szeged matrices
- **edge type count** $\rightarrow$ edge adjacency matrix
- **edge-valence-weighted Schultz distance matrix** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **edge-vertex incidence matrix** $\rightarrow$ incidence matrices
- **edge-vertex-valence-weighted Schultz distance matrix** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **edge-vertex-weighted Schultz distance matrix** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **edge-weighted Schultz distance matrix** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **edge-weighted detour matrix** $\equiv$ *weighted detour matrix* $\rightarrow$ detour matrix
- **edge-weighted Harary index** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **edge-weighted Harary matrix** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **edge Wiener index** $\rightarrow$ edge distance matrix
- **edge-Wiener matrix** $\rightarrow$ Wiener matrix
- **edge-XI matrix** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **edge-Zagreb matrix** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **edge- χ matrix** $\equiv$ χ *matrix* $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **EEVA descriptors** $\rightarrow$ EVA descriptors

- **Effective Dose** → biological activity indices (⊙ pharmacological indices)
- **effective resonance constant** → electronic substituent constants (⊙ resonance electronic constants)
- **effective solute hydrogen-bond acidity** → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **effective solute hydrogen-bond basicity** → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **EFVCI** ≡ *External Factor Variable Connectivity Index* → variable descriptors
- **eigenvalue-based descriptors** ≡ *spectral indices*
- **EigenValue descriptors** ≡ *EVA descriptors*
- **eigenvalues** → algebraic operators (⊙ characteristic polynomial)
- **eigenvalues of the adjacency matrix** → spectral indices
- **eigenvalues of the distance matrix** → spectral indices
- **eigenvectors** → algebraic operators (⊙ characteristic polynomial)
- **eigenvector centrality** → center of a graph
- **EIM** ≡ *Electronic Indices Methodology* → quantum-chemical descriptors (⊙ EIM descriptors)
- **EIM descriptors** → quantum-chemical descriptors
- **electrical conductance matrix** ≡ *conductance matrix* → resistance matrix
- **electric dipole moment** ≡ *dipole moment* → electric polarization descriptors
- **electric permittivity** → physico-chemical properties (⊙ dielectric constant)

■ electric polarization descriptors

Electric polarization, dipole moments, and other related physical quantities, such as multipole moments and polarizabilities, constitute a group of both local and molecular descriptors, which can be defined either in terms of classical physics or in terms of quantum mechanics. They encode information about the charge distribution in molecules [Böttcher, van Belle *et al.*, 1973]. They are particularly important in modeling solvation properties of compounds that depend on solute/solvent interactions and in effect frequently used to represent the → *dipolarity/polarizability term* in → *Linear Solvation Energy Relationships*. Moreover, they can be used to model the polar interactions that contribute to determine → *lipophilicity* of compounds.

The **dipole moment** μ (or **electric dipole moment**) is a vectorial quantity that encodes displacement with respect to the center of gravity of positive and negative charges in a molecule, defined as

$$\mu = \sum_i q_i \cdot \mathbf{r}_i$$

where q_i are point charges located at positions $\mathbf{r}_i$. The SI unit for dipole moments is the coulomb meter, but they are often expressed in debye.

The elements of the vector μ are called **dipole moment components**:

$$\mu_x = \sum_i q_i \cdot x_i \quad \mu_y = \sum_i q_i \cdot y_i \quad \mu_z = \sum_i q_i \cdot z_i$$

where x, y, z are the coordinates of the charges. Molecules with zero dipole moments are called *nonpolar*, others *polar*; moreover, dipole moments equal to zero indicate molecules with a center of symmetry.

In analogy to the definition of electric dipole moment, electric multipole moments are also defined. In particular, the **quadrupole moment** Q and the **octupole moment** U are defined as

$$\mathbf{Q} = \frac{1}{2!} \cdot \sum_{i=1}^A q_i \cdot \mathbf{r}_i \cdot \mathbf{r}_i \quad \mathbf{U} = \frac{1}{3!} \cdot \sum_{i=1}^A q_i \cdot \mathbf{r}_i \cdot \mathbf{r}_i \cdot \mathbf{r}_i$$

where $\mathbf{Q}$ is a tensor of second degree (a 3×3 symmetric matrix) and $\mathbf{U}$ a tensor of third degree (a $3 \times 3 \times 3$ symmetric matrix).

For example, when the charge distribution is spherically symmetrical, all the diagonal terms of the quadrupole moment are equal to zero. Therefore, the trace of the electric quadrupole moment is a measure of molecular charge distribution deviation from sphericity.

When a molecule is embedded in a uniform electric field $\mathbf{E}_0$ in vacuum, an **induced dipole moment** μ_{IND} arises, defined by the relationship:

$$\mu_{IND} = \alpha \cdot \mathbf{E}_0$$

where the scalar constant of proportionality is called **polarizability** α (or **static polarizability**). This scalar polarizability may be regarded as the sum of the **electronic polarizability** α_E and the **atom polarizability** α_A . A polarizable molecule shows an induced dipole moment different from zero.

For substituent groups, the **excess electron polarizability** $\Delta\alpha_E$ was also defined [Dearden, Bradburne *et al.*, 1991] as the difference between the calculated electron polarizability for straight chain alkyl groups by using a model based on the $\rightarrow$ McGowan characteristic volume V_X and the effective electron polarizability of the substituent:

$$\Delta\alpha_E = \alpha_E - [0.135 \cdot V_X + 0.052]$$

In general, the scalar polarizability α is not sufficient to describe the induced polarization; therefore, a **polarizability tensor** $\vec{\alpha}$ is used to better encode induced polarization and represents **molecular polarizability**. In such a general case, the induced dipole moment needs not to have the same direction as the applied field, but the direction will depend on the position of the molecule relative to the polarizing field [Miller and Savchik, 1979; Miller, 1990c].

Each molecule, polar or nonpolar, is polarizable, that is, its electrons can be shifted under an electric field $\mathbf{E}$ so that **polarization** $\mathbf{P}$ is induced in the molecule. The polarization is proportional to the electric field strength $\mathbf{E}$ and the simplest relationships between $\mathbf{P}$ and $\mathbf{E}$ are

$$\mathbf{P} = \chi^e \cdot \mathbf{E} = \frac{\epsilon - 1}{4\pi} \cdot \mathbf{E}$$

where the scalar proportionality constant χ^e is the $\rightarrow$ *dielectric susceptibility*. In the second equation, ϵ is the $\rightarrow$ *dielectric constant*.

Polarization can be factorized in two main contributions: **induced polarization** P_α , due to translation effects, and **dipole polarization** P_μ , due to orientation of permanent dipoles. Moreover, induced polarization can be viewed as being due to the contribution of **electronic polarization** P_E and **atomic polarization** P_A :

$$\mathbf{P} = \mathbf{P}_\alpha + \mathbf{P}_\mu = \mathbf{P}_E + \mathbf{P}_A + \mathbf{P}_\mu$$

Other important quantities related to polarization and dipole moments are listed below.

• molar polarization

The dipole moment induced per unit of volume V is called molar polarization P_M and is defined by the Clausius–Mossotti equation as

$$P_M = \frac{\epsilon - 1}{\epsilon + 2} \cdot \frac{MW}{\rho} = \frac{4\pi}{3} \cdot N_A \cdot \alpha = \frac{n_D^2 - 1}{n_D^2 + 2} \cdot \frac{MW}{\rho} = MR$$

where MW is the molecular weight, ρ the density, and ϵ the dielectric constant; in the second equation, N_A is the Avogadro number, α the scalar polarizability, and MR the $\rightarrow$ *molar refractivity*. For high frequency fields the relationship $\epsilon = n^2$ holds, where n is the $\rightarrow$ *refractive index*; the subscript D indicates the value of the refractive index corresponding to the sodium D-line, as it is usually used.

• atom polarizability

This is the polarization effect at atomic level, where dipoles $\mu_{IND,i}$ are induced on each atom as

$$\mu_{IND,i} = \alpha_i \cdot E_i$$

where E_i is the electric field at the i th atom and α_i the corresponding polarizability, assumed to be isotropic. Atom polarizabilities are linearly correlated with their $\rightarrow$ *hardness* [Politzer, 1987]. Atomic contributions to polarizability (Table E4) were estimated by several authors [Kang and Jhon, 1982; Miller, 1990b; No, Cho *et al.*, 1993] and are used to calculate the mean polarizability of a molecule by summing the atomic contributions.

Table E4 Atomic polarizability values: (1) Kang–Jhon atomic hybrid polarizabilities $\alpha_A(ahp)$; (2) Miller–Savchik average atomic polarizabilities $\alpha_A^*(ahc)$; (3) No–Cho–Jhon–Sheraga atomic polarizabilities $\alpha_{ij,0}^*$; and linear charge coefficient α_{ij} .

Atom/hybrid	1 α_A	2 α_A^*	3 $\alpha_{ij,0}^*$	4 α_{ij}
H bonded to Xsp ₃	0.386	0.392	0.396	0.219
H bonded to Xsp ₂	0.386	0.392	0.298	0.404
C sp ₃	1.064	1.116	1.031	0.590
C sp ₂ arom.	1.382	1.369	1.450	0.763
C sp ₂ carbonyl			1.253	0.862
C sp ₂ ethylene			1.516	0.568
C sp	1.279	1.294		
N sp ₃	1.094	1.077	0.966	0.437
N sp ₂ amide			0.821	0.422
N sp ₂ pyrrole	1.090	0.851	0.871	0.424
N sp ₂ pyridine	1.030	0.910	0.656	0.436
N sp	0.852	0.972		
O sp ₃	0.664	0.780	0.623	0.281
O sp ₂ carbonyl	0.460	0.739	0.720	0.347
O sp ₂ aromatic	0.422	0.586	0.720	0.347
S sp ₃	3.000 ^a	3.056	2.688	1.319

(Continued)

Table E4 (Continued)

Atom/hybrid	1 α_A	2 α_A^*	3 $\alpha_{ij,0^*}$	4 α_{ij}
S sp ₂ thione	3.729 ^a	3.661		
S sp ₂ aromatic	2.700 ^a	2.223		
P sp ₃	1.538 ^a	1.647		
F	0.296 ^a	0.527	0.226	0.144
Cl	2.315 ^a	2.357	2.180	1.089
Br	3.013 ^a	3.541	3.114	1.402
I	5.415 ^a	5.573	5.166	2.573

Data from Miller [Miller, 1990b].

• mean polarizability

The molecular polarizability is a tensor when the molecule is not perfectly spherical. The mean polarizability α of a molecule is calculated by the relation

$$\alpha = \frac{\alpha_{xx} + \alpha_{yy} + \alpha_{zz}}{3}$$

where α_{xx} , α_{yy} , and α_{zz} are the polarizability along each principal component axis of the molecule, obtained by diagonalization of the polarizability tensor [Cartier and Rivail, 1987]. When, for practical purposes, the effect due to anisotropy of polarizability is small, the mean polarizability is calculated simply as

$$\alpha = \frac{tr(\vec{\alpha})}{3}$$

where $tr(\vec{\alpha})$ is the trace of the nondiagonalized polarizability tensor $\vec{\alpha}$.

Molecular polarizability can also be approximated by simply summing atomic polarizabilities over all the molecule atoms. Moreover, molecular polarizability α expressed as polarizability volume is called **polarizability volume** and is defined as

$$\alpha' = \frac{\alpha}{4\pi\epsilon_0}$$

where ϵ_0 is the $\rightarrow$ dielectric constant in vacuum.

• atom-atom polarizability

Index of chemical reactivity, denoted as π_{ab} , is among the $\rightarrow$ quantum-chemical descriptors and calculated from the perturbation theory as

$$\pi_{ab} \equiv AAP_{ab} = 4 \cdot \sum_{i=1}^{N_{occ}} \sum_{j=1}^{N_{occ}} \sum_{\mu} \sum_{\nu} \frac{c_{i\mu,a} \cdot c_{j\mu,a} \cdot c_{i\nu,b} \cdot c_{j\nu,b}}{\epsilon_i - \epsilon_j}$$

where i and j run over the molecular orbitals, ϵ_i and ϵ_j denote their corresponding energies, μ and ν run over the atomic orbitals of the atoms a and b , and c denotes the coefficients of the linear combination of atomic orbitals ϕ defining each molecular orbital ψ .

The **self-atom polarizability** π_{aa} is analogously defined as

$$\pi_{aa} \equiv SAP_{aa} = 4 \cdot \sum_{i=1}^{N_{occ}} \sum_{j=1}^{N_{occ}} \sum_{\mu} \sum_{\nu} \frac{c_{i\mu,a}^2 \cdot c_{j\nu,a}^2}{\epsilon_i - \epsilon_j}$$

An index for **total self-atom polarizability** is obtained as

$${}^T\pi = \sum_{a=1}^A \pi_{aa}$$

where π_{aa} is the self-atom polarizability of the a th atom and A is the number of atoms in the molecule.

- **anisotropy of the polarizability (β^2)**

A measure of the deviation of the molecular polarizability from a spherical shape, defined as

$$\beta^2 = \frac{(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{yy} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{xx})^2}{2}$$

where α_{xx} , α_{yy} , and α_{zz} are the polarizabilities along each principal component axis of the molecule, obtained by diagonalization of the polarizability tensor $\vec{\alpha}$.

Some empirical **polarity/polarizability descriptors**, which were proposed to measure the ability of the compound to influence a neighboring charge or dipole by virtue of dielectric interactions, are discussed below.

- **electrostatic factor**

A molecular descriptor proposed for solvent classification and defined as

$$EF = \epsilon \cdot \mu$$

where ϵ is the dielectric constant and μ the magnitude of the dipole moment of the solvent.

In general, values between 0 and 2 indicate hydrocarbon solvents, between 3 and 20 electron-donor solvents, between 20 and 50 hydroxylic solvents, and values greater than 50, dipolar aprotic solvents.

- **Polarizability Effect Index (PEI)**

The polarizability effect index is based on the stabilizing energy E_X caused by the polarizability effect for a substituent X interacting with a point charge q [Cao and Li, 1998]. For alkyl and aliphatic alcohol substituents, the stabilizing energy E_X is defined as:

$$E_X = K \cdot \sum_i \left[N_i \cdot \frac{1 + \cos \theta}{q - \cos \theta} - \frac{2 \cdot \cos \theta \cdot (1 - (\cos \theta)^{N_i})}{(1 - \cos \theta)^2} \right]^{-2} = K \cdot \text{PEI}(X)$$

where $K = -2.16q^2/(2\epsilon r_{CC}^4)$, the constant -2.16 being the average polarizability α value of the four basic alkyl units (CH_3 , CH_2 -, CH -, and $-\text{C}$ -), ϵ the effective dielectric constant and r_{CC} the carbon-carbon bond length; the summation goes over all the basic units of the substituent X . N_i is the number of heavy atoms between the i th alkyl unit and the probe with charge q ; for each atom located at unit distance from the probe, N_i represents the topological distance from the probe. θ is the supplementary angle of bond angle CCC (i.e. $\theta = 180^\circ - 109.5^\circ = 70.5^\circ$ for sp^3 hybridization). In the last term of the expression of E_X , $\text{PEI}(X)$ is the relative order of

Table E5 Values of the ΔPEI contributions at different topological distance from the point charge center.

N_i	ΔPEI	N_i	ΔPEI	N_i	ΔPEI	N_i	ΔPEI
1	1.000000	6	0.009052	11	0.002375	16	0.001073
2	0.140526	7	0.006388	12	0.001972	17	0.000945
3	0.048132	8	0.004748	13	0.001628	18	0.000838
4	0.025503	9	0.003666	14	0.001421	19	0.000749
5	0.013800	10	0.002916	15	0.001229	20	0.000673

polarizability effect of alkyl substituents, called polarizability effect index; this can be calculated as the following:

$$\text{PEI}(\text{X}) = \sum_i \Delta\text{PEI}_i(\text{X})$$

where $\Delta\text{PEI}_i(\text{X})$ is the contribution of the i th basic alkyl unit of the substituent. In Table E5, values of $\Delta\text{PEI}_i(\text{X})$ are reported for different values of N_i .

Taking each i th carbon atom of an alkane molecule as the beginning one in the action of the probe, a $\text{PEI}(\text{X}_i)$ value can be calculated in the same way as $\text{PEI}(\text{X})$ for the substituent X ; then, the **Molecular Polarizability Effect Index** (MPEI) is derived as [Cao and Li, 1998]:

$$\text{MPEI} = \sum_{i=1}^A \text{PEI}(\text{X}_i)$$

where the summation is over all the carbon atoms in the molecule.

The **Geometric Mean Polarizability Effect Index** (GMPEI) was also proposed for alkanes as [Cao and Yuan, 2002]:

$$\text{GMPEI} = \left[\prod_{i=1}^A \text{PEI}(\text{X}_i) \right]^{1/A}$$

An analogous approach was proposed for alkene derivatives, where the **Geometric Mean Polarizability Effect Index of π bond** (GMPEI π) [Cao and Yuan, 2002] is calculated taking only into account the PEI contributions for the π bond.

The **Inner Molecular Polarizability Index** (IMPI) was defined as [Cao, Liu *et al.*, 1999]:

$$\text{IMPI} = \sum_{i=1}^A \text{PEI}_i$$

where A is the number of carbon atoms in the molecule and PEI_i the sum of the polarizability effect index of the alkyl groups bonded to the i -th atom. This last index differs from MPEI because PEI_i is the polarizability effect index of alkyl groups connected to the i -th carbon atom, whereas $\text{PEI}(\text{X}_i)$ is the polarizability effect index of the whole molecule seen as a substituent X with the i -th carbon atom as the beginning one.

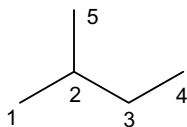
By comparing the IMPI value of an alkane with the IMPI value of the corresponding n -alkane, the quasi-length carbon chain $N_{\text{C}(\text{eff})}$ was defined as:

$$N_{\text{C}(\text{eff})} = \frac{\text{IMPI}_{n\text{-alkane}}}{\text{IMPI}_{\text{isomer}}} \cdot N_{\text{C}}$$

where N_{C} is the number of carbon atoms of the two isomers.

Example E3

Calculation of MPEI, GMPEI, and IMPI for 2-methylbutane.



$$C_1: \text{PEI}(X_1) = 1.00000 + 0.14053 + 2 \times 0.04813 + 0.02350 = 1.2603$$

$$C_2: \text{PEI}(X_2) = 1.00000 + 3 \times 0.14053 + 0.04813 = 1.4697$$

$$C_3: \text{PEI}(X_3) = 1.00000 + 2 \times 0.14053 + 2 \times 0.04813 = 1.3773$$

$$C_4: \text{PEI}(X_4) = 1.00000 + 0.14053 + 0.04813 + 2 \times 0.02350 = 1.2357$$

$$C_5: \text{PEI}(X_5) = 1.00000 + 0.14053 + 2 \times 0.04813 + 0.02350 = 1.2603$$

$$C_1: \text{PEI}_1 = 1.00000 + 2 \times 0.14053 + 0.04813 + 0.02350 = 1.3292$$

$$C_2: \text{PEI}_2 = 3 \times 1.00000 + 0.14053 = 3.1405$$

$$C_3: \text{PEI}_3 = 2 \times 1.00000 + 2 \times 0.14053 = 2.2811$$

$$C_4: \text{PEI}_4 = 1.00000 + 0.14053 + 2 \times 0.04813 = 1.2368$$

$$C_5: \text{PEI}_5 = 1.00000 + 2 \times 0.14053 + 0.04813 = 1.3292$$

$$\text{MPEI} = 1.2603 + 1.4697 + 1.3773 + 1.2357 + 1.2603 = 6.6033$$

$$\text{GMPEI} = [1.2603 \times 1.4697 \times 1.3773 \times 1.2357 \times 1.2603]^{1/5} = 1.3177$$

$$\text{IMPI} = 1.3292 + 3.1405 + 2.2811 + 1.2368 + 1.3292 = 9.3167$$

☞ [Cao, Yuan *et al.*, 2000, 2003; Liu, Liang *et al.*, 2006]

- **Kier–Hall solvent polarity index (${}^1\chi_f^v$)**

It is a solvent polarity index defined as the first-order $\rightarrow$ *valence connectivity index* ${}^1\chi^v$ divided by the number N_f of discrete isolated functional groups to account for multiple interaction sites as [Kier and Hall, 1986]:

$${}^1\chi_f^v = {}^1\chi^v / N_f$$

It was assumed that functional groups influencing solvent polarity are π -electron systems and lone pairs. Thus, for example, N_f (benzene) = 1, N_f (nitrobenzene) = 2 (one π -electron system and one lone pair), N_f (pyridine) = 2 (one π -electron system and one lone pair), N_f (nitro group) = 1. Halogen atoms are considered do not give a significant contribution to a molecule in enhancing its solvent polarity and thus for a halogen-containing molecule ${}^1\chi^v$ is used unmodified.

☞ [Sekusak and Sabljic, 1992]

- **local polarity index (Π)**

It is defined as the average deviation of the surface $\rightarrow$ *molecular electrostatic potential* calculated as [Brinck, Murray *et al.*, 1993; Murray, Brinck *et al.*, 1993]:

$$\Pi = \frac{1}{SA} \cdot \int_0^{SA} |V(\mathbf{r}) - \bar{V}| dSA = \lim_{n \rightarrow \infty} \frac{1}{n} \cdot \sum_{i=1}^n |V(\mathbf{r}_i) - \bar{V}|$$

where $V(\mathbf{r})$ is the potential energy value at i th grid point on the molecular surface SA , $\bar{V}$ is average potential energy value on the surface for the considered molecule, and n is the number of grid points on the molecular surface.

Π ranges from zero for a neutral atom to 21.6 Kcal/mol for water. It is a measure of charge separation or local polarity; it has been shown to correlate with the $\rightarrow$ *dipolarity/polarizability term* Π^* as well as with the $\rightarrow$ *dielectric constant* ϵ . The local polarity index is among the descriptors used in the $\rightarrow$ *GIPF approach* and was used to calculate $\log P$ by the $\rightarrow$ *Politzer hydrophobic model*.

• Q polarity index

This is a topological polarity index derived by the electrotopological $\rightarrow$ *intrinsic state* I of the atoms in the molecule and defined as [Kier and Hall, 1999b]:

$$Q = \frac{A^2 \cdot \sum_{i=1}^A I_i^{ALK}}{\left(\sum_{i=1}^A I_i \right)^2}$$

where A is the number of atoms, I_i^{ALK} is the intrinsic state of the i th atom in the skeleton structure of the molecule in which each nonhydrogen atom is replaced with an sp^3 carbon atom on the corresponding isoconnective alkane as reference structure, and I_i is the intrinsic state of the i th atom in the actual compound. The basic idea is that Q value for the considered molecule lies between two extremes of minimal and maximal polarity: the minimal polarity is given by a molecule only constituted of sp^3 carbon atoms and the maximal polarity is approximated by the square of the number of atoms A in the molecule.

• polar hydrogen factor (Q_H)

This is an index of the molecular polarity due to C–H bonds restricted to halogenated hydrocarbons [Di Paolo, Kier *et al.*, 1979]. It is calculated as the sum of the contributions to the polarity of all the C–H bonds in a molecule. For each C–H bond, three different contributions are considered due to halogens linked to the same carbon atom of the C–H bond, halogens in α -position and halogens in β -position with respect to the considered C–H bond:

$$Q_H = \sum_b \left[\sum_C k_C + \sum_\alpha k_\alpha + \sum_\beta k_\beta \right]$$

where the external summation runs over all the C–H bonds and the three internal summations run over all halogens directly attached to the carbon atom of C–H bond, in α - and β -positions, respectively. When no halogen is attached to a C–H bond, its contribution to Q_H is taken as zero. The only exception is made for a methylene group CH_2 flanked by two halogen-substituted methyl groups: in this case C–H bonds are considered in the calculation.

The constant values k (Table E6) are defined according to the Swain–Lupton $\rightarrow$ *field-inductive constant* F , that is, they represent the relative field effect of an atom or a group.

Table E6 *k* parameters for the polar hydrogen factor Q_H .

Halogen	<i>k</i> Contribution	Halogen	<i>k</i> Contribution
F	0.43	α -Cl	0.10
Cl	0.41	α -Br	0.09
Br	0.44	β -F	0.05
I	0.40	β -Cl	0.05
α -F	0.13	β -Br	0.05

For examples, Q_H values for CHCl_3 , CH_2Cl_2 , $\text{CF}_3\text{-CHFCl}$, and $\text{CF}_3\text{-CH}_2\text{-CF}_2\text{Cl}$ are $3 \times 0.41 = 1.23$, $4 \times 0.41 = 1.64$, $0.43 + 0.41 + 3 \times 0.13 = 1.23$, $2 \times (5 \times 0.13 + 0.10) = 1.50$, respectively.

📖 [Hannay and Smyth, 1946; McClellan, 1963; Buckingham, 1967; Böttcher, van Belle *et al.*, 1973; Exner, 1975; Lien, Liao *et al.*, 1979; Lien, Guo *et al.*, 1982; Li, Guo *et al.*, 1984; Topsom, 1987c; Lewis, 1989; Miller, 1990c; Beck, Glen *et al.*, 1996; Stuer-Lauridsen and Pedersen, 1997; Beck, Horn *et al.*, 1998; Norinder, Sjöberg *et al.*, 1998; Norinder, Österberg *et al.*, 1999; Hansch, Steinmetz *et al.*, 2003]

- **electromeric effect** → electronic substituent constants
- **electron affinity** → quantum-chemical descriptors
- **electron charge density weight** → connectivity indices (⊙ charge-weighted vertex connectivity indices)
- **Electron-Conformational Approach** → Electronic-Topological method
- **Electron-Conformational Matrix of Congruity** → Electronic-Topological method
- **Electron-Conformational method** → Electronic-Topological method
- **Electron-Conformational Submatrix of Activity** → Electronic-Topological method
- **electronic delocalization entropy** → MARCH-INSIDE descriptors
- **electron density** → quantum-chemical descriptors
- **electron donor-acceptor substituent constant** → electronic substituent constants

■ electronegativity

Atom electronegativity is among the most important → *atomic properties*. The concept of atom electronegativity was recognized as a useful basic principle in chemistry more than 150 years ago [Pritchard and Skinner, 1955]. Originally defined by Pauling [Pauling, 1939], electronegativity is “the power of an atom in a molecule to attract electrons to itself.”

The classical definition of atomic electronegativity is due to Mulliken [Mulliken, 1934, 1955a, 1955b] – **Mulliken electronegativity** –:

$$\chi^{\text{MU}} = \frac{\text{IP} + \text{EA}}{2}$$

that is, the arithmetic mean of the → *ionization potential* IP and the → *electronic affinity* EA of the atom. In this definition, the use of ionization potentials and electron affinities of valence states was proposed.

The Mulliken scale of electronegativity (in volts) can be converted in the Pauling scale χ^{PA} (Pauling units) by the empirical relation:

$$\chi^{\text{PA}} = 0.303 \cdot \chi^{\text{MU}}$$

Electronegativity scales, other than Pauling and Mulliken scales, are the Allred–Rochow scale χ^{AR} [Allred and Rochow, 1958, 1961] based on estimated effective nuclear potentials and covalent radii; the Gordy scale χ^{G} [Gordy, 1946, 1951] based on the number of electrons in the valence shell of the atom and the covalent radius; the Sanderson scales χ^{SA} [Sanderson, 1952, 1954, 1955, 1971] based on covalent radii; the Hinze–Jaffé scale χ^{HJ} [Hinze and Jaffé, 1962, 1963b, 1963a; Hinze, Whitehead *et al.*, 1963] based on orbital energies and effective charges; Zhang scale χ^{Z} based on ionization energies and covalent radii [Zhang, 1982b, 1982a]. Atom electronegativity can also be estimated by functions of the $\rightarrow$ vertex degree, as proposed by Kier–Hall ($\rightarrow$ Kier–Hall electronegativity) and by Roy–Ghosh in the framework of $\rightarrow$ ETA indices [Roy and Ghosh, 2003]. An extended review of electronegativity scales was published by Luo and Benson [Luo and Benson, 1990] and some of them are listed in Table E7.

Table E7 Electronegativity values from different sources (Pauling units).

	H	B	C	N	O	F	Si	P	S	Cl	Br	I
Pauling 1	2.1	2.0	2.5	3.0	3.5	4.0	1.8	2.1	2.5	3.0	2.8	2.5
Pauling 2	2.20	2.04	2.55	3.04	3.44	3.98	1.90	2.19	2.58	3.16	2.76	2.66
Mulliken	2.28	2.01	2.63	2.33	3.17	3.91	2.44	1.81	2.41	3.00	2.76	2.56
Allred–Rochow	2.20	2.01	2.50	3.07	3.50	4.10	1.74	2.06	2.44	2.83	2.74	2.21
Sanderson 1	2.31	1.88	2.47	2.93	3.46	3.92	1.74	2.16	2.66	3.28	2.96	2.50
Sanderson 2	2.592	2.275	2.746	3.194	3.654	4.000	2.138	2.515	2.957	3.475	3.219	2.778
Mullay	2.08	1.85	2.47	3.41	3.15	4.00	1.91	1.99	2.49	3.07	2.81	2.47
Gordy	2.17		2.52	2.98	3.45	3.95			2.58	3.50	2.75	2.50
Wells	2.28		2.30	3.35	3.70	3.95			2.80	3.03	2.80	2.47
Boyd–Markus	1.94	1.95	2.53	3.23	3.53	4.00	1.81	2.34	2.65	3.14	2.78	2.48
Inamoto–Masuda	2.00		2.21	2.71	3.02	3.05	1.72	1.93	2.15	2.37	2.32	2.15
Diudea	1.680	1.501	1.831	2.240	2.680	3.024	1.424	1.646	2.026	2.512	2.279	1.879
Zhang	2.271	1.966	2.536	3.062	3.642	4.188	1.769	2.131	2.479	2.835	2.529	2.142

(1) Pauling 1 [Pauling, 1939]; (2) Pauling 2 [Allred and Rochow, 1961]; (3) Mulliken [Mulliken, 1934, 1935a]; (4) Allred–Rochow [Allred and Rochow, 1958]; (5) Sanderson 1 [Sanderson, 1952, 1955]; (6) Sanderson 2 [Sanderson, 1988]; (7) Mullay [Mullay, 1984]; (8) Gordy [Gordy, 1946]; (9) Wells [Wells, 1968a]; (10) Boyd–Markus [Boyd and Markus, 1981]; (11) Inamoto–Masuda [Inamoto and Masuda, 1982]; (12) Diudea [Diudea, Kacso *et al.*, 1996]; (13) Zhang [Zhang, 1982a].

The concept of electronegativity has become increasingly general (perhaps, even ambiguous) during its revision in the different quantum-chemistry frameworks, ranging from atomic electronegativity to orbital and functional group electronegativity up to molecular electronegativity. Electronegativity is a property of the state of the system; electrons tend to flow from a region of low electronegativity to a region of high electronegativity. With the formation of a

molecule, electronegativities of the constituent atoms or fragments equalize, all becoming equal to the electronegativity of the final state of the molecule (**Sanderson's electronegativity equalization principle** [Sanderson, 1951, 1971; Zefirov, Kirpichenok *et al.*, 1987]).

The concept of electronegativity equalization led to the calculation of dipole moments [Malone, 1933; Ferreira, 1963a], bond dissociation energies [Ferreira, 1963b], atomic charges [Stoklosa, 1973], and force constants [Gordy, 1946; Polansky and Derflinger, 1963]. Moreover, based on electronegativity, several $\rightarrow$ *bond ionicity indices* were proposed. To avoid chemically unacceptable results due to the total equalization of the electronegativity in a molecule [Gasteiger and Marsili, 1980], a partial equalization principle was also taken into account as a result of changes in orbital overlap [Pritchard, 1963]. In particular, a **partial equalization of orbital electronegativity (PEOE)** for the calculation of partial atomic charges was proposed using a topological iterative approach [Gasteiger and Marsili, 1980; Marsili and Gasteiger, 1980]; a **modified partial equalization of orbital electronegativity (M-PEOE)** was also successively proposed [No, Grant *et al.*, 1990a, 1990b]. Moreover, equalization based on $\rightarrow$ *density functional theory* was also proposed [Itskowitz and Berkowitz, 1997].

Group electronegativity was proposed for molecular substituents, functional groups, and fragments, where the center atom is the atom connected to the body of the molecule.

Group electronegativity is defined as the electronegativity of the central atom of the substituent and is affected by the neighbors of the central atom. Several methods have been proposed to estimate group electronegativities [Clifford, 1959; McDaniel and Yingst, 1964; Huheey, 1965; Huheey, 1966; Inamoto and Masuda, 1982; Mullay, 1984, 1985; Bratsch, 1985; Xie, Sun *et al.*, 1995]. For example, **Sanderson group electronegativity ESG** is calculated as the geometric mean of the electronegativities of atoms belonging to the considered group [Sanderson, 1983]:

$$\text{ESG}_i = (\chi_1^{\text{SA}} \cdot \chi_2^{\text{SA}} \cdot \dots \cdot \chi_m^{\text{SA}})^{1/m}$$

where χ^{SA} is the Sanderson electronegativity and m the number of atoms of the i th molecular group. Valence group carbon-related electronegativities EC were derived from Sanderson's ESG values by taking into account the covalent radii (as mean bond lengths) relative to the sp^3 carbon atom [Diudea, Kacso *et al.*, 1996].

Another approach to the calculation of group electronegativities χ_G is based on a stepwise addition method [Zhou, Nie *et al.*, 2007]:

$$\chi_G(i) = \frac{1}{n_{i1}} \cdot \sum_{j=1}^{n_{i1}} \chi_{j1} \quad \text{where} \quad \chi_{j1} = \frac{1}{n_{j2}} \cdot \sum_{k=1}^{n_{j2}} \chi_{k2} \cdots \chi_{m(L-1)} = \frac{1}{n_{mL}} \cdot \sum_{q=1}^{n_{mL}} \chi_{qL}$$

where i represents the group focused atom, n_{i1} is the number of atoms directly bonded to the i th atom, χ_{j1} is the Pauling electronegativity of the j th atom at distance one from the i th atom, n_{j2} is the number of atoms directly bonded to the j th atom, which are at distance two from the i th atom, χ_{k2} is the Pauling electronegativity of the k th atom at distance two from the i th atom and bonded to the j th atom, and so on, until the last level corresponding to a distance equal to L from the i th atom (Figure E1).

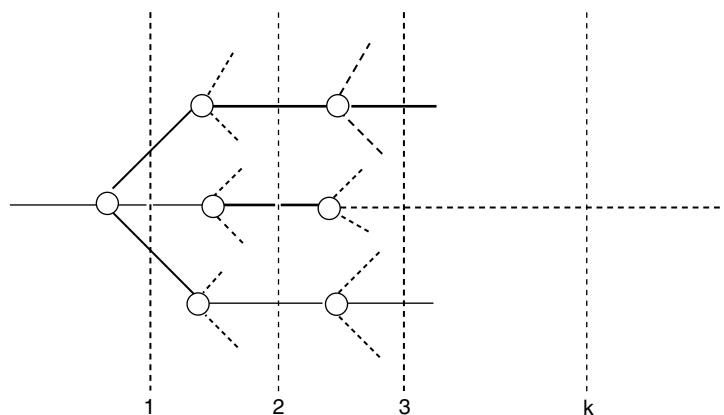


Figure E1 Levels of a group structure.

The equilibrium electronegativity is then defined as

$$\chi_i^{Eq} = \frac{\chi_i + \sum \chi_G(j)}{1 + k}$$

where the summation goes over the electronegativity of the groups directly bonded to the i th atom and k is the number of these contributions.

Electronegativity values, derived from the Pauling electronegativity scale, are shown in Table E8 for some chemical groups.

Table E8 Electronegativity values for some chemical groups from Zhou, Nie *et al.* [Zhou, Nie *et al.*, 2007].

Groups	χ_G	Groups	χ_G	Groups	χ_G	Groups	χ_G
–CH ₃	2.2875	–CH ₂ CH ₃	2.3094	–CH=CH ₂	2.3556	–C≡CH	2.4625
–CN	2.7950	–NCO	3.0100	–OCN	3.1175	–C ₆ H ₅	2.4333
–CHO	2.7300	–COOH	2.9467	–SiH ₃	2.1250	–NH ₂	2.4800
–NO	3.2400	–NO ₂	3.3067	–OPh	2.9367	–OCH ₃	2.8638
–OH	2.8200	–CCl ₃	3.0075	–COCH ₃	2.7592	–COPh	2.8078

Based on equilibrium electronegativity, some molecular descriptors were proposed as, for example, the $\rightarrow$ *Nt index*.

Using group electronegativities as local invariants $\rightarrow$ *electronegativity-based connectivity indices* were proposed as molecular descriptors. Atomic electronegativities are also used in the definition of $\rightarrow$ *MARCH-INSIDE descriptors*.

📖 [Boyd and Edgecombe, 1988; Diudea and Silaghi-Dumitrescu, 1989b; Cherkasov, Galkin *et al.*, 1999, 2000; Cherkasov, 2003; Gilson, Gilson *et al.*, 2003; Agrawal, Gupta *et al.*, 2005; Leyssens, Geerlings *et al.*, 2005]

➤ **electronegativity-based inductive constant** $\rightarrow$ **electronic substituent constants** (⊖ inductive electronic constants)

■ electronegativity-based connectivity indices

These are molecular descriptors defined by analogy with the $\rightarrow$ *Randić connectivity index* and calculated on hydrogen-included molecular graphs where heavy vertices are weighted by valence group electronegativities [Diudea, Kacso *et al.*, 1996].

The first proposed electronegativity-based connectivity index is the **DSI index** [Diudea and Silaghi-Dumitrescu, 1989a] based on the $\rightarrow$ *Sanderson group electronegativity* ESG_i used as the $\rightarrow$ *local vertex invariant*. It is defined as

$$DSI = \sum_{b=1}^B (EVG_i \cdot EVG_j)^{-1/2}$$

where the summation goes over all edges in the graph.

EVG_i is the valence group electronegativity of the i th vertex calculated from the Sanderson group electronegativity (geometric mean of electronegativities of the atoms belonging to the considered group G_i), accounting for atom valence as

$$EVG_i = (ESG_i)^{1/(1+\delta_i)} = \left[\left(\chi_i^{SA} \cdot (\chi_h^{SA})^{h_i} \right)^{1/(1+h_i)} \right]^{1/(1+\delta_i)}$$

where χ_i^{SA} and χ_h^{SA} are the Sanderson electronegativities of the i th heavy atom and the hydrogen atom, respectively, δ_i is the $\rightarrow$ *vertex degree* of the i th atom, and h_i is the number of hydrogen atoms belonging to the group G_i of the i th vertex and calculated as

$$h_i = 8 - L_i - \delta_i$$

where L is the principal quantum number. When $\delta_i > 8 - L_i$, then $h_i = 0$ by definition; moreover, if multiple bonds are present, the vertex degree should be replaced by the sum of the conventional bond orders (i.e., $\rightarrow$ *bond vertex degree* δ_i^b).

An extension of DSI index to paths of higher order was also proposed as

$${}^mDSI = \sum_{m_{p_{ij}}} (EVG_i \cdot EVG_k \cdot \dots \cdot EVG_j)^{-1/2}$$

where the sum is over all the paths of length m in the graph and the product over the valence group electronegativities of all the vertices involved in each path.

Other two electronegativity-based connectivity indices are the **ECP index** and **ECN index** based on the valence group carbon-related electronegativities EC_i proposed by Diudea [Diudea, Kacso *et al.*, 1996]:

$$ECP = \sum_{i=1}^A ecp_i = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (EC_i \cdot EC_j)^{1/2}$$

$$ECN = \sum_{i=1}^A ecn_i = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (EC_i \cdot EC_j)^{-1/2}$$

where ecp_i and ecn_i are local vertex invariants calculated by summing the products of the electronegativity of the considered i th vertex by the electronegativities of its bonded atoms, a_{ij} denotes the elements of the $\rightarrow$ *adjacency matrix* equal to 1 for adjacent vertices and otherwise zero, and A is the number of heavy atoms in the molecule.

- **electronegativity ETA measure** → ETA indices
- **electronegativity scales** → electronegativity
- **electronegativity-weighted adjacency matrix** → weighted matrices ($\odot$ weighted adjacency matrices)
- **electronegativity-weighted edge connectivity index** → edge adjacency matrix
- **electronegativity-weighted walk degrees** → walk counts
- **electronic charge index** $\equiv$ *total absolute atomic charge* → charge descriptors
- **electronic chemical potential** → quantum-chemical descriptors

■ electronic descriptors

These are local or global molecular descriptors related to the electronic distribution in the molecule; they are fundamental to many chemical reactions, physico-chemical properties, and ligand–macromolecule interactions. The theory of electronic density is based on a quantum-mechanical approach; however, → *electronegativity* and charges, which are not physical observables, are also important quantities for the definition of several electronic descriptors.

A lot of → *quantum-chemical descriptors* are derived from the charge distribution in a molecule or the → *electron density* of specified atoms or molecular regions, and from conformational energy values such as the → *Joshi electronic descriptors*. Several → *charge descriptors* and → *electric polarization descriptors* are calculated from atomic charge estimations.

Electronic information is combined with shape and steric information to characterize molecules in → *charged partial surface area descriptors*. Other approaches, different from those closely related to quantum-chemistry, refer to electronic distribution in molecules, such as → *electronic substituent constants*, → *electrotopological state indices*, → *topological charge indices*. → *Reactivity indices* and → *delocalization degree indices* are also related to electronic properties of molecules.

A simple example of electronic descriptors is the **lone-pair electron index**, denoted by LEI, calculated from the → *H-depleted molecular graphs* as [Cheng and Yuan, 2006]:

$$\text{LEI} = \frac{1}{A} \cdot \sum_{i=1}^{n_{\text{het}}} \left[\text{LE}_i \cdot \sum_{j=1}^A \frac{1}{d_{ij}^2} \right] \quad j \neq i$$

where the first summation runs over the heteroatoms and the second one over all the heavy atoms; A is the number of heavy atoms, d_{ij}^2 is the square → *topological distance* between atoms v_i and v_j , and LE_i is the **lone-pair electrostatic interaction** of the i th heteroatom, defined as

$$\text{LE}_i = \frac{\sqrt{L_i} \cdot (Z_i^v - Z_i^b)}{\chi_i^{\text{PA}}}$$

where L is the principal quantum number, χ^{PA} the Pauling → *electronegativity*, Z^v the number of valence electrons, and Z^b the number of bonding electrons.

The values of the lone-pair electrostatic interaction are provided in Table E9.

Table E9 Lone-pair electrostatic interaction values (LE) of atoms.

Atoms	LE	Atoms	LE
C	0	Cl	3.2887
N	0.9304	Br	4.0541
O	1.6444	I	5.0438
F	2.1320		

- **Electronic EigenValue descriptors** $\equiv$ *EEVA descriptors* $\rightarrow$ EVA descriptors
- **Electronic Indices Methodology** $\rightarrow$ quantum-chemical descriptors ($\odot$ EIM descriptors)
- **electronic polarizability** $\rightarrow$ electric polarization descriptors
- **electronic polarization** $\rightarrow$ electric polarization descriptors

■ **electronic substituent constants** ($\equiv$ *Hammett substituent constants*, σ electronic constants)

Derived from the $\rightarrow$ *Hammett equation*, σ electronic constants are calculated for different molecular substituents from the rate or equilibrium constant of specific reactions, with respect to a reference compound [Topsom, 1976, 1987b; Charton, 1981; Taft and Topsom, 1987].

These substituent descriptors are usually used in $\rightarrow$ *Hansch analysis* as the molecular electronic properties in the case of a monosubstituted series of compounds, while they are summed up over all the molecule substituents to obtain global molecular descriptors in the general case of polysubstituted compounds.

The possible modes of action of substituents in modifying the electron distribution of the parent molecule can be distinguished in two main effects: a **polar effect** (or **field-inductive effect** or **localized effect**) and a **resonance effect** (or **delocalized effect**) (Figure E2). The term polar effect is used to characterize the influence of unconjugated, sterically remote substituents on equilibrium or rate processes. The polar substituent effect is transmitted through-bonds (**inductive effect** or **static inductive effect**) involving a polarization either of the σ -bond network (**σ -inductive effect**) or of the π -bond network (**π -inductive effect**). It can also be transmitted through solvent space according to classical laws of electrostatics (**field effect**) by the formation of bond dipole moments. Transmission efficiency of the localized effects is empirically measured by a **transmission coefficient** ε ($0 \leq \varepsilon \leq 1$).

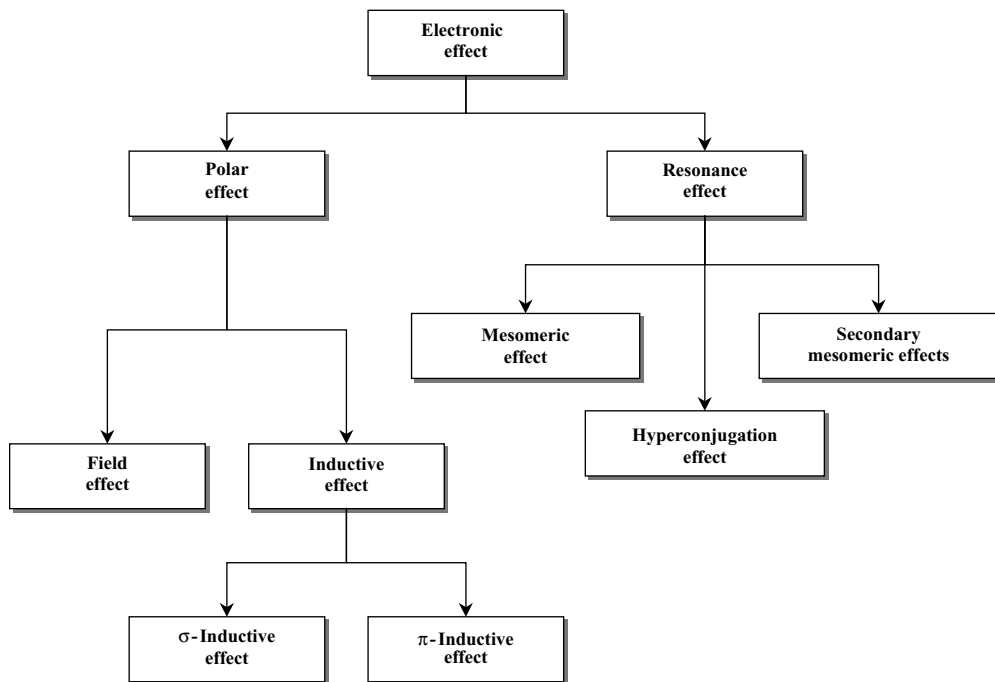


Figure E2 Scheme of the relationships among the electronic effects.

Note that the separability of the two contributions of the polar effect is difficult to attain; in effect, attempts to separate the polar effect contributions have been unsuccessful so they are usually considered together. However, from a theoretical point of view, field effects have been studied using the Kirkwood–Westheimer [Kirkwood and Westheimer, 1938; Westheimer and Kirkwood, 1938] and Tanford models [Tanford, 1957].

Resonance effect is an energy stabilization caused by delocalization of electrons in the bond network of the molecule and can be attributed to a **mesomeric effect**, that is, the delocalization of π electrons on the π orbital network, a **hyperconjugation effect**, that is, a delocalization of σ electrons in a π orbital aligned with the σ bond, and **secondary mesomeric effects**, such as repulsion of the π electrons by nonbonded electrons on a substituent or solvent, or by time-dependent effects due to polarizabilities (for the last, the term **electromeric effect** is sometimes used).

If the substituent is bonded to an sp^3 carbon atom not involved in the π molecular orbitals formation, its electrical effect is only local, assuming that σ electron delocalization is negligible. Substituents bonded to sp or sp^2 carbon atoms can exert both localized and delocalized effects.

Taking into account the basic contributions defined above, the overall electronic effect σ of the substituent can be represented by the following equation:

$$\sigma = \ell \cdot \sigma_I + d \cdot \sigma_R$$

where σ_I and σ_R are the polar and resonance contributions, respectively; ℓ and d weight their importance in determining the overall effect. Equivalent expressions with the same meaning are represented by different symbols by other authors, such as

$$\sigma = f \cdot F + r \cdot R \quad \text{or} \quad \sigma = \lambda \cdot \sigma_D + \delta \cdot \sigma_L$$

From these coefficients, the percentage of the resonance effect $D\%$ can be obtained as

$$D\% = \frac{d}{d + \ell} \cdot 100$$

Several electronic substituent constants were defined so as to represent both global and particular electronic effects. The σ values obtained unambiguously from experimentally accessible data or from the many possible reaction series are called *primary values* and the corresponding set *primary standard*. The σ values derived from the primary values, by rescaling with modified ρ constants or correlation equations, are called *secondary values* and the corresponding set *secondary standard*.

The most popular electronic substituent descriptors are listed below; Table E10 collects the main information concerning all the electronic substituents.

Table E10 Summary of the electronic substituent constants.

ID	Symbol	Reaction	ρ	$D\%$	References
1	σ_p	K; 4-XC ₆ H ₄ COOH	1.00	53	[McDaniel and Brown, 1958]; Lewis and Johnson, 1959]
2	σ_m	K; 3-XC ₆ H ₄ COOH	1.00	22	[McDaniel and Brown, 1958]; [Lewis and Johnson, 1959]
3	σ_p^+	k; 4-XC ₆ H ₄ C(CH ₃) ₂ Cl	-4.54	66	[Brown and Okamoto, 1958]

(Continued)

Table E10 (Continued)

ID	Symbol	Reaction	ρ	D%	References
4	σ_m^+	k ; 3-XC ₆ H ₄ C(CH ₃) ₂ Cl	-4.54	33	[Brown and Okamoto, 1958]
5	σ_p^-	K ; p-XC ₆ H ₄ OH	2.23	56	[Lewis and Johnson, 1959; Hine, 1962; Cohen and Jones, 1963]
6	σ_m^0	K and k ; unbiased values	1.00	23	[Taft, 1960]
7	$\bar{\sigma}_p$	K and k ; effective values	1.00	42	[Taft, 1960]
8	σ^0	k ; HO ⁻ + XC ₆ H ₄ CH ₂ COOEt	0.98	37	[Yukawa, Tsuno <i>et al.</i> , 1966]
9	σ_p^H	K and k ; "normal" values	1.00	47	[van Bekkum]
10	σ^*	k ; XCH ₂ COOY	2.48	6	[Taft, 1952, 1953b, 1956]; [Taft and Lewis, 1958]
11	σ^*	k ; XCOOY	2.48	37	[Taft, 1952, 1953b, 1956]; [Taft and Lewis, 1958]
12	σ^*	k ; X ₂ CHCOOY	2.48	0	[Taft, 1952, 1953b, 1956]; [Taft and Lewis, 1958]
13	σ^*	k ; X ₂ (CH ₂) ₂ COOY	2.48	15	[Taft, 1952, 1953b, 1956]; [Taft and Lewis, 1958]
14	σ^*	k ; X ₂ (CH ₂) ₃ COOY	2.48	33	[Taft, 1952, 1953b, 1956]; [Taft and Lewis, 1958]
15	σ^*	k ; 2-XC ₆ H ₄ COOY	1.00	53	[Taft, 1952, 1953b, 1956]; [Taft and Lewis, 1958]
16	σ'	K ; 4-XC ₈ H ₁₂ COOH	1.464	3	[Roberts and Moreland, 1953]
17	σ'	K ; 4-XC ₈ H ₁₂ COOH	1.65	1	[Holtz and Stock, 1964]; [Baker, Parish <i>et al.</i> , 1967]
18	σ_I	$\sigma_I = 0.45 \sigma^*$ k ; XCH ₂ COOY	6.23	0	[Taft, 1952, 1953b, 1956]; [Taft and Lewis, 1958]
19	σ_R	$\sigma_R = \sigma - \sigma_I$	1.00	92	[Taft, Ehrenson <i>et al.</i> , 1959]; [Taft and Lewis, 1959]; [Taft Price <i>et al.</i> , 1963a]
20	σ_R^0	$\sigma_R^0 = \sigma^0 - \sigma_I$	1.00	84	[Taft, Ehrenson <i>et al.</i> , 1959]; [Taft and Lewis, 1959]; [Taft, Price <i>et al.</i> , 1963a]
21	σ_I^q	k ; XC ₇ H ₁₂ N	1.00	—	[Grob and Schlageter, 1976]
22	σ''	K ; 4-XC ₆ H ₁₀ COOH	1.00	0	[Siegel and Komarmy, 1960]
23	σ_j^{DG}	K ; j-XC ₁₀ H ₈ COOH	1.46	29	[Dewar and Grisdale, 1962b]
24	σ_I	$\sigma_I = b \cdot pK_a^R + a$	—	—	[Charton, 1964]
25	σ_m^I	ionization potential, XC ₆ H ₄ CH ₂ [•]	1.00	26	[Harrison, Kebarle <i>et al.</i> , 1961]
26	σ_p^I	ionization potential, XC ₆ H ₄ CH ₂ [•]	1.00	65	[Harrison, Kebarle <i>et al.</i> , 1961]
27	σ^Q	$\sigma^Q = (f - 34.826)/1.024$ ³⁵ Cl quadrupole resonance	1.00	26	[Bray and Barnes, 1957]
29	σ_p^F	¹⁹ F nmr chemical shift, XC ₆ H ₄ F	1.00	65	[Taft, 1960]
30	σ_p^C	¹³ C nmr chemical shift, XC ₆ H ₅	1.00	68	[Maciel and Natterstad, 1965]

(Continued)

Table E10 (Continued)

ID	Symbol	Reaction	ρ	D%	References
31	ι (iota)	$\iota = 1.755 - \delta^C/54.9$	—	—	[Inamoto and Masuda, 1977]; [Inamoto, Masuda <i>et al.</i> , 1978]
32	σ_x	^{13}C nmr shift, $\text{XC}_6\text{H}_4\text{Y}$ Theoretical calculations	—	—	[Marriott and Topsom, 1982]; [Marriott, Reynolds <i>et al.</i> , 1984]
33	I	K; acids	1.00	5	[Branch and Calvin, 1941]
34	σ_{R}^0	$\sigma_{\text{R}}^0 = 0.0079 A^{1/2} - 0.027$	1.00	96	[Brownlee, Katritzky <i>et al.</i> , 1965]; [Brownlee, Katritzky <i>et al.</i> , 1966]
35	σ_{R}^0	Infrared spectroscopy $\sigma_{\text{R}}^0 = (\delta_{\text{p}}^{\text{F}} - \delta_{\text{m}}^{\text{F}})/2.97$ ^{19}F nmr chemical shift, $\text{XC}_6\text{H}_4\text{F}$	—	—	[Taft, Ehrenson <i>et al.</i> , 1959]
36	R	K; 4-X-pyridinium ions	1.00	—	[Taft and Grob, 1974]
37	F	$f(\sigma_{\text{m}}, \sigma_{\text{p}})$	1.00	22	[Dewar and Grisdale, 1962b]
38	M	$f(\sigma_{\text{m}}, \sigma_{\text{p}})$	1.00	93	[Dewar and Grisdale, 1962b]
39	F'	$f(\sigma_{\text{m}}, \sigma_{\text{p}})$	1.00	27	[Dewar and Grisdale, 1962b]
40	M'	$f(\sigma_{\text{m}}, \sigma_{\text{p}})$	1.00	93	[Dewar and Grisdale, 1962b]
41	σ_{p}^+	$f(\sigma_{\text{m}}, \sigma_{\text{p}})$	1.00	70	[Swain and Lupton Jr., 1968]
42	$\bar{F}$	$f(\sigma_{\text{m}}, \sigma_{\text{p}})$	1.00	0	[Swain and Lupton Jr., 1968]
43	$\mathcal{R}$	$f(\sigma_{\text{m}}, \sigma_{\text{p}})$	1.00	100	[Swain and Lupton Jr., 1968]
44	$\Delta\sigma$	$\sigma_{\text{p}} - \sigma_{\text{m}}$	1.00	92	[McDaniel and Brown, 1958]; [Lewis and Johnson, 1959]
45	C_{T}	X-cianoethylenes	—	—	[Hetnarski and O'Brien, 1975]
46	σ^{Φ}	Dialkylphosphinic acids	—	—	[Mastryukova and Kabachnik, 1971]
47	σ_{a}	Heterocyclic rings	—	—	[Otsuji, Kubo <i>et al.</i> , 1960]

The symbols K and k in the reaction column represent equilibrium $(1/\rho) \log (K_{\text{X}}/K_{\text{H}})$ values and kinetic $(1/\rho) \log (k_{\text{X}}/k_{\text{H}})$ values, respectively; subscripts X and H refer to the X- and H-substituted compounds, respectively. For all the series the unsubstituted compounds have $\sigma = 0$ except for series 15–17, 27, 29, and 30, whose values are: (15) $\sigma = 7.760$; (16) $\sigma = 7.760$; (17) $\sigma = 34.622$; (27) $\sigma = 0.490$; (29) $\sigma = -0.100$; (30) $\sigma = -0.115$. The values of the sensitivity to resonance effect (D%) are taken from Swain–Lupton [Swain and Lupton Jr.,

1968]. σ_{j} electronic constants of Dewar–Grisdale (37–40) represent σ values determined from j - $\text{XC}_{10}\text{H}_8\text{COOH}$ substituted compounds (1-naphtoic acids) with $j = 3, 4, 5, 6, 7$; the corresponding percentages of resonance (D%) are 29, 57, 38, 43, and 48. σ^{Q} values (27) were estimated from the ^{35}Cl quadrupole resonance frequency f in *ortho*-substituted chlorobenzenes; it is related to the Taft σ^+ polar constant. σ_{R}^0 values (34) were estimated from the integrated intensity A of the ν_{16} band in IR spectra.

• overall electronic constants σ_{m} and σ_{p}

These are the original Hammett substituent constants [Hammett, 1937, 1970] measuring the overall electronic effect of the *meta*- and *para*-substituents of benzene derivatives having the functional group in the side chain. They were originally calculated from the variation of the acid

dissociation constant K_a of substituted benzoic acids (*m*-, *p*- $\text{XC}_6\text{H}_4\text{COOH}$) in water at 25 °C, with respect to the unsubstituted compound (i.e., benzoic acid):

$$\sigma_{m,p} = \frac{1}{\rho} \cdot \log\left(\frac{K^X}{K^0}\right) = \frac{1}{\rho} \cdot (pK_a^0 - pK_a^X)$$

where the reaction constant ρ is arbitrarily assumed equal to 1. K^X and K^0 denote the acid dissociation constants of the X-substituted and parent compound, respectively. The subscripts *m* and *p* are used to identify *meta*- and *para*-substituents, respectively. Other values of $\sigma_{m,p}$ were also obtained from the hydrolysis of benzoic esters and other reaction series, based on both equilibrium and rate constants (K and k , respectively).

σ values measure the substituent total electronic effect with respect to hydrogen and are, in principle, independent of the nature of the reaction; however, for large and/or charged substituents the σ values are estimated with less precision.

• unbiased constants σ^0

The unbiased constants σ^0 were defined by Taft to avoid overestimating the resonance effect caused by direct resonance interactions between substituent and reaction site [Taft, 1960], that is, they are a measure only of the interaction between the substituent and the molecular skeleton as felt at the reaction site.

These σ^0 constants were evaluated by the dissociation of *meta*-substituted phenylacetic acids and esters and were defined for a selected group of *m*- XC_6H_4 - substituents which exhibit a precise linear free energy relationship. Their values are independent of whether the process is rate or equilibrium, the solvent and reaction conditions.

They represent inductive constants for the *meta*-substituted phenyl groups (*m*- XC_6H_4 -) relative to the unsubstituted phenyl group (C_6H_5 -), because substituent X in the *meta* position is not directly conjugated with the side chain reaction center Y, thus specific $-\text{C}_6\text{H}_4\text{Y}$ resonance effects do not contribute to the overall electronic effect. However, σ^0 values contain contributions from resonance interactions within *m*- XC_6H_4 - groups. The same does not hold for the corresponding *para*-substituted phenyl substituents (*p*- XC_6H_4 -). Therefore, by using the linear free energy relationships for selected *meta*-substituted groups to determine the ρ reaction constant, the effective $\bar{\sigma}$ values were obtained for all *p*- XC_6H_4 -:

$$\bar{\sigma} = 1/\rho \cdot \log(k/k_0)$$

The difference $\sigma - \sigma^0$ can be considered as a measure of the resonance effect between the aromatic group and the reaction center Y for the dissociation of *meta*-substituted benzoic acids in water. Moreover, deviations from the relationship $\log(k/k_0) = \sigma^0\rho$ provide a useful measure of the specific polar and resonance effects dependent upon both solvent conditions and reaction type.

• inductive electronic constants

These are electronic substituent constants representing the polar effect exerted by the substituent on the active site of a molecule. Different reaction series and data derived from various statistical procedures led to several proposals for inductive constants; only the most popular ones are considered in the following discussion.

The **Roberts–Moreland inductive constant** σ' is based on the dissociation constants K of 4-substituted bicyclo[2,2,2]octane-1-carboxylic acids in 50% ethanol by volume at 25 °C

defined as

$$\sigma' = \frac{1}{1.464} \cdot (\log K_X - \log K_H)$$

where $\rho = 1.464$ is assumed in the Hammett equation [Roberts and Moreland, 1953].

As the X substituent in bicyclo[2,2,2]octane-1-carboxylic acids is bonded to a sp^3 carbon atom, it exerts only an inductive effect. Moreover, the chosen reference compound is free from conformational effects and no steric effect is observed, as the substituent and the active site are not in close proximity to each other.

The **Holtz–Stock inductive constant** σ' was calculated by the dissociation constants K of 4-substituted bicyclo[2,2,2]octane-1-carboxylic acids in 50% ethanol by weight at 25 °C using $\rho = 1.65$ in the Hammett equation [Holtz and Stock, 1964].

Taft σ^* constant (or σ^* electronic constant, Taft polar constant) was proposed by Taft [Taft, 1956] to measure the inductive effect in the aliphatic series:

$$\sigma^* = \frac{1}{2.48} \cdot \left[\log \left(\frac{k_X}{k_{Me}} \right)_B - \log \left(\frac{k_X}{k_{Me}} \right)_A \right] = \frac{1}{2.48} \cdot \left[\log \left(\frac{k_X}{k_{Me}} \right)_B - E_S \right]$$

where A and B, respectively, stand for acid-catalyzed and base-catalyzed, k_X and k_{Me} denote the rate constants of acid- and base-catalyzed hydrolysis or esterification of substituted and unsubstituted esters, respectively, and E_S is the $\rightarrow$ Taft steric constant.

The factor 2.48 corresponds to the average of the available ρ values of alkaline hydrolysis from the Hammett equation and attempts to place σ^* on the same scale as the σ_m and σ_p electronic constants. Since base-catalyzed hydrolysis involves both inductive and steric effects and acid-catalyzed hydrolysis involves only the steric effect, removing the steric effect leaves only the inductive effect, assuming that in both reactions the steric effects are the same.

When defining the σ^* values as a measure of inductive effects, the choice in reactivity types is such that specific steric, resonance, and other effects are apparently constant and a linear free energy relationship of the Hammett type holds.

An **Additive Model of Inductive Effect** was proposed to estimate the σ^* inductive constant of substituents on the basis of the fundamental characteristics of the constituent atoms [Cherkasov, Galkin *et al.*, 1998; Cherkasov and Jonsson, 1998; Cherkasov, 2005]:

$$\sigma^* = \sum_{i=1}^n \frac{\sigma_i^A}{r_i^2} = 7.84 \times \sum_{i=1}^n \frac{\Delta\chi_i \cdot R_i^2}{r_i^2}$$

where the sum runs over all the atoms of the substituent, r_i is the distance of the i th atom of the substituent to the reaction center, and R_i is the covalent radius of the atom. σ^A is an empirical atomic parameter reflecting the ability of an atom to attract (or donate) electrons and its values were estimated by multivariate regression analysis on Taft σ^* constants for several substituents. The values were found to have good correlation with the difference in electronegativity $\Delta\chi$ between a given atom and the reaction center, reflecting the driving force of electron density displacement and, with the square of the covalent radius of the atom, reflecting the ability to delocalize the charge.

The **Taft–Lewis inductive constant** σ_I was proposed [Taft and Lewis, 1958] to measure the inductive effect in aliphatic series on a scale for direct comparison with aromatic σ values, derived from σ^* constant as

$$\sigma_I = 0.45 \times \sigma^* = \frac{1}{5.51} \times \left[\log \left(\frac{k_X}{k_H} \right)_B - \log \left(\frac{k_X}{k_H} \right)_A \right]$$

The derived ρ value of 5.51 was later modified by Taft into $\rho = 6.23$ [Taft, 1960]. In this way, σ_I values can also be considered as a measure of the inductive effect of substituents bonded to aromatic carbons.

These σ_I values of Taft and Lewis were used as a basis set by Charton [Charton, 1963, 1964] to obtain a large number of inductive constants. Acid dissociation constants of substituted acetic acids (XCH_2COOH) in water were correlated with σ_I constants of the basis set at temperatures from 5 ° to 50 °C in terms of the equation:

$$\sigma_I = b \cdot pK_a^X + a$$

The regression coefficients a and b were estimated separately for each reaction series, and then additional σ_I values (**Charton inductive constants**) were estimated and a set of recommended values also suggested. In general, steric and resonance effects in the acetic acid system can be considered negligible. Moreover, unlike the Taft inductive constants, those defined by Charton require only one experiment for their determination.

The **Siegel–Kormany inductive constant** σ'' was calculated by the dissociation constants of 4-substituted cyclohexanecarboxylic acids in three different solvents (in water, water/ethanol 50% by weight and by volume) at 25 °C using $\rho = 1$ in the Hammett equation [Siegel and Komarmy, 1960]. The expected relationship between this inductive constant and the inductive constant of Roberts–Moreland or Taft σ^* was confirmed.

Based on the acid dissociation of 4-substituted quinuclidines in water at 25 °C, the **Grob inductive constant** σ_I^q was proposed [Taft and Grob, 1974; Grob and Schlageter, 1976; Grob, 1985]:

$$\sigma_I^q = pK_a^X - pK_a^H$$

assuming $\rho = 1$ in the Hammett equation. The quinuclidine system is free of steric and conformational effects. Moreover, it is much more sensitive to electronic effects than the bicyclooctane system.

^{19}F inductive constant σ_m^F was estimated from ^{19}F -NMR for *meta*-substituents of F-benzene in very dilute CCl_4 solution [Taft, 1960; Taft, Price *et al.*, 1963a] by the equation:

$$\sigma_m^F \equiv \sigma_I = 0.084 - \frac{\delta_m^F}{7.1}$$

where δ_m^F is the ^{19}F chemical shift of *meta*-substituted fluorobenzenes. The NMR chemical shifts measure the inductive perturbation of *meta*-substituents on the charge density of the fluorine neighbor. This relationship is based on the correlation between chemical shift and the Taft–Lewis inductive constant σ_I .

The **Inamoto–Masuda inductive constant** ι (*iota*) [Inamoto and Masuda, 1977] was empirically defined by modifying the electronegativity as defined by Gordy [Gordy, 1946]. The inductive

constant ι is based on the $\rightarrow$ *electronegativity* of the substituent atom directly bonded to the skeletal group and defined as

$$\iota = \frac{Z_{\text{eff}} + 1}{L_{\text{eff}}}$$

where Z_{eff} is the effective nuclear charge in the valence shell of the considered atom and L_{eff} is the effective principal quantum number (Slater rule). If the atom belongs to the second period group, the inductive constant is estimated by the equation:

$$\iota = 0.64 \cdot \chi_X + 0.53$$

where χ_X is the substituent charge obtained from the bond dipole moment [Inamoto, Masuda *et al.*, 1978].

The **electronegativity-based inductive constant** σ_χ was derived from atomic charge densities on the hydrogen atom of XH, XCH₂H, and XCH₂CH₂H derivatives [Marriott, Reynolds *et al.*, 1984] defined as

$$\sigma_\chi = 1 - q_{\text{H}}$$

where q_{H} is the atomic charge on the H atom and is calculated from $\rightarrow$ *computational chemistry*. The electronegativity of a substituent is primarily determined by the electronegativity of the nearest attached atom. Moreover, the electronegativity of the atom is affected by both changes in hybridization and the polarity of other atoms in the group. The values $1 - q_{\text{H}}$ being related to $\rightarrow$ *electronegativity scales*, such values can be considered a measure of the inductive effect of the substituent.

• resonance electronic constants

These are electronic substituent constants representing the resonance effect exerted by the substituent on the active site of a molecule. Several proposals of resonance constants were made on the basis of different reaction series or derived from statistical procedures; only the most popular are considered in the following.

The **Taft resonance constants** were calculated from the overall electronic effect in specific reaction series by subtracting the inductive contribution based on the Taft–Lewis σ_{I} values [Taft and Lewis, 1959; Taft, Ehrenson *et al.*, 1959; Taft, 1960; Ehrenson, Brownlee *et al.*, 1973] as

$$\sigma_{\text{R}}^0 = \sigma^0 - \sigma_{\text{I}} \quad \bar{\sigma}_{\text{R}} = \bar{\sigma} - \sigma_{\text{I}} \quad \sigma_{\text{R}}^+ = \sigma^+ - \sigma_{\text{I}} \quad \sigma_{\text{R}}^- = \sigma^- - \sigma_{\text{I}}$$

In particular, σ_{R}^0 called **resonance polar effect** [Taft, 1956] is defined for any benzene derivative where there is no direct conjugation between substituent and reactive center; it can be considered constant for a particular solvent, therefore expressing resonance interactions between substituent and skeletal group. $\bar{\sigma}_{\text{R}}$ is usually referred to as the **effective resonance constant**; σ_{R}^+ and σ_{R}^- hold for electrophilic and nucleophilic reaction series, respectively.

Moreover, Taft also tried to propose a general σ_{R} -scale as

$$\sigma_{\text{R}} = \sigma - \sigma_{\text{I}}$$

However, the σ_R values for *para*-substituents show great variability with reaction type, and it was therefore inferred that a widely applicable and precise σ_R -scale could not be devised [Taft and Lewis, 1958, 1959].

Resonance constants σ_R^0 were also estimated from ^{19}F -NMR based on ^{19}F chemical shifts in *para*- and *meta*-substituted F-benzene in very dilute CCl_4 solution [Taft, Ehrenson *et al.*, 1959] by the equation:

$$\sigma_R^0 = \frac{1}{2.97} \cdot (\delta_p^{\text{F}} - \delta_m^{\text{F}})$$

where δ^{F} is the ^{19}F chemical shift. The quantity $\delta_p^{\text{F}} - \delta_m^{\text{F}}$ can be considered a measure of the perturbation in electron density detected by the fluorine atom caused by the resonance interaction between the *para*-substituent and fluorobenzene system; δ_m^{F} is mainly related to the perturbation derived from bond polarizations, and the inductive contributions to δ_m^{F} and δ_p^{F} values are assumed to be equal.

Analogously, a significant correlation was found between δ_p^{C} (and $\delta_p^{\text{C}} - \delta_m^{\text{C}}$) and σ_R^0 or σ_R , δ^{C} being the chemical shift of the ^{13}C atom of the aromatic ring in the *meta*- or *para*-position relative to the substituent [Maciel and Natterstad, 1965].

Dual electronic constants σ^+ and σ^- were proposed to measure the “exaltation” of the resonance effect that appears when the substituent and active site bonded to a skeletal group give origin to a direct conjugation between them.

The **electrophilic substituent constant** σ^+ measures the electronic effects for electron-releasing substituents (e.g., $-\text{OMe}$, $-\text{Me}$, $-\text{OH}$, $-\text{NH}_2$) and for a strong electron-acceptor active site, while the **nucleophilic substituent constant** σ^- measures the electronic effects for electron-acceptor substituents (e.g., $-\text{NO}_2$, $-\text{CN}$, $-\text{CO}_2\text{H}$) and strong electron-releasing active site.

Standard σ^+ values were estimated by Brown – Okamoto [Brown and Okamoto, 1958] from the rate constants k of the solvolysis of substituted *t*-cumyl chlorides ($\text{XC}_6\text{H}_4\text{C}(\text{CH}_3)_2\text{Cl}$) in 90% acetone–water at 25°C , as

$$\sigma^+ = \frac{1}{-4.54} \cdot \log \frac{k}{k_0}$$

where the reaction constant $\rho = -4.54$ was estimated from a set of *meta*-substituents.

The σ^+ values correlate with ionization potentials obtained from substituted benzyl radicals [Harrison, Kebarle *et al.*, 1961].

Standard σ_p^- values were obtained from acid dissociation constants K_a of *para*-substituted phenols ($p\text{-XC}_6\text{H}_4\text{OH}$) in water or water/ethanol 50% at 25°C [Cohen and Jones, 1963] and successively from those of *para*-substituted anilines.

When strong resonance interactions are less relevant, σ^+ and σ^- constants are equal to the normal σ values obtained from substituted benzoic acids.

The **Yukawa–Tsuno equation** (also referred to as **Linear Aromatic Substituent Reactivity relationship**, *LASR*) modifies the Hammett equation, taking into account the exaltation of the resonance effects of electron-releasing and electron-attracting substituents on the reaction center [Yukawa, Tsuno *et al.*, 1972a, 1972b].

For electron-releasing substituents, the Yukawa–Tsuno equation is

$$\log\left(\frac{k_X}{k_H}\right) = \rho \cdot [\sigma^0 + r \cdot (\sigma^+ - \sigma^0)] = \rho \cdot [\sigma^0 + r \cdot \Delta\sigma^+]$$

where k_X and k_H are the respective rate constants of the X-substituted and unsubstituted compounds, ρ is the Hammett reaction constant, and r is the contribution of the enhanced resonance effect of the substituent. σ^0 is the “normal” (i.e., unbiased) substituent constant derived by Yukawa *et al.* from the rate constants of alkaline hydrolysis of ethyl phenyl-acetates ($\text{HO}^- + \text{XC}_6\text{H}_4\text{CH}_2\text{COOEt}$) in water at 25 °C.

$\Delta\sigma^+$ was proposed as the substituent constant measuring the exaltation of the resonance effect of a *para*-substituent on an electrophilic reaction, while the first term $\rho\sigma^0$ accounts primarily for the electronic effect of *meta*- and *para*-substituents whose σ^0 and σ^+ are equivalent.

Analogously, for electron-attracting substituents, the Yukawa–Tsuno equation is

$$\log\left(\frac{k_R}{k_0}\right) = \rho \cdot [\sigma^0 + r \cdot (\sigma^- - \sigma^0)] = \rho \cdot [\sigma^0 + r \cdot \Delta\sigma^-]$$

Both these equations were originally proposed using σ values instead of σ^0 values [Yukawa and Tsuno, 1959]. If $r = 0$, the Yukawa–Tsuno equations reduce to the classical Hammett equation, while, if $r = 1$, it corresponds to the correlation with only σ^+ or σ^- constants.

A resonance constant R was calculated from two distinct reaction series as

$$R = \log\left(\frac{K}{K_0}\right)_I - \log\left(\frac{K}{K_0}\right)_{II}$$

where the subscript I represents the series of dissociation constants of 4-substituted pyridinium ions in water at 25 °C and II the series of dissociation constants of 4-substituted quinoclidinium ions in water at 25 °C [Taft and Grob, 1974].

This resonance constant is justified by the following correlation equations found separately for the two series:

$$\log\left(\frac{K}{K_0}\right)_I = 5.15 \cdot \sigma_I + 2.69 \cdot \sigma_R^+$$

$$\log\left(\frac{K}{K_0}\right)_{II} = 5.15 \cdot \sigma_I$$

Since the inductive effects are essentially the same in the two series, it follows that the difference in $\log(K_X/K_0)$ values for corresponding substituents in I and II reaction series gives a measure of the resonance effect for the dissociation of 4-substituted pyridinium ions.

• field/resonance effect separation

To give a uniform view of the different σ -scales, considering both field-inductive and resonance effect, a number of proposed approaches were aimed at separating the two main contributions within a unique theoretical framework.

According to the **Swain–Lupton approach** (SL), the σ electronic constant is defined as a linear combination of the two basic electronic contributions based on the equation [Swain and

Lupton Jr., 1968; Swain, Unger *et al.*, 1983; Reynolds and Topsom, 1984; Swain, 1984]:

$$\sigma^{\text{SL}} = f \cdot F + r \cdot \mathcal{R}$$

where F is the **field-inductive constant** (or **Swain–Lupton field constant**) and $\mathcal{R}$ is the **resonance constant** (or **Swain–Lupton resonance constant**), f and r being weighting factors that depend on the system used to define the particular σ -scale (analogous to the coefficients ℓ and d defined above).

Swain and Lupton defined the field constant F assuming that the polar effect was a component in both Hammett electronic constants σ_{m} and σ_{p} :

$$F = b_0 + b_1 \cdot \sigma_{\text{m}} + b_2 \cdot \sigma_{\text{p}}$$

where the coefficients were evaluated by least square regression using $\text{p}K_{\text{a}}$ values of 4-substituted bicyclo[2,2,2]octane-1-carboxylic acids of Holtz–Stock. While Swain–Lupton assumed $\rho = 1$ so as to put the F values on the same scale as Hammett constants, Hansch used σ' as originally calculated by $\rho = 1.65$, thus obtaining the following equation [Hansch, Leo *et al.*, 1973]:

$$F \equiv \sigma' = -0.009 + 1.369 \cdot \sigma_{\text{m}} - 0.373 \cdot \sigma_{\text{p}}$$

The resonance constant $\mathcal{R}$ was estimated by

$$\mathcal{R} = \sigma_{\text{p}} - 0.921 \cdot F$$

assuming $r = 1$ and the coefficient f was evaluated assuming that $\mathcal{R} = 0$ for $\text{N}^+(\text{CH}_3)_3$. The main assumption is that the substituents in the *para*-position give the primary resonance effect.

Based on the same previous assumptions, but using a different ρ value ($\rho = 1.56$) and an extended data set of σ_{I} values, the following equation was used to define the field constant F and the corresponding resonance constant $\mathcal{R}$:

$$F \equiv \sigma_{\text{I}} = 0.033 + 1.297 \cdot \sigma_{\text{m}} - 0.385 \cdot \sigma_{\text{p}}$$

Several sets of resonance constants were defined according to the different types of σ_{p} values. Resonance substituent constants $\mathcal{R}^+$, $\mathcal{R}^-$, and $\mathcal{R}^0$ are derived from the corresponding σ values as

$$\mathcal{R}^+ = \sigma_{\text{p}}^+ - f \cdot F \quad \mathcal{R}^- = \sigma_{\text{p}}^- - f \cdot F \quad \mathcal{R}^0 = \sigma_{\text{p}}^0 - f \cdot F$$

using the appropriate field constant F values calculated by Swain–Lupton (the coefficient f is usually taken as equal to 1).

Also in the Taft–Lewis approach, the resonance R and polar effects I can be viewed as additive contributions to the overall electronic effect, defined as

$$\log\left(\frac{k}{k_0}\right) = I + R = \rho_{\text{I}} \cdot \sigma_{\text{I}} + \rho_{\text{R}} \cdot \sigma_{\text{R}}$$

where the inductive effect I is defined as in the Hammett equation.

This equation is applied separately to the effects of *meta*- and *para*-substituents, both depending on σ and ρ positions.

However, assuming

$$\rho_{\text{I}} = \rho_{\text{I}}^{\text{m}} = \rho_{\text{I}}^{\text{p}} = \rho_{\text{R}}^{\text{m}} = \rho_{\text{R}}^{\text{p}} = \rho_{\text{R}}$$

for *meta*- and *para*-substituents of benzene derivatives, the two following equations were proposed:

$$\log\left(\frac{K_p}{K_0}\right) = \rho_I \cdot (\sigma_I + \sigma_R)$$

$$\log\left(\frac{K_m}{K_0}\right) = \rho_I \cdot (\sigma_I + \alpha \cdot \sigma_R)$$

Moreover, the inductive constants are assumed equal in both cases while the resonance constant of the substituent in *meta*-position is considered a fraction of its resonance effect in the *para*-position. The coefficient α was originally proposed as equal to 1/3 for the dissociation of benzoic acids; other values were proposed to account for enhanced resonance effects.

By combining the two expressions, a general equation to measure the inductive effect was proposed as

$$\rho_I \cdot \sigma_I = \left(\frac{1}{1-\alpha}\right) \cdot \left[\log \frac{K_m}{K_0} - \alpha \cdot \log \frac{K_p}{K_0}\right]$$

According to the original **Dewar–Grisdale approach** (DG), also called **FM method** (Field and Mesomeric method) [Dewar and Grisdale, 1962a, 1962b], delocalized and localized long-range field effects are represented by the quantities M and F , respectively. Electronic substituent constants σ_{ij} of a given substituent X , bonded at position i of the skeletal group, acting on a reaction center bonded at position j , are defined by the equation:

$$\sigma_{ij}^{\text{DG}} = \frac{F}{r_{ij}} + M \cdot q_{ij}$$

where q_{ij} and r_{ij} represent the charge in position j and the distance (in units of benzene bond length) between i and j atoms. F and M values are calculated from σ_m and σ_p constants by substituting appropriate values in the equation. Once F and M values are known for each substituent, σ_{ij}^{DG} constants can be calculated for any structure at the i - j positions (e.g., *ortho*, *meta*, *para* positions in benzene derivatives, or other positions in naphthalene derivatives), using appropriate values of r_{ij} and q_{ij} . The term q_{ij} is taken as a measure of the transmission of mesomeric effects between position i and j of a conjugated system. In this approach, both substituent and reaction site are approximated by single point charges located at points i and j (Figure E3).

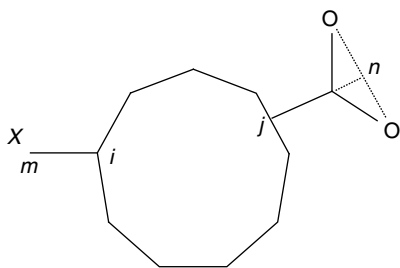


Figure E3 Geometrical scheme for the calculation of the Dewar–Grisdale electronic constants.

Substituting the charge term q_{ij} by the atom–atom polarizability term α_{ij} results in an analogous set of F' and M' values.

The **Dewar–Golden–Harris approach** (DGH), also called **FMMF method** (i.e., Field, Mesomeric, and Mesomeric–Field method) is a modification of the Dewar–Grisdale approach, where the substituent X is approximated by a finite dipole (represented by two point charges along the i –X bond) and the reaction site at position j as a single point charge [Dewar, Golden *et al.*, 1971]. This approach is based on the following equation:

$$\sigma_{ij}^{\text{DGH}} = \frac{F}{r_{ij}} + M \cdot q_{ij} + M_F \cdot \sum_{k \neq j} \frac{q_{ik}}{r_{kj}}$$

where F and M , respectively, represent localized and delocalized effects of the substituent X, and q_{ik} is the charge of the carbon atom at different positions k in the skeletal moiety, and r is the distance (in units of benzene bond length) between atom j and any other atom in the skeletal group. The third substituent parameter M_F is the mesomeric–field constant, describing the ability of the substituent to polarize adjacent π systems; if there is no direct resonance interaction between the substituent and the reaction center, M_F should be proportional to M .

In the event of there being substituent effects on the dissociation of the carboxylic acids, the previous equation becomes

$$\sigma_{ij}^{\text{DGH}} = F \cdot R_{ij} + M \cdot q_{ij} + M_F \cdot \sum_{k \neq j} \frac{q_{ik}}{r_{kn}}$$

where the quantity R_{ij} is defined as

$$R_{ij} = \frac{1}{r_{in}} - \frac{0.9}{r_{mn}}$$

where m is a point charge (with charge $q = -0.9$) at distance 1.40 Å from the point charge i (with charge $q = 1$) along the i –X bond; n is the point charge in the middle of the axis joining the two oxygens of the carbonyl group of the reaction site.

Further modifications of two previous approaches were proposed by Forsyth [Forsyth, 1973], calculating $\sigma_{ij}^{\pm}$ constants.

Other substituent electronic constants were defined for specific different reference compounds (e.g., heterocycles) and reactions.

Aryl electronic constants σ_a are electronic constants defined for substituents on an aromatic ring different from benzene, such as pyridine [Otsuji, Kubo *et al.*, 1960]. The concept may be generalized to various isocyclic and heterocyclic rings such as thiophene, furan, and so on, and various types of aryl electronic constants σ_a^0 , σ_a^+ , σ_a^- may be also defined. A particular set of this kind is given by **σ_r electronic constants** obtained by the protonation reaction of hydrocarbons, assuming as the reference compound α -naphthyl ($\sigma_r = 0$).

Phosphorus electronic constants σ^ϕ are electronic substituent constants derived from the dissociation constants of dialkylphosphinic acids for substituent groups directly bonded to a phosphorus atom [Mastryukova and Kabachnik, 1971]. Assuming the alkyl groups exert only an inductive effect, these electronic constants can be distinguished as σ_I^ϕ inductive constant and σ_R^ϕ resonance constant [Charton and Charton, 1978].

Radical electronic constants are substituent constants derived from free-radical reactions. The most popular are the **E_R radical parameter** defined on the basis of the radical abstraction of α -hydrogens of substituted cumenes [Yamamoto and Otsu, 1967] and the **σ_a^* radical substituent constants** defined by the benzylic α -hydrogen hyperfine coupling constants [Wayner and Arnold, 1984].

The **charge transfer constant** C_T (or **group charge transfer**) is an electronic substituent constant defined from the dissociation constant of a complex between tetracyanoethylene and a X-substituted parent compound [Hetnarski and O'Brien, 1975]:

$$C_T = \log K_X - \log K_H$$

This definition is analogous to that used for σ electronic substituent constants and accounts for the formation ability of a charge-transfer complex (CTC), such as the π -complex formation ability of aromatic systems.

The **electron donor–acceptor substituent constant** κ is an electronic constant proposed to measure the ability of a group to modify the stability of an electron donor–acceptor complex that is often referred to as charge transfer complex [Foster, Hyde *et al.*, 1978; Livingstone, Hyde *et al.*, 1979]. It is defined as

$$\kappa = \log K_X^{APP} - \log K_H^{APP}$$

where K^{APP} is the apparent equilibrium constant for the formation of a complex between the electron-acceptor 1,3,5-trinitrobenzene and the X-substituted benzene in CCl_4 solution at 33.4 °C. Such equilibrium constants were determined by the NMR technique. Benzene was chosen as the reference compound.

📖 Additional references are provided in the thematic bibliography (see Introduction).

- **electronic-topological descriptors** → charge descriptors
- **Electronic-Topological Matrix of Conjunction** → Electronic-Topological method

■ **Electronic-Topological method** ($\equiv$ *ET method*)

This is a QSAR approach based on a matrix representation of chemical compounds involving geometrical and electronic features. This approach is mainly aimed at identifying the → *pharmacophore* for a series of compounds having the same biological activity; the pharmacophore here means a set of common structural and electronic features in active compounds [Dimoglo, 1985; Dimoglo, Bersuker *et al.*, 1988; Bersuker and Dimoglo, 1991; Bersuker, Dimoglo *et al.*, 1991].

The **Electron-Conformational method** (or **EC method**) is an extension of the Electronic-Topological method, which treats more explicitly the role of different conformations of molecules in determining activity. Moreover, this method is aimed at quantitatively predicting the biological activity on the basis of the presence of the pharmacophore, also accounting for the influence of pharmacophore flexibility and the concept of Anti-Pharmacophore Shielding (APS) [Bersuker, Bahçeci *et al.*, 1999b, 1999a, 2000b, 2000a; Bersuker, 2003].

Both methods are based on the same matrix representation of molecules, which is calculated from conformational geometries and → *quantum-chemical descriptors* of atoms and bonds. This

matrix is called **Electronic-Topological Matrix of Conjunction** (ETMC) in the framework of the ET method and **Electron-Conformational Matrix of Congruity** (ECMC) in the framework of the EC method. The matrix is symmetric of dimension ($A \times A$), A being the number of atoms, and is constructed in the following way:

- the diagonal elements are electronic atomic parameters (often, atomic charges, HOMO and LUMO energies, values of the atomic $\rightarrow$ *Interaction Index*);
- the off-diagonal elements corresponding to bonded atoms are bond properties such as $\rightarrow$ *bond order*, bond energy, $\rightarrow$ *Wiberg index*, or polarizability;
- the off-diagonal elements corresponding to pairs of nonbonded atoms are their interatomic $\rightarrow$ *geometric distances* r_{ij} .

For a given conformation of a molecule the interatomic distances are fixed while the electronic parameters relative to atoms and bonds can be combined in different ways, each giving a different matrix. If all these matrices are taken together, a three-dimensional matrix is obtained with $A \times A \times K$ elements where A is the number of atoms and K is the number of all the considered combinations of electronic parameters (Figure E4).

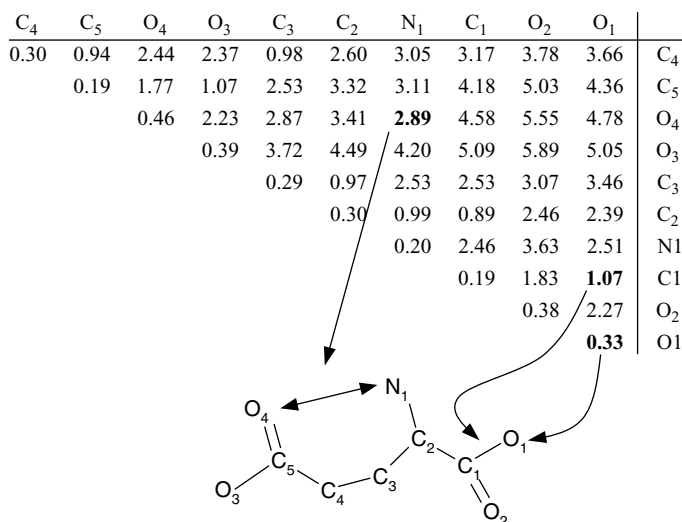


Figure E4 The ECMC for glutamic acid. Interaction Index = 0.33; bond order = 1.07; interatomic distance = 2.89 Å.

The **ET method** allows to select molecular fragments that represent valuable information to design new active compounds [Güzel, 1996; Güzel, Saripinar *et al.*, 1997; Terletsкая, Shvets *et al.*, 1999; Shvets, Terletsкая *et al.*, 1999; Dimoglo, Shvets *et al.*, 2001; Altun, Kumru *et al.*, 2001].

To this end a four-step procedure is adopted.

- (1) The Electronic-Topological Matrix of Conjunction is calculated for each molecule in the data set. If a molecule is found in a few stable conformations, all of them are treated as a

separate molecule and enter the analysis. After processing the ETMCs, only the conformation containing the pharmacophore is considered as the active one.

- (2) Then, one of the most active compounds is chosen as the reference molecule and its ETMC (the template) is compared with all other ETMCs. By this comparison those matrix elements that are present in all active compounds but are absent in the inactive ones (i.e., active features or pharmacophore) are derived and represented by the **Electronic-Topological Submatrix of Conjunction** (ETSC).
- (3) To derive this submatrix, which collects structural and electronic features responsible for activity, the flexibility of molecules is taken into account by choosing some tolerance limits for variation of diagonal (Δ_1) and off-diagonal (Δ_2 for bonded atoms and Δ_3 for nonbonded atoms) elements. Then, to decide which features are responsible for activity, two probabilistic functions are used:

$$(1) \ p_A(F_i) = \frac{n_1 + 1}{n_1 + n_3 + 2} \quad \text{and} \quad (2) \ p_A(F_i) = \frac{n_1 \cdot n_4 - n_2 \cdot n_3}{\sqrt{(n_1 + n_2) \cdot (n_1 + n_3) \cdot (n_2 + n_4) \cdot (n_3 + n_4)}}$$

where n_1 and n_2 are the numbers of molecules possessing and not possessing the feature F_i in the class of active compounds, respectively; n_3 and n_4 are the numbers of molecules possessing and not possessing the feature F_i in the class of nonactive compounds, respectively. The first function is similar to the $\rightarrow$ *Tversky similarity measure* with $\alpha = 1$ and $\beta = 0$ and can be interpreted as the fraction of active molecules possessing the activity feature F_i . The second function is the $\rightarrow$ *Pearson similarity coefficient*, which is related to both active and nonactive compounds.

- (4) To verify the stability of the selected features the procedure is generally repeated using some different reference molecule.

The same procedure can be also used to select inactivity features by choosing one of the most inactive compounds as the reference molecule.

In the **EC method**, to allow the identification of a proper pharmacophore, first data of conformational analysis and electronic structure need to be calculated for every molecule in the training set and on the basis of these data the ECMC matrices are calculated for all the possible conformations of all the compounds.

By comparing the ECMC matrices of all the active compounds with those of the inactive ones, the **Electron-Conformational Submatrix of Activity** (ECSA) is derived, which contains those matrix elements that, within the chosen tolerance, are the same for all the active compounds and are absent in the same combination from the inactive ones. This EC Submatrix of Activity represents the pharmacophore and is equivalent to the Electronic-Topological Submatrix of Conjunction used in the ET method. The information contained in this matrix allows the designing of new active compounds as well as the screening of several compounds with respect to their activity. However, the presence of the pharmacophore is a very important necessary condition of activity, but it may not be sufficient for practical prediction of activity, because there may be other atoms and/or atomic groups that are positioned outside the pharmacophore and influence the molecule activity. These groups are divided into Anti-Pharmacophore Shielding (APS) groups defined as groups of atoms and charges outside the pharmacophore, which hinder

the proper ligand–receptor docking, diminishing the activity partially or completely, and other Auxiliary Groups (AG) that influence the activity in other way (e.g., groups responsible for hydrophobicity). The influence of the APS/AG groups on the activity is accounted for by introducing in the final model for activity some specific structural and electronic parameters, which need to be optimized by statistical analysis. Therefore, the model for activity (A) prediction, based on the Boltzmann distribution of each conformation of drug molecules and a function S for the energy difference in the ligand–receptor-binding interaction, is the following:

$$A_i = A_0 \cdot \frac{\sum_{m=1}^{M_i} e^{-E_{im}/kT} \cdot e^{-S_{im}} \cdot \delta_{im}(Pha)}{\sum_{m=1}^{M_i} e^{-E_{im}/kT}}$$

where A_i is the activity of the i th molecule, A_0 is a constant, M_i is the number of conformations of the i th molecule, kT the Boltzmann term, E_{im} the energy of the m th conformation of the i th molecule, and δ is the Dirac delta function equal to 1 when the pharmacophore (Pha) is present and zero otherwise. The function S_{im} accounts for the influence of APS/AG groups on the molecule activity and is defined as

$$S_{im} \equiv \frac{E'_{im}}{kT} = \sum_{j=1}^N b_j \cdot x_{im,j}$$

where E' is the contribution of APS/AG groups and pharmacophore flexibility to the ligand–receptor interaction energy, $x_{im,j}$ is the parameter that describes the j th APS/AG group in the m th conformation of the i th molecule, N is the number of chosen groups, and b are the regression coefficients to be estimated. The choice of $x_{im,j}$ parameters depends on the given data set. The parameter that describes the pharmacophore flexibility may be taken as the atomic $\rightarrow$ Interaction Index and interatomic distance between the pharmacophore atoms. The parameters for APS groups may be taken as the geometrical steric factors determined by their outstanding position with regard to the pharmacophore main plane and the basic skeleton.

A simplified model for activity is obtained by considering only the lowest energy conformation where the pharmacophore is present (P) and using a reference molecule (R) for which the activity A_R is known to determine the A_0 constant:

$$A_i = A_R \cdot e^{-(E_{iP} - E_{RP})/kT} \cdot e^{-(S_{iP} - S_{RP})}$$

where S_{iP} and S_{RP} are linear combinations of the chosen parameters x for APS/AG groups and pharmacophore flexibility, as described above, for the lowest energy conformation with Pha of the i th molecule and the reference molecule, respectively.

The **Electron-Conformational Approach** (ECA) is based on the same procedure as the ET method [Chumakov, Terletskaia *et al.*, 2000]. The main difference is the representation of molecules, which, in the framework of ECA, is defined in terms of a Set of Electronic and Conformational Parameters (SECPs) in place of the Electronic-Topological Matrix of Conjunction. Then, after choosing a compound as the reference, its SECPs are compared with the SECPs

of all other compounds to select the set of parameters that are common to all active compounds (i.e., activity features). The electronic $\rightarrow$ *quantum-chemical descriptors* are the dipole moment of the molecule, HOMO and LUMO energies, the energies of HOMO minus one and minus two orbitals (HOMO-1 and HOMO-2), their corresponding gap energies, atomic orbital coefficients and the derived reactivity indices, atomic charges, and nonpolar water-accessible area of the van der Waals molecular surface.

The conformational parameters are distances and torsion angles between pseudoatoms (PAs) and chain's atoms as well as dihedral angles between PAs. Pseudoatoms are used to replace flat fragments in the molecular graph, while the remaining atoms of a molecule are referred to as chains.

- **Electronic-Topological Submatrix of Conjunction** $\rightarrow$ Electronic-Topological method
- **electron-ion interaction potential** $\rightarrow$ average quasivalence number
- **electron isodensity contour surface** $\rightarrow$ molecular surface
- **electron-transition stochastic matrix** $\rightarrow$ MARCH-INSIDE descriptors
- **electrophilic atomic frontier electron density** $\rightarrow$ quantum-chemical descriptors
- **electrophilic charge** $\rightarrow$ quantum-chemical descriptors ($\odot$ electrophilic atomic frontier electron density)
- **electrophilic frontier electron density index** $\rightarrow$ quantum-chemical descriptors ($\odot$ electrophilic atomic frontier electron density)
- **electrophilic indices** $\rightarrow$ reactivity indices
- **electrophilicity index** $\rightarrow$ quantum-chemical descriptors ($\odot$ hardness indices)
- **electrophilic substituent constant** $\rightarrow$ electronic substituent constants ($\odot$ resonance electronic constants)
- **electrophilic superdelocalizability** $\rightarrow$ quantum-chemical descriptors
- **electropositivity of an atom** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)

■ electropy index (ε)

It is an information index proposed to characterize the global electronic structure of molecules calculated from the molecular structure but avoiding quantum chemical approaches [Mekeny, Bonchev *et al.*, 1987]. It is defined as

$$\varepsilon = \log_2(N_{\text{el}}!) - \sum_{g=1}^G \log_2(n_g!)$$

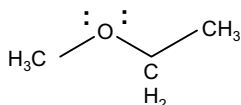
where N_{el} is the total number of electrons in the molecule, that is, the sum of all electrons (inner and valence electrons) of all atoms; n_g is the number of electrons in the g th molecular subspace and G is the number of all possible subspaces.

Molecular subspaces are (a) the core parts of each atom type (e.g., C_{1s} , O_{1s} , N_{1s}); (b) σ and π bond spaces, where σ space is further divided into different independent bond spaces such as C–C, C–H, CH_2 , CH_3 , C–O, etc.; (c) the lone pairs on each atom also constitute an independent subspace.

The electropy index may be viewed as a measure of the degree of freedom for electrons to occupy different subspaces during the process of molecular formation.

Example E4

Calculation of the electropy index for methylethyl ether.



$$\begin{aligned} \varepsilon &= \log_2(34!) - 3 \times \log_2(2!) - 1 \times \log_2(2!) + \\ &- 2 \times \log_2(6!) - 1 \times \log_2(4!) - 2 \times \log_2(2!) + \\ &- 1 \times \log_2(2!) - 1 \times \log_2(4!) = 92.642 \end{aligned}$$

Molecular subspace type	Subspace electrons, n_g	Subspace of each type, n
1s(C)	2	3
1s(O)	2	1
CH ₃	6	2
CH ₂	4	1
C-O	2	2
C-C	2	1
oxygen lone pairs	4	1
no. of electrons, $N_{el} = 34$		
no. of subspaces, $G = 11$		

- **electrostatic balance term** → GIPF approach
- **electrostatic factor** → electric polarization descriptors
- **electrostatic hydrogen-bond acidity** → Theoretical Linear Solvation Energy Relationships
- **electrostatic hydrogen-bond basicity** → Theoretical Linear Solvation Energy Relationships
- **electrostatic interaction fields** → molecular interaction fields
- **electrotopological descriptor** → charge descriptors (⊖ total absolute atomic charge)
- **electrotopological state index** $\equiv$ *E-state index* → electrotopological state indices

■ **electrotopological state indices**

The electrotopological state S_i of the i th atom in the molecule, called **E-state index** (or **electrotopological state index**) gives information related to the electronic and topological state of the atom in the molecule and is defined as [Kier and Hall, 1990a, 1999b; Ivanciuc, 2008]:

$$S_i = I_i + \Delta I_i = I_i + \sum_{j=1}^A \frac{I_i - I_j}{(d_{ij} + 1)^k}$$

where I_i is the **intrinsic state** of the i th atom and ΔI_i is the field effect on the i th atom calculated as perturbation of the intrinsic state of i th atom by all other atoms in the molecule, d_{ij} is the → *topological distance* between the i th and the j th atoms, and A is the number of atoms. The exponent k is a parameter to modify the influence of distant or nearby atoms for particular studies. Usually it is taken as $k = 2$.

The intrinsic state of the i th atom is calculated by

$$I_i = \frac{(2/L_i)^2 \cdot \delta_i^v + 1}{\delta_i}$$

where L_i is the principal quantum number (2 for C, N, O, F atoms, 3 for Si, S, Cl, . . .), δ_i^v is the number of valence electrons (→ *valence vertex degree*) and δ_i is the number of sigma electrons (→ *vertex degree*) of i th atom in the → *H-depleted molecular graph* (Table E11).

Table E11 Kier–Hall atom types.

No.	Atom group	Z	δ^v	δ	VSI	I_{AR}	<i>I</i>	KHE	Symbol
1	–CH ₃	6	1	1	2	0	2.000	0.00	sCH3
2	=CH ₂	6	2	1	3	0	3.000	0.25	dCH2
3	–CH ₂ –	6	2	2	4	0	1.500	0.00	ssCH2
4	≡CH	6	3	1	4	0	4.000	0.25	tCH
5	=CH–	6	3	2	5	0	2.000	0.25	dsCH
6	aCHa	6	3	2	5	1	2.000	0.25	aaCH
7	>CH–	6	3	3	6	0	1.333	0.00	sssCH
8	=C=	6	4	2	6	0	2.500	0.50	ddC
9	≡C–	6	4	2	6	0	2.500	0.50	tsC
10	=C<	6	4	3	7	0	1.667	0.25	dssC
11	aCa–	6	4	3	7	1	1.667	0.25	aasC
12	aaCa	6	4	3	7	1	1.667	0.25	aaaC
13	>C<	6	4	4	8	0	1.250	0.00	ssssC
14	–NH ₃ [+ 1]	7	2	1	3	0	3.000	0.25	sNH3
15	–NH ₂	7	3	1	4	0	4.000	0.50	sNH2
16	–NH ₂ –[+ 1]	7	3	2	5	0	2.000	0.25	ssNH2
17	=NH	7	4	1	5	0	5.000	0.75	dNH
18	–NH–	7	4	2	6	0	2.500	0.50	ssNH
19	aNHa	7	4	2	6	1	2.500	0.50	aaNH
20	≡N	7	5	1	6	0	6.000	1.00	tN
21	>NH–[+ 1]	7	4	3	7	0	1.667	0.25	sssNH
22	=N–	7	5	2	7	0	3.000	0.75	dsN
23	aNa	7	5	2	7	1	3.000	0.75	aaN
24	>N–	7	5	3	8	0	2.000	0.50	sssN
25	–N≪	7	5	3	8	0	2.000	0.50	ddsN (nitro)
26	aaNs	7	5	3	8	1	2.000	0.50	aasN (N-oxide)
27	>N<[+ 1]	7	5	4	9	0	1.500	0.25	ssssN (onium)
28	–OH	8	5	1	6	0	6.000	1.00	sOH
29	=O	8	6	1	7	0	7.000	1.25	dO
30	–O–	8	6	2	8	0	3.500	1.00	ssO
31	aOa	8	6	2	8	1	3.500	1.00	aaO
32	–F	9	7	1	8	0	8.000	1.50	sF
33	–PH ₂	15	3	1	4	0	2.333	0.22	sPH2
34	–PH–	15	4	2	6	0	1.388	0.22	ssPH
35	>P–	15	5	3	8	0	1.073	0.22	sssP
36	–>P=	15	5	4	9	0	0.806	0.11	dsssP
37	–>P<	15	5	5	10	0	0.644	0.00	sssssP
38	–SH	16	5	1	6	0	3.222	0.44	sSH
39	=S	16	6	1	7	0	3.667	0.55	dS
40	–S–	16	6	2	8	0	1.833	0.44	ssS
41	aSa	16	6	2	8	1	1.833	0.44	aaS
42	>S=	16	6	3	9	0	1.221	0.33	dssS (sulfone)
43	=>S=	16	6	4	10	0	0.916	0.22	ddssS (sulfate)
44	–>S<–	16	6	6	12	0	0.611	0.00	ssssssS
45	–Cl	17	7	1	8	0	4.111	0.67	sCl
46	–she	34	5	1	6	0	2.250	0.25	sSeH
47	=Se	34	6	1	7	0	2.500	0.31	dSe

(Continued)

Table E11 (Continued)

No.	Atom group	Z	δ^v	δ	VSI	I_{AR}	I	KHE	Symbol
48	—Se—	34	6	2	8	0	1.250	0.25	ssSe
49	>Se=	34	6	3	9	0	0.833	0.19	dssSe
50	>Se<<	34	6	4	10	0	0.625	0.13	ddssSe
51	—Br	35	7	1	8	0	2.750	0.38	sBr
52	—I	53	7	1	8	0	2.120	0.24	sI

Z, atomic number; δ^v , valence vertex degree; δ , vertex degree; VSI, valence state indicator; I_{AR} , aromatic indicator; I , intrinsic state; KHE, Kier–Hall electronegativity. Data from [Kier and Hall, 1999b].

The intrinsic state of an atom can be simply thought of as the ratio of π and lone-pair electrons over the count of the σ bonds in the molecular graph for the atom considered. Therefore, the intrinsic state reflects the possible partitioning of non- σ electron's influence along the paths starting from the considered atom; the less partitioning of the electron influence, the more available are the valence electrons for intermolecular interactions. The sum of the intrinsic states of all of the atoms is a molecular descriptor called **intrinsic state sum**; moreover, from the intrinsic state sum the $\rightarrow Q$ *polarity index* was derived.

The perturbation $\Delta_{ij} = I_i - I_j$ of the i th intrinsic state by the j th atom can be viewed as an “electronegative gradient” whose sign gives the direction of influence of surrounding atom intrinsic states.

From the definition of intrinsic states and field effects, it can be seen that large positive values of E -states S_i relate to atoms of high electronegativity and/or terminal atoms or atoms that lie on the mantle of the molecule; small or negative E -state values correspond to atoms possessing only σ electrons and/or buried in the interior of the molecule or close to higher electronegative atoms. Therefore, the E -state index is a measure of the electronic accessibility of an atom and can be interpreted as a probability of interaction with another molecule. However, the index cannot be considered a pure electronic descriptor: it is, in fact, a descriptor of atom polarity and steric accessibility.

Note that

$$\sum_{i=1}^A \Delta I_i = 0 \quad \rightarrow \quad \sum_{i=1}^A S_i = \sum_{i=1}^A I_i$$

the sum of the field effects over all atoms in the molecule being equal to zero.

This corresponds to an electronegativity equalization principle and means that the sum of the E -states in the molecule depends only on the number and type of atoms, not on their mutual interactions.

The electrotopological states are $\rightarrow$ *local vertex invariants*. After rescaling, they are also used as atomic weighting factors for the calculation of the $\rightarrow$ *WHIM descriptors*.

Since the E -state values derive from an H-depleted graph, they are calculated for each $\rightarrow$ *hydride group*, that is, they encode electronic and topological information about both heavy atoms and their bonded hydrogens. For molecules with high polar groups, these two contributions can be treated separately by the **hydrogen electrotopological state index** HS_i (or **HE-state index**), which was defined to complement the E -state index to encode electronic and topological

information about the hydrogens. It is defined as [Kier and Hall, 1999b]:

$$\begin{aligned} HS_i &= KHE_i + [KHE_i - KHE(H_i)] + \sum_{j \neq i} \frac{KHE_j - KHE(H_i)}{(d_{ij} + 1)^2} \\ &= KHE_i + [KHE_i + 0.2] + \sum_{j \neq i} \frac{KHE_j + 0.2}{(d_{ij} + 1)^2} \end{aligned}$$

where KHE_i is the $\rightarrow$ Kier–Hall electronegativity of the i th heavy atom in the H-depleted molecular graph chosen as a measure of the intrinsic state of the attached hydrogen atom whose electronegativity $KHE(H_i)$ is taken to be -0.2 ; the perturbation term is given by the sum over all other heavy atoms in the molecule of the difference between their electronegativity and hydrogen electronegativity divided by the square of the topological distance d ; note that the distance is calculated between each j heavy atom and the i th heavy atom to which the hydrogen is bonded.

Therefore, given a H-depleted molecular graph two state values can be calculated for each vertex, the E -state value, which measures the electron density and accessibility of the atom, and the HE -state value, which measures the reaction and interaction ability of the bonded hydrogens, that is, the polarity of X–H bonds. Obviously, the HS values for atoms with no attached hydrogens are always equal to zero.

In this approach to the calculation of hydrogen E -states HS_i , topology is not considered a relevant factor in determining the E -state of a hydrogen, only the relative electronegativity is used to characterize the polarity of bonds with hydrogen in the molecule. However, in other definitions of hydrogen electrotopological states, the topology is also accounted for. In particular, in the first approach to HE -states the state of the hydrogen atom in the X–H bond is mainly determined by the electronegativity of the X atom and, to a lesser extent, by its topology [Kellogg, Kier *et al.*, 1996]. It is defined as

$$HS_i = I(H_i) + \sum_j \frac{\Delta I_{ij}}{(d_{ij} + 1)^2} = \frac{(\delta_i^v - \delta_i)^2}{\delta_i} + \sum_j \frac{\Delta I_{ij}}{(d_{ij} + 1)^2}$$

where $I(H_i)$ is the intrinsic state of the hydrogen bonded to the i th atom, defined in terms of square electronegativity of the i th atom to accentuate the electronic influence on H in the X–H bond. ΔI_{ij} is the perturbation of the hydrogen intrinsic state by the intrinsic state of the j th atom in the molecule and d is the topological distance. Hydrogen intrinsic states seem to be a measure of the H-donor ability of X–H groups.

Another approach [Kier and Hall, 1999b] is to apply the E -state definition for each vertex of the molecular graph where the hydrogens of polar groups ($-\text{OH}$, $-\text{NH}$, $-\text{COOH}$, etc.) are explicitly considered as independent vertices. In this case, the intrinsic state of polar hydrogens will be

$$I_i(\text{H}) = \frac{(2/L_i)^2 \cdot \delta_i^v + 1}{\delta_i} = 5$$

where L , δ^v , and δ always equal one. For each polar group, the difference between the HE -state and the E -state of the bonded atoms H–X reflects the polarity of the bond considered.

Based on the same approach used to define E -state indices, a **bond E -state index** BS_b was also tentatively proposed as [Kier and Hall, 1999b]

$$BS_b = BI_b + \Delta BI_{bt} = \sqrt{(I_i \cdot I_j)_b} + \sum_{t \neq b} \frac{BI_b - BI_t}{(d_{bt} + 1)^2}$$

where b is the considered bond formed by the atoms i and j , t runs over all the remaining bonds different from the bond b , BI is the bond intrinsic state defined by the intrinsic states I of the adjacent vertices, ΔBI the perturbation term, d_{bt} the $\rightarrow$ topological edge distance between bonds b and t . An alternative expression for the bond E -state index was proposed as [Tetko, Tanchuk *et al.*, 2001c]

$$BES_b = BI_b + \Delta BI_{bt} = \frac{(I_i + I_j)_b}{2} + \sum_{t \neq b} \frac{BI_b - BI_t}{(\bar{r}_{bt} + 1)^2}$$

where $\bar{r}_{bt}$ is the average bond length of the bonds b and t .

E -state and HE -state values were also used as atomic properties to calculate $\rightarrow$ molecular interaction fields [Kellogg, Kier *et al.*, 1996]. The **E -state fields** are defined by superimposing a 3D fixed grid over the molecule and calculating at each k th grid point an interaction energy value:

$$E_k = \sum_i S_i \cdot f(r_{ik})$$

where S_i is the electrotopological state value of the i th atom and $f(r_{ik})$ is a function of the distance r between each i th atom of the target molecule and the considered k th grid point. This function is not defined *a priori* but has to be empirically searched for; explored functions were $1/r$, $1/r^2$, $1/r^3$, $1/r^4$, and e^{-r} . Since the $1/r^n$ functions are discontinuous at grid points close to atoms, field default values of zero are set for grid points within the van der Waals envelope of the molecule. In the same way **HE -state fields** are calculated by using hydrogen electrotopological state values HS for each molecule atom instead of S values.

Moreover, **atom-type E -state indices** were proposed as molecular descriptors encoding topological and electronic information related to particular atom types in the molecule [Hall and Kier, 1995; Hall, Kier *et al.*, 1995]. They are calculated by summing the E -state values of all atoms of the same atom-type in the molecule or, alternatively, as average of the E -state values. Each atom-type is first defined by atom identity, based on the atomic number Z , and valence state, itself identified by the **valence state indicator (VSI)** defined as

$$VSI = \delta^v + \delta$$

where δ^v and δ are the $\rightarrow$ valence vertex degree and the $\rightarrow$ vertex degree of the atom; an aromatic indicator variable I_{AR} is also used to discriminate atoms of an aromatic system ($I_{AR} = 1$) from nonaromatic atoms ($I_{AR} = 0$). To distinguish particular atoms, classified as the same atom type according to atom number, valence state indicator and aromatic indicator, the analysis of bonded atoms and the difference between valence δ^v and simple vertex degree δ are used. The symbol of each atom-type E -state index is a composite of three parts (Table E11). The first part is “S” which refers to the sum of the E -states of all atoms of the same type. The second part is a string representing the bond types associated with the atom (“s”, “d”, “t”, “a” for single, double, triple, and aromatic bonds, respectively). The third part is the symbol identifying the chemical element and eventual bonded hydrogens, such as CH_3 , CH_2 , F, and so on.

The atom-type E -state indices combine structural information about the electron accessibility associated with each atom type, an indication of the presence or absence of a given atom type and a count of the number of atoms of a given atom type.

Atom-type E -state counts were proposed as simple counts of the E -state atom types in a molecule [Butina, 2004].

Moreover, **atom-type HE-state indices** were proposed as molecular descriptors calculated by summing hydrogen electrotopological states of all atoms of the same atom-type [Kier and Hall, 1999b]. **Bond-type E-state indices** were analogously defined by summing bond *E*-state index values of all the edges of the same type [Tetko, Tanchuk *et al.*, 2001c].

A $\rightarrow$ *Balaban-like index*, called $\rightarrow$ *E-state topological parameter*, was derived from *E*-state indices.

Two sets of molecular descriptors, called **intrinsic state pseudoconnectivity indices** Ψ , one based on the intrinsic states *I* and the other on the scaled electrotopological state values *E*, were proposed as [Pogliani, 2000b, 2004]:

$$\begin{array}{ll}
 1. \quad {}^S\Psi_I = \sum_{i=1}^A I_i & {}^S\Psi_E = \sum_{i=1}^A E_i \\
 2. \quad {}^0\Psi_I = \sum_{i=1}^A (I_i)^{-1/2} & {}^0\Psi_E = \sum_{i=1}^A (E_i)^{-1/2} \\
 3. \quad {}^1\Psi_I = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (I_i \cdot I_j)^{-1/2} & {}^1\Psi_E = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (E_i \cdot E_j)^{-1/2} \\
 4. \quad {}^T\Psi_I = \prod_{i=1}^A (I_i)^{-1/2} & {}^T\Psi_E = \prod_{i=1}^A (E_i)^{-1/2} \\
 5. \quad {}^0\Psi_{Id} = (-0.5)^A \cdot \prod_{i=1}^A I_i & {}^1\Psi_{Ed} = (-0.5)^A \cdot \prod_{i=1}^A E_i \\
 6. \quad {}^1\Psi_{Id} = (-0.5)^{(A+C-1)} \cdot \prod_b (I_i + I_j)_b & {}^1\Psi_{Ed} = (-0.5)^{(A+C-1)} \cdot \prod_b (E_i + E_j)_b \\
 7. \quad {}^1\Psi_{Is} = \prod_b (I_i + I_j)_b^{-1/2} & {}^1\Psi_{Es} = \prod_b (E_i + E_j)_b^{-1/2}
 \end{array}$$

where *A* is the total number of atoms; in the equations 1, 2, 4, and 5 the summation/product runs over all the atoms, while, in equation 3, it is over all the atoms, but only pairs of bonded atoms give contributions different from zero, a_{ij} being the elements of the $\rightarrow$ *adjacency matrix*. In equation 6, *C* is the $\rightarrow$ *cyclomatic number*; in equations 6 and 7, only contributions from pairs of bonded atoms are considered.

The pseudoconnectivity descriptors Ψ_E are calculated from the electrotopological *E*-state values S_i transformed to avoid negative *S* values. Then, scaled electrotopological values *E* are calculated as $E_i = S_i + 5.5$, where -5.5 is the *S* value for the carbon atom in CF_4 , which is the lowest *S* value a carbon atom can assume. These descriptors were used to model melting point and crystal density of aminoacids and alkanes, enthalpies of metal halides, as well as some biological activities [Pogliani, 2001a, 2002a, 2002b, 2002c, 2003a, 2003b, 2005a, 2005b, 2006b].

Other descriptors based on a modification of intrinsic and *E*-states are $\rightarrow$ *MEDV descriptors*.

📖 Additional references are listed in the thematic bibliography (see Introduction).

- **elongation** → graph
- **elongation** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **elongation/elongation-fixed ratio** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **embedded cluster graph** → biodescriptors (⊙ proteomics maps)
- **embedded correlation** → statistical indices (⊙ correlation measures)
- **embedded graph of partial order** → biodescriptors (⊙ proteomics maps)
- **embedded neighborhood graph** → biodescriptors (⊙ proteomics maps)
- **embedded zigzag curve** → biodescriptors (⊙ proteomics maps)
- **embedding frequencies** → cluster expansion of chemical graphs

■ empirical descriptors

The class of the empirical descriptors is a fuzzy, not well-defined class. In principle, empirical descriptors are those not defined on the basis of a general theory such as, for example, quantum chemistry or graph theory. Rather they are defined by practical rules derived from chemical experience, thus considering specific or local structural factors present in the molecules, often sets of congeneric compounds. As a consequence, in most of the cases, empirical descriptors represent limited subsets of compounds and cannot be extended to classes of compounds different from those for which they were defined. Empirical descriptors have not to be confused with experimentally derived descriptors even if it is well known that several of them are empirically derived.

Empirical descriptors can be user-defined values for discriminating among special molecular fragments or number of occurrences of local specific atom/fragment within a molecule.

An example of useless empirical descriptor is the **Sadhana index** (or *Sadhna index*) [Khadikar, Agrawal *et al.*, 2002; Khadikar, Joshi *et al.*, 2004] defined for polyacene molecules as $Sd = 2h \times (5h + 1)$, where h is the number of benzene units. This index is almost perfectly correlated to the → *PI index*, to the square count of benzene units, and any quadratic combination of the coefficients of h and h^2 , such as $5h^2 + h$ or $20h^2 + 4h$.

Other examples of empirical descriptors are the → *Taillander index* (restricted to substituted benzenes), → *second-grade structural parameters* (restricted to alkenes), → *polar hydrogen factor* (restricted to halogenated hydrocarbons), → *hydrophobic fragmental constants*, → *six position number*, → *Idoux steric constant*, → *hydrophilicity index*, → *adsorbability index*, → *bond flexibility index*, and → *atomic solvation parameter*.

📖 [Carlton, 1998; Chiorboli, Piazza *et al.*, 1993c]

- **endospectral graph** → self-returning walk counts
- **endospectral vertices** → self-returning walk counts
- **end point mean square distance index** → distance matrix
- **end-to-end distance** → size descriptors
- **energy-based descriptors** → quantum-chemical descriptors
- **energy moments** → quantum-chemical descriptors
- **enthalpic fields** → molecular interaction fields
- **enthalpic hydrophobic substituent constant** → lipophilicity descriptors (⊙ Hansch–Fujita hydrophobic substituent constants)

- **enthalpies** → physico-chemical properties
- **entropic fields** → molecular interaction fields
- **entropic hydrophobic substituent constant** → lipophilicity descriptors (⊙ Hansch–Fujita hydrophobic substituent constants)

■ environmental indices

In recent years, a great importance has been given to QSAR approaches for modeling and predicting the environmental behavior, fate, and toxicity of chemicals, that is, **environmental QSAR** [Dearden, 2002; Devillers and Karcher, 1991; Karcher and Devillers, 1990a, 1990b; Karcher and Karabunarliev, 1996; Sabljic, 1990]. Research has mainly focused on some chemical classes of environmental interest, such as Persistent Organic Pollutants (POPs), Volatile Organic Compounds (VOC), Hazardous Air Pollutants (HAP), Persistent-Bioaccumulative-Toxic (PBT) pollutants.

Numerical quantities, experimentally determined or estimated by statistical or computational approaches, which measure the environmental behavior, fate, and toxicity of chemicals, are molecular descriptors usually referred to as environmental indices. The → *octanol–water partition coefficient* (K_{ow} , $\log P$) is the most well-known environmental index used as the measure of lipophilicity of compounds. Together with $\log P$ and the → *soil sorption partition coefficient* (K_{oc}) [Baker, Mihelcic *et al.*, 2001; Uddameri and Kuchanur, 2004], other quantities have to be considered as relevant for environmental studies. Some of these have been defined to describe mobility, biodegradability, bioaccumulation, metabolism, partition, and toxicity of chemicals, thus becoming relevant to human health and environmental safety assessment.

Some definitions of the most popular environmental descriptors are given below.

• half lifetime ($t_{1/2}$)

The half lifetime of a compound, subject to exponential decay, is the time required for the compound to decay to half of its initial value. Although this concept originated from the study of radioactive decay, it applies to many other fields as well, including phenomena that are described by nonexponential decays. It is mathematically defined as

$$t_{1/2} = \ln(2) \cdot \tau$$

where τ is the mean lifetime.

The half lifetime of a chemical is calculated as the length of time it takes for the concentration of that chemical to be reduced by one-half relative to its initial level, assuming first-order decay kinetics. It can be estimated for all major environmental compartments (water, air, soil, sediments, and biota).

The half lifetime in air is related to the atmospheric degradation ability of a chemical, measured by the rate constant of its reactions with free radicals (e.g., $\text{OH}^\bullet$, $\text{NO}_3^\bullet$) and ozone O_3 or of photochemical reactions [Gramatica and Papa, 2007; Gramatica, Pilutti *et al.*, 2003b].

In water, soil, and sediments, besides physical and chemical reactions, the enzymatic biological activity, mainly due to microorganisms (biodegradation) plays a relevant role.

• persistence

The persistence of a chemical is the length of time it remains in a particular environment (water, soil, air, and sediment) in an unchanged form before it is physically transported to another

compartment and/or is chemically or biologically transformed [Gramatica, Consolaro *et al.*, 2001; Gramatica, Papa *et al.*, 2004; Leip and Lammel, 2004; Pavan and Worth, 2008]. The longer a chemical persists, the higher the potential for human or environmental exposure to it.

The experimental determination of the persistence is generally based on the **degradability** of the chemical, that is, a chemical that is degraded in an experimental test system is usually considered not persistent.

The degradation process is often characterized by the extent of degradation and the nature of the degradation process. Then, *primary degradation* refers to the production of organic derivatives that exhibit their own degradation properties; *mineralization* refers to the complete degradation of an organic chemical to stable inorganic species; *abiotic degradation* refers to transformations such as reduction, oxidation, hydrolysis, and photodegradation; *biodegradation* refers to transformation by enzymatic reactions in microorganisms [Pavan and Worth, 2008].

• bioconcentration factor (BCF)

The bioconcentration factor is the concentration of a chemical in a tissue per concentration of the chemical in water (generally adimensional) [Pavan, Netzeva *et al.*, 2008]. This physical property characterizes the uptake of pollutants due to chemical partitioning from environmental phase (e.g., air or water) into an organic phase (e.g., lipids or proteins) through an exchange surface (e.g., gills of fish).

A simple scale for *BCF* values is the following: high potential: $BCF > 1000$; moderate potential: $1000 > BCF > 250$; low potential: $BCF < 250$.

The bioconcentration factor for aquatic organisms is related to the octanol–water partition coefficient (K_{ow}) by the following equation:

$$\log(BCF) = a + b \cdot \log K_{ow}$$

where coefficients *a* and *b* depend upon the type of organism (e.g., fish, crustaceans).

Comparable equations can be developed for bioconcentration from air, by using K_{oa} (octanol–air partition coefficient).

📖 [Bermúdez-Saldaña, Escuder-Gilabert *et al.*, 2005; Cheng, Kontogeorgis *et al.*, 2005; Dearden, 2002; Devillers, Domine *et al.*, 1998; Feng, Han *et al.*, 1996b; Govers, Rupert *et al.*, 1984; Gramatica, 2001; Gramatica and Papa, 2003, 2005; Ivanciuc, 1998b; Khadikar, Singh *et al.*, 2003; Papa, Dearden *et al.*, 2007; Roy, Sanyal *et al.*, 2006; Russom, Breton *et al.*, 2003; ; Sabljčić, 1988; Sabljčić and Protic, 1982a; Vighi, Gramatica *et al.*, 2001; Wei, Zhang *et al.*, 2001; Zhao, Yuan *et al.*, 1997]

• bioaccumulation

Bioaccumulation represents the uptake of a chemical through all routes of entry into the organism, in particular through food. Therefore, it is not a simple physical–chemical process, because active absorption in the digestive system may occur.

Biomagnification occurs when the concentration of a chemical increases substantially (often many orders of magnitude) at different levels of the food chain, from primary producers to top predators. Biomagnification is possible for chemicals that are very persistent in living organisms (neither metabolized nor excreted) and that can be efficiently stored in some tissues (usually lipids or proteins).

- **leaching indices**

They are environmental indices specifically proposed to study the environmental fate of pesticides.

The **GUS index** assesses the leachability of molecules and the possibility of finding these chemicals in groundwater [Gustafson, 1989; Papa, Castiglioni *et al.*, 2004]. It is defined as

$$\text{GUS} = \log_{10}(t_{1/2}) \cdot [4 - \log_{10}(K_{oc})]$$

where $t_{1/2}$ is the half lifetime (in days), quantifying the soil persistence, and K_{oc} is the organic carbon partition coefficient, quantifying the mobility in soil.

The **LEACH index** is a leaching index assessing the potential degree of groundwater and river water contamination [Laskowski, Goring *et al.*, 1982]. It is defined as

$$\text{LEACH} = \frac{S_w \cdot t_{1/2}}{V_p \cdot K_{oc}}$$

where S_w is the water solubility (mg/l), $t_{1/2}$ is the degradation half lifetime in soil (in days), V_p the vapour pressure (Pa), and K_{oc} is the organic carbon partition coefficient. The lower the LEACH value the lower the risk of contamination. LEACH values are expressed on a logarithmic scale to allow comparison with other indices. A modified version of the LEACH index, **modified LEACH index**, was also proposed without taking vapour pressure into account, to avoid a double counting of volatilization that is already considered in disappearance half lifetime:

$$\text{LEACH}_{\text{MOD}} = \frac{S_w \cdot t_{1/2}}{K_{oc}}$$

Based on the Principal Component Analysis, the **LIN index** (leaching index) and the **VIN index** (volatility index) were defined in terms of the first and second PCs, respectively, explaining 92.7% of the total variance [Gramatica and Di Guardo, 2002]. PCs were calculated on a data set of 135 pesticides, described by vapour pressure (V_p), Henry's law constant (H), water solubility (S_w), and octanol/water (K_{ow}) and organic carbon (K_{oc}) partition coefficients. The LIN and VIN indices are defined as the following:

$$\begin{aligned} \text{LIN} &= -0.531 \cdot \log K_{ow} + 0.518 \cdot \log S_w - 0.495 \cdot \log K_{oc} - 0.023 \cdot \log V_p - 0.452 \cdot \log H \\ \text{VIN} &= -0.034 \cdot \log K_{ow} + 0.211 \cdot \log S_w + 0.202 \cdot \log K_{oc} + 0.837 \cdot \log V_p + 0.461 \cdot \log H \end{aligned}$$

The **Global Leachability Index (GLI index)** [Papa, Castiglioni *et al.*, 2004] was defined by Principal Component Analysis, condensing information derived from GUS index, modified LEACH index, and LIN index.

The first PC, explaining 87.5% of the total variance, was assumed as representing the risk (higher positive values higher the risk):

$$\text{GLI} = 0.579 \cdot \text{LIN} + 0.558 \cdot \text{GUS} + 0.595 \cdot \text{LEACH}_{\text{MOD}}$$

GLI values less than -0.5 indicate low-risk compounds, values between -0.5 and 1 medium-risk compounds, and values greater than 1 high-risk compounds.

- **Global Warming Potential (GWP)**

Global warming potential is a measure of how much a given mass (a ton) of greenhouse gas (GHG) is estimated to contribute to global warming, evaluated as its accumulated

radiative effect. It is a relative scale that compares the gas in question to that of the same mass of carbon dioxide CO₂ (whose GWP is by definition 1). A GWP is calculated over a specific time interval (usually 20, 100, or 500 years) [Ivanciuc and Ivanciuc, 2002].

Other properties of chemicals, encountered in environmental studies, are *Long Range Transport* (LRT) [Beyer, Mackey *et al.*, 2000; Leip and Lammel, 2004; Wania and Dugani, 2003], defined as the atmospheric transport of air pollutants within a moving air mass for a distance greater than 100 km, the *mobility index* [Gramatica, Papa *et al.*, 2004], and *Atmospheric Persistent Index* (ATPIN) to evaluate the atmospheric degradability of chemicals [Gramatica, Pilutti *et al.*, 2002, 2003b].

📖 [Basak and Mills, 2005; Cronin, Walker *et al.*, 2003; Drefahl and Reinhard, 1993; Govers, 1990; Halfon, Galassi *et al.*, 1996; Halfon and Reggiani, 1986; Jaworska, Comber *et al.*, 2003; Koch, 1982; Okouchi, Saegusa *et al.*, 1992; Rorije, Van Wezel *et al.*, 1995; Sabljčić, 2001; Sabljčić and Piver, 1992; Todeschini and Gramatica, 1997c]

- **environmental QSAR** → environmental indices
- **environment connectivity descriptors** → connectivity indices
- **E_K polarity scale** → Linear Solvation Energy Relationships (⊙ dipolarity/polarizability term)
- **E_T polarity scale** → Linear Solvation Energy Relationships (⊙ dipolarity/polarizability term)
- **equilibrium constants** → physico-chemical properties
- **equilibrium electronegativity** → electronegativity
- **equipoise random walks** → walk counts
- **equipotent walks** → walk counts

■ equivalence classes

Subsets of equivalent elements according to a specified equivalence relation.

A **decomposition** *N* of a system containing *N* elements is any partition of these elements into disjoint subsets of equivalent elements by a specified equivalence relation.

A **finite probability scheme** can be associated with this decomposition as the following:

Equivalence classes	1, 2, ..., <i>G</i>
Element partition	<i>n</i> ₁ , <i>n</i> ₂ , ..., <i>n</i> _{<i>g</i>} , ..., <i>n</i> _{<i>G</i>}
Probability distribution	<i>p</i> ₁ , <i>p</i> ₂ , ..., <i>p</i> _{<i>g</i>} , ..., <i>p</i> _{<i>G</i>}

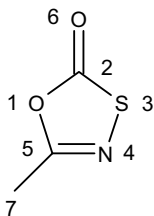
where *G* is the total number of equivalence classes, *p_g* = *n_g*/*N* is the probability of a randomly chosen element belonging to the *g*th subset having *n_g* elements and

$$N = \sum_{g=1}^G n_g.$$

→ *Information content* is a fundamental measure derived from the partitioning of elements in equivalence classes; several molecular descriptors are derived as → *information indices*.

Example E5

Equivalence classes of 5-methyl-1,3,4-oxathiazol-2-one based on the atom types.



Atom	C	O	N	S
n	3	2	1	1
p	3/7	2/7	1/7	1/7

- **E_R radical parameter** → electronic substituent constants
- **error rate** → classification parameters
- **error standard deviation** $\equiv$ residual standard deviation → regression parameters
- **error sum of squares** $\equiv$ residual sum of squares → regression parameters
- **ESSR** $\equiv$ Extended Set of Smallest Rings → ring descriptors
- **E -state index** → electrotopological state indices
- **E -state fields** → electrotopological state indices
- **E -state topological parameter** → Balaban distance connectivity index
- **Estrada Generalized Topological Indices** → variable descriptors
- **Estrada index** → spectral indices ($\odot$ subgraph centrality)

■ **ETA indices** ($\equiv$ Extended Topochemical Atom indices)

ETA indices [Roy and Ghosh, 2003] are both local vertex and graph invariants, defined in the framework of the **Valence Electron Mobile environment (VEM environment)** theory [Pal, Sengupta *et al.*, 1988; Pal, Sengupta *et al.*, 1989], according to which a vertex in the $\rightarrow H$ -depleted molecular graph is considered to be consisted of a core and a valence electronic environment.

The **core count** α_i is a local vertex invariant calculated as

$$\alpha_i = \frac{Z_i - Z_i^v}{Z_i^v} \cdot \frac{1}{L_i - 1}$$

where Z , Z^v , and L are the atomic number, the valence electron number, and the principal quantum number, respectively. The sum of the α values of the nonhydrogen atoms is a simple molecular descriptor related to molecular bulk.

By combining the core count α of an atom with its valence electron number Z^v , the **electronegativity ETA measure** ϵ_i for the i th atom was defined as

$$\epsilon_i = -\alpha_i + 0.3 \cdot Z_i^v$$

The **VEM count** β_i is another local vertex invariant defined as

$$\beta_i = [\beta_s]_i + [\beta_{ns}]_i = \left[\sum_{\sigma(i)} f_\sigma \right] + \left[\sum_{\pi(i)} f_\pi + f_{LP(i)} \right]$$

where $[\beta_s]_i$ and $[\beta_{ns}]_i$ are two other local vertex invariants accounting for σ bonds and π bonds plus lone pairs of each i th atom, respectively, f_σ is the contribution of a σ bond and f_π the

contribution of a π bond, and f_{LP} a correction factor accounting for lone pair electrons; the two summations run over bonds formed by the i th atom.

The contribution of a sigma bond f_{σ} is equal to 0.5 for two bonded atoms of similar ETA electronegativity ($\Delta\epsilon \leq 0.3$) and 0.75 for two bonded atoms of different electronegativity ($\Delta\epsilon > 0.3$). The contribution of a π bond f_{π} arises only from multiple bonds (i.e. double and triple bonds) and is equal to (a) $f_{\pi} = 1$ for two bonded atoms of similar electronegativity ($\Delta\epsilon \leq 0.3$); (b) $f_{\pi} = 1.5$ for two bonded atoms of different electronegativity ($\Delta\epsilon > 0.3$) or for conjugated (nonaromatic) π systems; and (c) $f_{\pi} = 2$ for aromatic π systems. The term f_{LP} is equal to 0.5 per atom with a lone pair of electrons capable of resonance with aromatic ring (e.g., nitrogen of aniline, oxygen of phenol, etc.).

From the atomic VEM counts, two molecular descriptors are derived:

$$\beta_s = \frac{\sum_{i=1}^A [\beta_s]_i}{A} \quad \beta_{ns} = \frac{\sum_{i=1}^A [\beta_{ns}]_i}{A}$$

where A is the number of nonhydrogen atoms; β_s accounts for all the σ bonds in the molecule and can be considered a relative measure of the number of electronegative atoms in the molecule, while β_{ns} is calculated from all π bonds and electron lone pairs and can be considered a relative measure of electron-richness (unsaturation) of the molecule.

From the two vertex invariants α and β , the **VEM vertex count** γ_i is defined as

$$\gamma_i = \frac{\alpha_i}{\beta_i}$$

The **composite ETA index** η is a molecular descriptor defined as

$$\eta = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \left(\frac{\gamma_i \cdot \gamma_j}{d_{ij}^2} \right)^{1/2}$$

where d_{ij} is the topological distance between i - j atoms.

Defined analogously to the composite ETA index, but considering only contributions from pairs of bonded atoms, the **local ETA index** is calculated as

$$\eta^{loc} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\gamma_i \cdot \gamma_j)^{1/2}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix*, equal to 1 only for pairs of bonded atoms and zero otherwise.

Moreover, the **composite reference ETA index**, denoted as η_R , is calculated as the composite ETA index from a molecular graph where all heteroatoms are substituted by carbon atoms and all the multiple bonds by single bonds. Then, the **functionality index** η_F is defined as the difference between the composite reference ETA index and the composite ETA index:

$$\eta_F = \eta_R - \eta$$

This index was proposed to measure the molecule functionality, here intended as the presence of heteroatoms and multiple bonds. To avoid molecular size dependence, this functionality index is normalized by the number of atoms A :

$$\eta'_F = \frac{\eta_R - \eta}{A}$$

In a similar way to the functionality index, the **local functionality index** is defined as

$$\eta_F^{loc} = \eta_R^{loc} - \eta^{loc}$$

where η_R^{loc} is the local index for the corresponding reference alkane.

The **branching ETA index** η_B is calculated from the local ETA index of the reference alkane η_R^{loc} of the considered compound compared to the index value η_N^{loc} of the corresponding normal alkane, that is, the straight chain graph containing the same number of vertices as the compound considered:

$$\eta_B = \eta_N^{loc} - \eta_R^{loc} + 0.086 \cdot NRG$$

where NRG is the number of rings in the molecular graph of the reference alkane. η'_B is the branching ETA index normalized by the number of atoms A .

The calculation of η_N^{loc} can be easily performed by the following relationship:

$$\eta_N^{loc} = 1.414 + (A - 3) \cdot 0.5$$

where A is the total number of nonhydrogen atoms and the relationship holds only for compounds for which $A > 2$.

The **shape ETA indices** are derived from the core counts α as

$$\frac{[\sum_i \alpha_i]_p}{\sum_{i=1}^A \alpha_i}, \quad \frac{[\sum_i \alpha_i]_y}{\sum_{i=1}^A \alpha_i}, \quad \frac{[\sum_i \alpha_i]_x}{\sum_{i=1}^A \alpha_i}$$

where p , y , and x stand for the summation of α values of the vertices that are joined to one, three, and four other nonhydrogen atoms, respectively.

The contribution of a specific i th atom to the composite ETA index was also defined as local vertex invariant and called **atom-level composite ETA index**:

$$[\eta]_i = \sum_{j=1}^A \left(\frac{\gamma_i \cdot \gamma_j}{d_{ij}} \right)^{1/2} \quad j \neq i$$

Analogously, the reference and functionality atom-level indices and their normalized counterparts were also derived and denoted as $[\eta_R]_i$, $[\eta_F]_i$, and $[\eta'_F]_i$.

ETA indices are an extension of the **TAU indices** (or *Topochemically Arrived Unique indices*) [Pal, Purkayastha *et al.*, 1992; Pal, Sengupta *et al.*, 1988, 1989, 1990], which were defined some years before in the framework of a previous version of the Valence Electron Mobile environment (VEM environment). TAU indices are calculated from previous definitions of core count and VEM count and include four indices: the composite topochemical index, denoted by T (similar to the composite ETA index), the functionality index, denoted by F , the skeletal index, denoted by T_R , and the simple branching index, denoted by B . In QSAR studies, these indices were used in combination with $\rightarrow$ STIMS indices, $\rightarrow$ connectivity indices, and some $\rightarrow$ information indices [Roy, Pal *et al.*, 1999, 2001].

▢ [Pal, Purkayastha *et al.*, 1992; Pal, Sengupta *et al.*, 1990; Roy and Ghosh, 2004a, Roy and Ghosh, 2004b, Roy and Ghosh, 2004c, Roy and Ghosh, 2005, Roy and Ghosh, 2006a, Roy and Ghosh, 2006b, Roy and Ghosh, 2007; Roy and Saha, 2003a, 2003b, 2004; Roy, Sanyal *et al.*, 2006, 2007; Roy and Sanyal, 2006; Roy and Toropov, 2005]

- **ET method** $\equiv$ *Electronic-Topological method*
- **Euclidean-adjacency map matrix** $\rightarrow$ biodescriptors ($\odot$ proteomics maps)
- **Euclidean connectivity index** $\rightarrow$ connectivity indices
- **Euclidean degree** $\equiv$ *geometric distance degree* $\rightarrow$ molecular geometry
- **Euclidean distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)
- **Euclidean distance matrix** $\rightarrow$ similarity/diversity
- **Euclidean-distance map matrix** $\rightarrow$ biodescriptors ($\odot$ proteomics maps)
- **Eulerian walk** $\rightarrow$ graph

■ Euler formula

The fundamental relation between the number of vertices V , edges E , and faces F of convex polyhedra was proposed in 1758 by Euler [Euler, 1758] as

$$V - E + F = 2$$

This relationship holds for any division of a sphere into polygons.

Euler's formula relating the number of edges, vertices, and faces of a convex polyhedron was studied and generalized by Cauchy [Cauchy, 1813] and [L'Huillier, 1861] and is at the origin of topology.

From these primary indices, that is, vertices, edges, and faces, two secondary topological indices were derived, called **Schläfli indices** [Bucknum and Castro, 2005a], namely, the polygonality (n) and the connectivity (p).

Polygonality refers to the weighted average number of sides of the polygonal faces of a polyhedron, computed by drawing polygons and also considering the terminal atoms. *Connectivity* refers to the average of connectivity of the vertices of a polyhedron. Moreover, the ratio of polygonality to connectivity was proposed as a compactness index and called **Schläfli topological form index** [Bucknum and Castro, 2005b]:

$$I = \frac{n}{p}$$

Based on the two topological identities, (1) each edge of a polyhedron is shared by two faces, then $n \times F = 2 \times E$, and (2) each edge terminates with two vertices, then $p \times V = 2 \times E$, the following Schläfli relationship was found:

$$\frac{1}{n} + \frac{1}{p} - \frac{1}{2} = \frac{1}{E}$$

The Schläfli relation establishes a connection between secondary topological indices, n and p , and primary topological indices V , E , and F .

Polygonality and connectivity can be calculated not only for 3D pattern but also for 2D patterns, introducing the concept of cell. A cell is a topological concept, like a sphere, which involves the division of the plane into fused polygons in which some edges form a boundary.

In effect, the Euler relation in a 2D space, a cell, is

$$V - E + F = 1$$

Noting that faces are exactly what in graph theory are called cycles, the previous Euler formula provides the definition of $\rightarrow$ *cyclomatic number* C :

$$F \equiv C = E - V + 1 \equiv B - A + 1$$

where A and B are the number of vertices and edges in a $\rightarrow$ *molecular graph*.

■ EVA descriptors ($\equiv$ EigenVAlue descriptors)

EVA descriptors were proposed by Ferguson *et al.* [Ferguson, Heritage *et al.*, 1997; Turner, Willett *et al.*, 1997] as an approach to extract chemical structural information from mid- and near-infrared spectra. The approach is to use, as a multivariate descriptor, the vibrational frequencies of a molecule, a fundamental molecular property characterized reliably and easily from the potential energy function. The EigenVAlue (EVA) descriptor is a function of the eigenvalues obtained from the normal coordinate matrix; it corresponds to the fundamental vibrational frequencies of the molecule, which can be calculated using standard quantum or molecular mechanical methods from $\rightarrow$ *computational chemistry*.

The eigenvectors, corresponding to atomic displacement, are not considered as molecular descriptors.

Since the number of vibrational normal modes varies with the number of atoms A in a molecule (actually $3A-6$ for a molecule without axial symmetry or $3A-5$ for a linear molecule), each set of eigenvalues will generally be of different dimensionality. Thus, to obtain comparability among the molecules and $\rightarrow$ *uniform-length descriptors*, the frequency range chosen is 0 and 4000 cm^{-1} to encompass the frequencies of all fundamental molecular vibrations, and the eigenvalues are projected onto a bounded frequency scale (BFS) where the vibrational frequencies are represented by points along this axis, obtaining a scale of fixed dimensionality. Then a Gaussian function of fixed standard deviation σ is centered at each eigenvalue projection over the BFS axis, resulting in a series of $3A-6$ (or $3A-5$) identical and overlapping Gaussians (Figure E5).

The value of the **EVA function** at any point x on the BFS axis is determined by summing the contributions from each and every one of the $3A-6$ (or $3A-5$) overlaid Gaussians at that point:

$$\text{EVA}_x = \sum_{i=1}^{3A-6} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-(x-\lambda_i)^2/2\sigma^2}$$

where λ_i is the i th vibrational frequency (eigenvalue) for the molecular structure.

Finally, the EVA function is sampled at fixed increments of $L\text{ cm}^{-1}$ along the BFS axis; this sampling results in $4000/L$ values that define the EVA uniform-length descriptor.

The choice of σ defines the degree to which the fundamental vibrations overlap: σ values determine the number of and extent to which vibrations of a particular frequency in one structure can be statistically related to those in the other structures (interstructural overlap); moreover, such values govern the extent to which vibrations within the same structure may overlap at nonnegligible values (intrastructural overlap).

After the frequency range is fixed, the sampling parameter L determines the total number of EVA descriptor elements; L should be maximized so as to reduce computational overhead and minimized to catch all the useful information. The optimal L value depends on the selected σ value.

Characteristic value of the Gaussian standard deviation σ is 10 cm^{-1} (range $10\text{--}20\text{ cm}^{-1}$) and of the sampling increment L is 5 cm^{-1} (range $2\text{--}20\text{ cm}^{-1}$), resulting in 800 descriptor variables.

The EVA descriptors are among $\rightarrow 3D$ descriptors, independent of any molecular alignment, giving information about molecular size, shape, and electronic properties. Moreover, the EVA descriptors show only a moderate dependence on conformations.

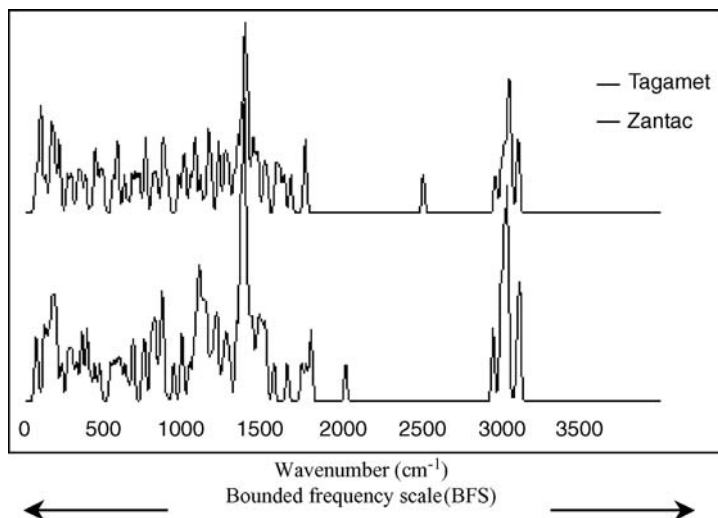


Figure E5 Bounded Frequency Scale with superimposed EVA descriptors for two compounds.

EEVA descriptors (or **Electronic EigenVALUE descriptors**) are $\rightarrow$ *vectorial descriptors* proposed as a modification of EVA [Tuppurainen, 1999a]. Semiempirical molecular orbital energies, that is, the eigenvalues of the Schrödinger equation, are used instead of the vibrational frequencies of the molecule. Each molecular orbital energy of the molecule is first projected onto a bounded energy scale (the range can be $-45 \div 10\text{ eV}$, but this depends on the quantum-chemical method used to calculate orbital energies). Then a Gaussian function of fixed standard deviation σ (0.50 eV was proposed) is centered at each MO energy projection resulting in a series identical and overlapping Gaussians. Once an appropriate sampling interval L (0.25 eV was proposed) has been chosen, the whole range is sampled; the EEVA descriptors at each sampling point x are defined as

$$\text{EEVA}_x = \sum_i \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-(x-\epsilon_i)^2/2\sigma^2}$$

where the summation goes over all Gaussian functions and ϵ_i is the i th molecular orbital energy of the molecule. By using the parameter values defined above, this procedure provides a descriptor vector consisting of 220 (i.e., $55/0.25$) elements so that dimensionality is much lower than that of the EVA descriptor vector.

📖 [Baumann, 1999; Benigni, Passerini *et al.*, 1999a, 1999b; Borosy, Balogh *et al.*, 2005; Devillers, 2000; Ford, Phillips *et al.*, 2004; Ginn, Turner *et al.*, 1997; Heritage, Ferguson *et al.*, 1998; Makhija and Kulkarni, 2001a; Takane and Mitchell, 2004; Tuppurainen and

Ruuskanen, 2000; Tuppurainen, Viisas *et al.*, 2002; Turner, Willett *et al.*, 1999; Turner and Willett, 2000a, 2000b]

- **EVA function** → EVA descriptors
- **evaluation set** $\equiv$ *external evaluation set* → data set
- **Evans extended connectivity indices** → connectivity indices
- **EV_{TYPE} descriptors** → van der Waals excluded volume method
- **EV_{WHOLE} descriptors** → van der Waals excluded volume method
- **even/odd distance indices** → Wiener index
- **even/odd Wiener polynomial descriptors** → Wiener index
- **excess electron polarizability** → electric polarization descriptors
- **excess molar refractivity** → physico-chemical properties ($\odot$ molar refractivity)
- **Exner statistics** → regression parameters
- **expanded distance Cluj matrices** → expanded distance matrices
- **expanded distance indices** → expanded distance matrices

■ **expanded distance matrices** ($\equiv$ *distance-extended matrices*)

The original *expanded distance matrix*, denoted as Π , proposed by Tratch [Tratch, Stankevitch *et al.*, 1990] is a square symmetric $A \times A$ matrix representing a $\rightarrow H$ -depleted molecular graph with A vertices whose diagonal entries are equal to zero and each off-diagonal entry is defined as

$$[\Pi]_{ij} = \begin{cases} \mu_{ij} \cdot n_{ij} \cdot d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} denotes the $\rightarrow$ *topological distance* between vertices v_i and v_j , n_{ij} the number of external paths including the shortest path between the considered vertices with length $m \geq d_{ij}$, and μ_{ij} is the number of shortest paths between v_i and v_j that is equal to 1 for any pair of vertices in acyclic graphs. The number of external paths n_{ij} with respect to the shortest path $i-j$ is just the same for each of the shortest paths connecting vertices v_i and v_j ; therefore, the product $n_{ij} \times \mu_{ij}$ gives the total number of external paths in cyclic graphs.

Applying the $\rightarrow$ *Wiener operator* Wi to the expanded distance matrix, a molecular descriptor called **expanded Wiener number** $\tilde{W}$ is defined as

$$\tilde{W} \equiv Wi(\Pi) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Pi]_{ij}$$

For acyclic graphs, the expanded distance matrix can be obtained simply as

$$\Pi = \mathbf{D} \otimes \mathbf{W}$$

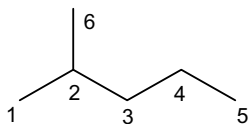
where $\mathbf{D}$ and $\mathbf{W}$ denote the $\rightarrow$ *distance matrix* and the $\rightarrow$ *Wiener matrix*, respectively, and $\otimes$ indicates the $\rightarrow$ *Hadamard matrix product*. Therefore, the corresponding expanded Wiener number is calculated as

$$\tilde{W} \equiv \sum_{i=1}^{A-1} \sum_{j=i+1}^A d_{ij} \cdot N_i \cdot N_j$$

where N_i and N_j are the number of vertices on each side of the path $i-j$, including both vertices i and j , respectively. If only the bond contributions are considered ($d_{ij} = 1$), the expanded Wiener number coincides with the $\rightarrow$ *Wiener index*.

Example E6

Calculation of the expanded Wiener index and the Wiener index for 2-methylpentane.



vertex pair (i, j)	d_{ij}	N_i	N_j	$d_{ij} \cdot N_i \cdot N_j$	vertex pair (i, j)	d_{ij}	N_i	N_j	$d_{ij} \cdot N_i \cdot N_j$
(1, 2)	1	1	5	5	(2, 6)	1	5	1	5
(1, 3)	2	1	3	6	(3, 4)	1	4	2	8
(1, 4)	3	1	2	6	(3, 5)	2	4	1	8
(1, 5)	4	1	1	4	(3, 6)	2	3	1	6
(1, 6)	2	1	1	2	(4, 5)	1	5	1	5
(2, 3)	1	3	3	9	(4, 6)	3	2	1	6
(2, 4)	2	3	2	12	(5, 6)	4	1	1	4
(2, 5)	3	3	1	9					

Expanded Wiener number:

$$\tilde{W} = 5 + 6 + 6 + 4 + 2 + 9 + 12 + 9 + 5 + 8 + 8 + 6 + 5 + 6 + 4 = 95$$

$$\text{Wiener index: } W = d_{12} \cdot N_1 \cdot N_2 + d_{23} \cdot N_2 \cdot N_3 + d_{26} \cdot N_2 \cdot N_6 + d_{34} \cdot N_3 \cdot N_4 + d_{45} \cdot N_4 \cdot N_5 = 5 + 9 + 5 + 8 + 5 = 32$$

Based on different powers of the topological distance, **generalized expanded Wiener numbers** $\tilde{W}^\lambda$, also called **Tratch–Stankevitch–Zefirov-type indices**, were proposed as [Klein and Gutman, 1999]

$$\tilde{W}^\lambda = \sum_{i=1}^{A-1} \sum_{j=i+1}^A d_{ij}^\lambda \cdot N_i \cdot N_j$$

where λ is any integer and the relation is valid only for trees.

Note that $\lambda = 0$ and $\lambda = 1$ result in, respectively, the $\rightarrow$ *hyper-Wiener index* and the expanded Wiener number; formally, for $\lambda \rightarrow -\infty$, the generalized expanded Wiener number should coincide with the well-known Wiener index. For cycle-containing graphs, the generalized expanded Wiener numbers were calculated as

$$\tilde{W}^\lambda = \sum_{i=1}^{A-1} \sum_{j=i+1}^A d_{ij}^\lambda \cdot \#_{ij}$$

where $\#_{ij}$ is the number of pairs of vertices (v_p, v_q) of the graph such that there is a geodesic (that is, the shortest path) between them containing both vertices v_i and v_j . The number $\#_{ij}$ is equal to the number of all shortest paths containing the path p_{ij} as a subpath, that is, the total number of external paths.

A generalization of the expanded distance matrix was proposed by Diudea [Diudea and Gutman, 1998] to define new matrices derived from the Hadamard matrix product between the distance matrix $\mathbf{D}$ and a general square $A \times A$ matrix $\mathbf{M}$ as

$$\mathbf{D\_M} = \mathbf{D} \otimes \mathbf{M}$$

If $\mathbf{M}$ is one among the $\rightarrow$ Cluj matrices then **expanded distance Cluj matrices** are obtained. Next, if $\mathbf{M}$ is the $\rightarrow$ Szeged matrix then the **expanded distance Szeged matrix** is derived; analogously, **expanded distance Szeged property matrices** [Minailiuc, Katona *et al.*, 1998] and **expanded distance walk matrices** are derived from $\rightarrow$ Szeged property matrices and $\rightarrow$ walk matrices.

From these matrices, two kinds of molecular indices are obtained. The **expanded distance indices**, denoted by DM_p or $D^U M_p$, depending on whether the matrix is symmetric or unsymmetric, are calculated by applying the $\rightarrow$ Wiener operator Wi to both expanded distance symmetric $\mathbf{M}$ and unsymmetric $\mathbf{UM}$ matrices:

$$DM_p \equiv Wi(\mathbf{D\_M}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D\_M}]_{ij} \quad D^U M_p \equiv Wi(\mathbf{D\_UM}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D\_UM}]_{ij}$$

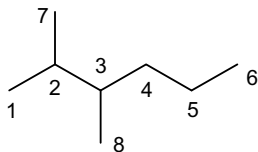
The **expanded square distance indices** are calculated, only from unsymmetrical $\mathbf{D\_UM}$ matrices, by applying the $\rightarrow$ orthogonal Wiener operator $Wi^\perp$ as

$$D^2 M_p \equiv Wi^\perp(\mathbf{D\_UM}) = \sum_{i=1}^A \sum_{j=i}^A ([\mathbf{D\_UM}]_{ij} \cdot [\mathbf{D\_UM}]_{ji})$$

Note that $D^2 M_p$ indices involve square topological distances.

Example E7

Calculation of the expanded Wiener index and the Wiener index for 2,3-dimethylhexane.



$\mathbf{D}$									
	1	2	3	4	5	6	7	8	VS_i
1	0	1	2	3	4	5	2	3	20
2	1	0	1	2	3	4	1	2	14
3	2	1	0	1	2	3	2	1	12
4	3	2	1	0	1	2	3	2	14
5	4	3	2	1	0	1	4	3	18
6	5	4	3	2	1	0	5	4	24
7	2	1	2	3	4	5	0	3	20
8	3	2	1	2	3	4	3	0	18
CS_j	20	14	12	14	18	24	20	18	140

Wiener index (W) = 70

UCJD _p										D_UCJD _p									
	1	2	3	4	5	6	7	8	VS _i		1	2	3	4	5	6	7	8	VS _i
1	0	1	1	1	1	1	1	1	7	1	0	1	2	3	4	5	2	3	20
2	7	0	3	3	3	3	7	3	29	2	7	0	3	6	9	12	7	6	50
3	5	5	0	5	5	5	5	7	37	3	10	5	0	5	10	15	10	7	62
4	3	3	3	0	6	6	3	3	27	4	9	6	3	0	6	12	9	6	51
5	2	2	2	2	0	7	2	2	19	5	8	6	4	2	0	7	8	6	41
6	1	1	1	1	1	0	1	1	7	6	5	4	3	2	1	0	5	4	24
7	1	1	1	1	1	1	0	1	7	7	2	1	2	3	4	5	0	3	20
8	1	1	1	1	1	1	1	0	7	8	3	2	1	2	3	4	3	0	18
CS _j	20	14	12	14	18	24	20	18	140	CS _j	44	25	18	23	37	60	44	35	286
Wiener index (W) = 70										D ^U CJD _p = 143									
hyper-Cluj-distance index (CJD _p) = 143										D ² CJD _p = 605									
SCJD _p										D_SCJD _p									
	1	2	3	4	5	6	7	8	VS _i		1	2	3	4	5	6	7	8	VS _i
1	0	7	5	3	2	1	1	1	20	1	0	7	10	9	8	5	2	3	44
2	7	0	15	9	6	3	7	3	50	2	7	0	15	18	18	12	7	6	83
3	5	15	0	15	10	5	5	7	62	3	10	15	0	15	20	15	10	7	92
4	3	9	15	0	12	6	3	3	51	4	9	18	15	0	12	12	9	6	81
5	2	6	10	12	0	7	2	2	41	5	8	18	20	12	0	7	8	6	79
6	1	3	5	6	7	0	1	1	24	6	5	12	15	12	7	0	5	4	60
7	1	7	5	3	2	1	0	1	20	7	2	7	10	9	8	5	0	3	44
8	1	3	7	3	2	1	1	0	18	8	3	6	7	6	6	4	3	0	35
CS _j	20	50	62	51	41	24	20	18	286	CS _j	44	83	92	81	79	60	44	35	518
hyper-Cluj-distance index (CJD _p) = 143										DCJD _p = 259									

Using the reciprocal distance matrix $\mathbf{D}^{-1}$ instead of the simple distance matrix in the Hadamard matrix product, **expanded reciprocal distance matrices** $\mathbf{H\_M}$ are obtained as

$$\mathbf{H\_M} = \mathbf{D}^{-1} \otimes \mathbf{M}$$

where $\mathbf{M}$ can be any square $A \times A$ matrix as defined above.

From these matrices, the **expanded reciprocal distance indices** (HM_p and $H^U M_p$) and **expanded reciprocal square distance indices** ($H^2 M_p$) are derived as

$$HM_p \equiv Wi(\mathbf{H\_M}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{H\_M}]_{ij} \quad H^U M_p \equiv Wi(\mathbf{H\_UM}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{H\_UM}]_{ij}$$

and, only for unsymmetrical $\mathbf{H\_UM}$ matrices,

$$H^2 M_p \equiv Wi^{\perp}(\mathbf{H\_UM}) = \sum_{i=1}^A \sum_{j=i}^A ([\mathbf{H\_UM}]_{ij} \cdot [\mathbf{H\_UM}]_{ji})$$

Moreover, following the same procedure, other expanded matrices are defined [Minailiuc, Katona *et al.*, 1998] replacing the topological distance matrix **D** by the $\rightarrow$ *geometry matrix G* as

$$\mathbf{G\_M} = \mathbf{G} \otimes \mathbf{M} \quad \text{and} \quad \mathbf{G^{-1}\_M} = \mathbf{G^{-1}} \otimes \mathbf{M}$$

where **M** can be any square $A \times A$ matrix as defined above and $\mathbf{G^{-1}}$ is the $\rightarrow$ *reciprocal geometry matrix*. Therefore, **G_M** and **G⁻¹_M** matrices can be called **expanded geometric distance matrices** and **expanded reciprocal geometric distance matrices**, respectively. The corresponding molecular indices are defined as

$$\begin{aligned} GM_p &\equiv Wi(\mathbf{G\_M}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{G\_M}]_{ij} & G^U M_p &\equiv Wi(\mathbf{G\_UM}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{G\_UM}]_{ij} \\ H_G M_p &\equiv Wi(\mathbf{G^{-1}\_M}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{G^{-1}\_M}]_{ij} \\ H_G^U M_p &\equiv Wi(\mathbf{G^{-1}\_UM}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{G^{-1}\_UM}]_{ij} \end{aligned}$$

and, only for unsymmetrical **G_UM** and **G⁻¹_UM** matrices,

$$\begin{aligned} G^2 M_p &\equiv Wi^\perp(\mathbf{G\_UM}) = \sum_{i=1}^A \sum_{j=i}^A ([\mathbf{G\_UM}]_{ij} \cdot [\mathbf{G\_UM}]_{ji}) \\ H_G^2 M_p &\equiv Wi^\perp(\mathbf{G^{-1}\_UM}) = \sum_{i=1}^A \sum_{j=i}^A ([\mathbf{G^{-1}\_UM}]_{ij} \cdot [\mathbf{G^{-1}\_UM}]_{ji}) \end{aligned}$$

The symbols of the molecular descriptors derived from the most common expanded distance matrices are listed in the Table E12.

Table E12 Wiener-type indices derived from some expanded distance matrices.

Molecular descriptor	... obtained from the matrix ...
$\tilde{W}$	Expanded distance matrix
HW_p	Expanded reciprocal distance path-Wiener matrix
GW_p	Expanded geometric distance path-Wiener matrix
$H_G W_p$	Expanded reciprocal geometric distance path-Wiener matrix
$DCJD_p$	Expanded distance symmetric path-Cluj-distance matrix
$D^U CJD_p$ and $D^2 CJD_p$	Expanded distance unsymmetric path-Cluj-distance matrix
$HCJD_p$	Expanded reciprocal distance symmetric path-Cluj-distance matrix
$H^U CJD_p$ and $H^2 CJD_p$	Expanded reciprocal distance unsymmetric path-Cluj-distance matrix
$GCJD_p$	Expanded geometric distance symmetric path-Cluj-distance matrix
$G^U CJD_p$ and $G^2 CJD_p$	Expanded geometric distance unsymmetric path-Cluj-distance matrix
$H_G CJD_p$	Expanded reciprocal geometric distance symmetric path-Cluj-distance matrix
$H_G^U CJD_p$ and $H_G^2 CJD_p$	Expanded reciprocal geometric distance unsymmetric path-Cluj-distance matrix
$DCJ\Delta_p$	Expanded distance symmetric path-Cluj-detour matrix
$D^U CJ\Delta_p$ and $D^2 CJ\Delta_p$	Expanded distance unsymmetrical path-Cluj-detour matrix
$HCJ\Delta_p$	Expanded reciprocal distance symmetric path-Cluj-detour matrix
$H^U CJ\Delta_p$ and $H^2 CJ\Delta_p$	Expanded reciprocal distance unsymmetrical path-Cluj-detour matrix
$GCJ\Delta_p$	Expanded geometric distance symmetric path-Cluj-detour matrix
$G^U CJ\Delta_p$ and $G^2 CJ\Delta_p$	Expanded geometric distance unsymmetrical path-Cluj-detour matrix

(Continued)

Table E12 (Continued)

Molecular descriptor	... obtained from the matrix ...
$H_G C J \Delta_p$	Expanded reciprocal geometric distance symmetric path-Cluj-detour matrix
$H_G^U C J \Delta_p$ and $H_G^2 C J \Delta_p$	Expanded reciprocal geometric distance unsymmetrical path-Cluj-detour matrix
DSZ_p	Expanded distance symmetric path-Szeged matrix
$D^U SZ_p$ and $D^2 SZ_p$	Expanded distance unsymmetrical path-Szeged matrix
HSZ_p	Expanded reciprocal distance symmetric path-Szeged matrix
$H^U SZ_p$ and $H^2 SZ_p$	Expanded reciprocal distance unsymmetrical path-Szeged matrix
GSZ_p	Expanded geometric distance symmetric path-Szeged matrix
$G^U SZ_p$ and $G^2 SZ_p$	Expanded geometric distance unsymmetrical path-Szeged matrix
$H_G SZ_p$	Expanded reciprocal geometric distance symmetric path-Szeged matrix
$H_G^U SZ_p$ and $H_G^2 SZ_p$	Expanded reciprocal geometric distance unsymmetrical path-Szeged matrix

If the calculation of indices in Table E12 is performed by summing only the matrix entries corresponding to pairs of adjacent vertices, then similar edge-defined indices can be calculated.

📖 [Diudea, Pârv *et al.*, 1997b]

- **expanded distance Szeged matrix** → expanded distance matrices
- **expanded distance Szeged property matrices** → expanded distance matrices
- **expanded distance walk matrices** → expanded distance matrices
- **expanded geometric distance matrices** → expanded distance matrices
- **expanded matrices** → matrices of molecules
- **expanded reciprocal distance indices** → expanded distance matrices
- **expanded reciprocal distance matrices** → expanded distance matrices
- **expanded reciprocal geometric distance matrices** → expanded distance matrices
- **expanded reciprocal square distance indices** → expanded distance matrices
- **expanded square distance indices** → expanded distance matrices
- **expanded Wiener number** → expanded distance matrices
- **expected square error** $\equiv$ *mean square error* → regression parameters
- **experimental design** → chemometrics

■ experimental measurements

Experimental measurements are the basis from which numerical or graphical information can be extracted by experiments. An experiment is a well-defined operational procedure that measures a quantity for a given sample; hence, experimental quantities are quantities measured by experimental measurement.

→ *Physico-chemical properties* and spectroscopic signals constitute the most important class of experimental chemical measurements, also playing a fundamental role as → *molecular descriptors* both for their availability as well as their interpretability. Examples of physico-chemical measurable quantities are refractive indices, molar refractivities, parachors, densities, solubilities, partition coefficients, dipole moments, chemical shifts, retention times, spectroscopic signals, rate constants, equilibrium constants, vapour pressures, boiling and melting points, acid dissociation constants, and so on [Baum, 1997; Horvath, 1992; Lyman, Reehl *et al.*, 1982; Reid, Prausnitz *et al.*, 1988].

→ *Biological activities*, → *toxicological indices*, → *environmental indices* are other basic experimental quantities that are often considered as responses in QSAR modeling. Examples of biological measurable quantities are effects due to a concentration of a compound, binding affinities, toxicities, inhibition constants, carcinogenicity, mutagenicity, and teratogenicity. Examples of environmental measurable quantities are biochemical oxygen demand, chemical oxygen demand, biodegradability, bioconcentration factors, atmospheric residence time, volatilization from soil, and rate constants of atmospheric degradation reactions.

From a theoretical point of view, experimental properties $\mathcal{P}$ can be distinguished with respect to their behavior in a system S . A molecular property $\mathcal{P}$ may be categorized in terms of its behavior under the hypothesis that a system $S = \mathcal{A} \cup \mathcal{B}$ breaks up into two separate non-interacting subsystems $\mathcal{A}$ and $\mathcal{B}$ [Trinajstić, Randić *et al.*, 1986].

The physical behavior of the property has at least four mathematical possibilities of interest:

- (1) $\mathcal{P}(S) = \mathcal{P}(\mathcal{A}) + \mathcal{P}(\mathcal{B})$
- (2) $\mathcal{P}(S) = \mathcal{P}(\mathcal{A})$ or $\mathcal{P}(\mathcal{B})$
- (3) $\mathcal{P}(S) = \mathcal{P}(\mathcal{A}) \cdot \mathcal{P}(\mathcal{B})$
- (4) $\mathcal{P}(S) = \partial \mathcal{P}(\mathcal{A}) \cdot \mathcal{P}(\mathcal{B}) - \mathcal{P}(\mathcal{A}) \cdot \partial \mathcal{P}(\mathcal{B})$

where $\mathcal{A} \cap \mathcal{B} = \emptyset$. These properties are termed (1) *additive*, (2) *constantive*, (3) *multiplicative*, and (4) *derivative*, respectively. Additive and constantive properties correspond to those properties called, in physical language, extensive and intensive properties.

Many physico-chemical properties and biological activities seem to fall within the domain of additive properties. Examples of constantive properties are local molecular properties, such as dissociation energy for a localized bond or ionization potential. Characteristic multiplicative “properties” are wave functions, Kekulé structure counts, and probabilities. The derivative properties are associated with the corresponding multiplicative property $\mathcal{P}$.

The choice of suitable QSAR/QSPR approaches as well as effective molecular descriptors depends on the characteristic behavior of the property studied.

For each experimental quantity, several estimation methods are known; moreover, for many of them, theoretical methods and/or empirical models were proposed to obtain reliable estimates avoiding experimental measurements.

In any case, experimental values or their estimated values are commonly used as molecular descriptors (i.e., as predictors in X block) or constitute the response (i.e., in Y block) that has to be modeled by other descriptors, that is, reproduced by theoretical models.

Some experimental quantities, such as spectra, need to be transformed in some way before they are used as descriptors. For example, infrared spectra signals (IR spectra) sampled at 10 cm^{-1} in the fingerprint region ($1500\text{--}600 \text{ cm}^{-1}$) were used as → *vectorial descriptors*, each spectrum being scaled in the range 0–100 [Benigni, Passerini *et al.*, 1999a].

📖 [Chaumat, Chamel *et al.*, 1992; Crebelli, Andreoli *et al.*, 1992; Dearden, 1990; Gasteiger, 1988; Horvath, 1988, 1992; Jochum, Hicks *et al.*, 1988; Müller and Klein, 1991]

- **explanatory variables** $\equiv$ *independent variables* → data set
- **exploratory data analysis** → chemometrics
- **exponential cost function** → regression parameters

- **exponential product-sum connectivities** → exponential sum connectivities
- **exponential similarity index** → quantum similarity

■ exponential sum connectivities

These are → *local vertex invariants* proposed by Balaban [Balaban and Catana, 1993] with the aim of obtaining high discrimination among the vertices of a → *H-depleted molecular graph*. They are denoted by c_i and are expressed in logarithmic units as

$$\log c_i = \left(\sum_{k=1}^{\eta_i} k^z \cdot \prod_{v_j \in V_{ik}} G_{jk} \right) \cdot \log G_i$$

where the sum runs over all the distances from vertex v_i to any other vertex in the graph up to its → *atom eccentricity* η_i , that is, the maximum distance, and it involves the products of G values of the atoms v_j , located at distance k from the vertex v_i and constituting the vertex subset V_{ik} ; G_i is the value of local invariant G for the considered vertex v_i defined as

$$\log G_i = \left(\prod_{j=1}^{\delta_i} g_j \right) \cdot \log g_i \quad \text{being} \quad g_i = (\delta_i)^{-1/2} \cdot \left(\sum_{j=1}^{\delta_i} \delta_j \right)^{-1}$$

where the product involves all g values of the δ_i first neighbors of the vertex v_i , the local invariant g being defined as a function of the → *vertex degree* δ_i of the considered atom and the vertex degrees δ_j of its first neighbors. The exponent z in the formula of the c invariant can be either equal to $+1$, leading to **distance-enhanced exponential sum connectivities** or equal to -1 leading to **distance-normalized exponential sum connectivities**.

Analogously, other highly discriminating local invariants were proposed, denoted by c'_i and called **exponential product-sum connectivities**:

$$\log c'_i = \left(\sum_{k=1}^{\eta_i} k^z \cdot \prod_{v_j \in V_{ik}} G'_{jk} \right) \cdot \log G'_i$$

where

$$\log G'_i = \left(\prod_{j=1}^{\delta_i} g'_j \right) \cdot \log g'_i \quad \text{and} \quad g'_i = (\delta_i)^{-1/2} \cdot \left(\sum_{j=1}^{\delta_i} \delta_j + \prod_{j=1}^{\delta_i} \delta_j \right)^{-1}$$

Exponential sum connectivities c_i and c'_i take real values in the 0–1 range. Larger values are assigned to vertices that have higher degrees, are closer to the → *graph center*, and are closer to a vertex of higher degree.

By summing the exponential sum connectivities of all vertices, the corresponding molecular descriptors are derived:

$$XC = \sum_{i=1}^A c_i \quad \text{and} \quad XC' = \sum_{i=1}^A c'_i$$

where A is the number of vertices in the molecular graph. These topological indices decrease asymptotically toward zero with increasing number of vertices in linear chain graphs, while they increase toward infinity with increasing the number of vertices in highly branched graphs.

Exponential sum connectivities c_i and c'_i can also be calculated for graph fragments or organic substituents X , considering the dimer molecule $X-X$ (e.g., for the ethyl group *n*-butane is taken)

[Balaban and Catana, 1994]. The values of c_i or c'_i are calculated for each half of such a dimer and assigned as LOVIs to each fragment. Moreover, the following descriptors for the whole fragment were proposed:

$$\begin{aligned} G_1 &= \sum_i e^{-d_i} \cdot c_i & G'_1 &= \sum_i e^{-d_i} \cdot c'_i \\ G_2 &= \sum_i 2^{-d_i} \cdot c_i & G'_2 &= \sum_i 2^{-d_i} \cdot c'_i \\ G_3 &= \sum_i 10^{1-d_i} \cdot c_i & G'_3 &= \sum_i 10^{1-d_i} \cdot c'_i \\ G_4 &= \sum_i d_i^{-3} \cdot c_i & G'_4 &= \sum_i d_i^{-3} \cdot c'_i \end{aligned}$$

where the summation runs over all the vertices of the fragment, d is the topological distance of the considered atom from the root vertex.

For groups containing heteroatoms or multiple bonds, the LOVIs of each vertex are multiplied by the parameter f_i :

$$f_i = \frac{R_i}{R_{C_{sp^3}}}$$

where R_i is the covalent radius of the i th atom in its hybridization state and $R_{C_{sp^3}}$ is the covalent radius of C_{sp^3} (0.77 Å).

Moreover, for monocyclic substituents, the following descriptor was proposed:

$$G_1 = \sum_i c_i \cdot f_i \cdot \exp(-d_i - d_i / N_{BR})$$

where N_{BR} is the number of ring adjacencies (i.e., six of cyclohexane, nine for benzene). An analogous formula holds for G'_1 on replacing c_i by c'_i .

➤ **extended adjacency ID number** → ID numbers

■ extended adjacency matrices

The extended adjacency matrices **EA** are → *weighted adjacency matrices* $A \times A$ whose elements are defined as a function of → *local vertex invariants* of the → *adjacency matrix* **A** and of some → *atomic properties* [Yang, Xu *et al.*, 1994]. The defined functions aim at removing degeneracy of the entries of the adjacency matrix that is a binary matrix and resemble to some extent the function used in defining the → χ matrix.

The **extended vertex adjacency matrix** (or simply **extended adjacency matrix**) is an adjacency matrix **EA** whose entries are defined as

$$[EA]_{ij} = \begin{cases} a_{ij} \cdot \frac{\delta_i/\delta_j + \delta_j/\delta_i}{2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where a_{ij} are the entries of the adjacency matrix and δ is the → *vertex degree*.

A correction factor can be introduced to account for heteroatoms, such as the → *atomic electronegativity* χ_i ; then, the entries of the **heteroatom-corrected extended adjacency matrix** EA^h are the following:

$$[EA^h]_{ij} = \begin{cases} a_{ij} \cdot \frac{\delta_i^h/\delta_j^h + \delta_j^h/\delta_i^h}{2} & \text{if } i \neq j \\ \chi_i & \text{if } i = j \end{cases}$$

where $\delta_i^h = \delta_i \cdot \chi_i$.

A further extension of the extended adjacency matrix can also be made to consider $\rightarrow$ *bond multiplicity*; the entries of the **heteroatom/multiplicity-corrected extended adjacency matrix**, denoted by $\mathbf{EA}^{hb}$, are the following:

$$[\mathbf{EA}^{hb}]_{ij} = \begin{cases} \frac{1}{2} \cdot \left(\frac{\delta_i^h + 1 - 1/\pi_{ij}^*}{\delta_j^h + 1 - 1/\pi_{ij}^*} + \frac{\delta_j^h + 1 - 1/\pi_{ij}^*}{\delta_i^h + 1 - 1/\pi_{ij}^*} \right) & \text{if } (i,j) \in E(\mathcal{G}) \\ \chi_i & \text{if } i=j \\ 0 & \text{if } (i,j) \notin E(\mathcal{G}) \end{cases}$$

where the increment to the modified vertex degree being due to the multiplicity is

$$\begin{cases} 0.00 & \text{if } \pi^* = 1 \\ 0.33 & \text{if } \pi^* = 1.5 \\ 0.50 & \text{if } \pi^* = 2 \\ 0.67 & \text{if } \pi^* = 3 \end{cases}$$

where π_{ij}^* is the $\rightarrow$ *conventional bond order*.

Note. In the original paper, the heteroatom/multiplicity degrees are defined as the following:

$$\begin{cases} \delta_i^{hb} = \delta_i^h & \text{if } \pi^* = 1 \\ \delta_i^{hb} = \delta_i^h + 1/2 & \text{if } \pi^* = 2 \\ \delta_i^{hb} = \delta_i^h + 1/3 & \text{if } \pi^* = 3 \end{cases}$$

However, by this definition, the ranking of the corrected degrees due to different multiplicities is doubtful, the triple bond increment being intermediate to single and double bonds.

From the extended adjacency matrix defined above two $\rightarrow$ *spectral indices*, called $\rightarrow$ *extended adjacency matrix indices*, were proposed as molecular descriptors. Moreover, the $\rightarrow$ *extended adjacency ID number* was derived from an extended adjacency matrix defined in terms of atomic covalent radii and local vertex invariants computed from $\rightarrow$ *layer matrices*.

The **extended edge adjacency matrix**, denoted as $\mathbf{E}^E\mathbf{A}$, was also defined by analogy with the extended vertex adjacency matrix. This is a symmetric $B \times B$ matrix, B being the number of edges in the graph, where the vertex degrees are replaced by the $\rightarrow$ *edge degrees* ϵ [Janežič, Miličević *et al.*, 2007]. The extended edge adjacency matrix of a molecular graph $\mathcal{G}$ is the extended vertex adjacency matrix of the corresponding $\rightarrow$ *line graph* $L(\mathcal{G})$.

- **extended adjacency matrix indices** $\rightarrow$ spectral indices
- **extended adjacency matrix** $\equiv$ *extended vertex adjacency matrix* $\rightarrow$ extended adjacency matrices
- **Extended Connection Table Representation** $\rightarrow$ molecular graph
- **extended connectivity** $\rightarrow$ canonical numbering ($\odot$ Morgan's extended connectivity algorithm)
- **extended connectivity algorithm** $\equiv$ *Morgan's extended connectivity algorithm* $\rightarrow$ canonical numbering
- **Extended Connectivity FingerPrints** $\rightarrow$ substructure descriptors ($\odot$ fingerprints)
- **extended connectivity indices** $\rightarrow$ canonical numbering ($\odot$ Morgan's extended connectivity algorithm)

- **extended edge adjacency matrix** → extended adjacency matrices
- **extended edge connectivity indices** → edge adjacency matrix
- **extended Ivanciuc–Balaban operator** → Balaban distance connectivity index
- **extended local information on distances** → topological information indices
- **extended Madan degree** → vertex degree
- **Extended Set of Smallest Rings** → ring descriptors
- **extended vertex adjacency matrix** → extended adjacency matrices
- **extended vertex degree** $\equiv$ *extended connectivity* → canonical numbering ($\odot$ Morgan's extended connectivity algorithm)
- **extended Wiener–Hosoya indices** → Wiener index
- **external evaluation set** → data set
- **External Factor Variable Connectivity Indices** → variable descriptors
- **external fragment topological indices** → fragment topological indices
- **external validation** → validation techniques

■ extrathermodynamic approach

Sometimes used as a synonym of → *Hansch analysis*, the extrathermodynamic approach refers to models based on empirical relationships of → *physico-chemical properties* with thermodynamic parameters such as free energies, enthalpies, and entropies for various reactions.

These relationships, based on thermodynamic parameters but not requiring the formal thermodynamic theory, are therefore “extrathermodynamic.”

The Hammett, Taft, and Hansch–Fujita equations defining electronic, steric, and hydrophobic constants are examples of extrathermodynamic relationships, being based on logarithms of rate or equilibrium constants, that is, free energy-related quantities of standard organic reactions of congeneric compound series. Since these correlation equations are often linear with respect to at least one variable, they are called **Linear Free Energy Relationships** (LFER) [Fujita, 1990; Mekenyan and Bonchev, 1986; Wells, 1968b].

Between 1961–1964 Hansch and coworkers used, for the first time, an extrathermodynamic approach to mathematically relate biological activity to the physico-chemical properties of molecules.

The basic assumption is that the introduction of different substituents into a reference compound modifies its biological activity, which can be expressed to a first-order approximation by the following relationship:

$$\Delta K = f(\Delta\Phi_1, \Delta\Phi_2, \dots, \Delta\Phi_j)$$

where Φ are molecular physico-chemical properties, usually electronic, steric, and hydrophobic properties, and K (or k) is the equilibrium (or rate) constant of the biological interaction.

📖 [Abraham, Whiting *et al.*, 1998; Chapman and Shorter, 1978; Charton, 1978a; Drmanić, Jovanović *et al.*, 2000; Gironés and Carbó-Dorca, 2003; Hansch, Quinlan *et al.*, 1968; Jenkins, Samuel *et al.*, 1995; Konovalov, Coomans *et al.*, 2007; Kubinyi, 1993b; Ohlenbusch and Frimmel, 2001; Platts, Abraham *et al.*, 2000; Platts, Butina *et al.*, 1999; Ponec, Gironés *et al.*, 2002; Roberts, 1995; Simón-Manso, 2005; Verloop, 1972; Zissimos, Abraham *et al.*, 2002c]

- **E weighting scheme** → weighting schemes

F

- **false negative rate** → classification parameters
- **false positive rate** → classification parameters
- **farness** → center of a graph
- **FCFC fingerprints** → substructure descriptors (⊙ fingerprints)
- **FCFP fingerprints** ≡ *Functional Connectivity FingerPrints* → substructure descriptors (⊙ fingerprints)
- **feature maps** ≡ *topological feature maps*
- **feature reduction** ≡ *variable reduction*
- **feature tree** → molecular graph
- **Ferreira–Kiralj hydrophobicity parameters** → lipophilicity descriptors
- **FHACA index** → charged partial surface area descriptors
- **FHASA index** ≡ *RSAM index* → charged partial surface area descriptors
- **FHASA₂ index** → charged partial surface area descriptors (⊙ *RSAM index*)
- **FHBCA index** → charged partial surface area descriptors
- **FHBSA index** → charged partial surface area descriptors
- **FHDCA index** → charged partial surface area descriptors
- **FHDSA index** ≡ *RSHM index* → charged partial surface area descriptors
- **Fibonacci numbers** → symmetry descriptors (⊙ Merrifield–Simmons index)
- **Field characterization for Reaction Analysis and Understanding** ≡ *FRAU Features*
- **field effect** → electronic substituent constants
- **field-fitting alignment** → alignment rules
- **field-inductive constant** → electronic substituent constants (⊙ field/resonance effect separation)
- **field-inductive effect** ≡ *polar effect* → electronic substituent constants
- **field/resonance effect separation** → electronic substituent constants
- **¹⁹F inductive constant** → electronic substituent constants (⊙ inductive electronic constants)
- **finite probability scheme** → equivalence classes
- **first eigenvector algorithm** → canonical numbering
- **fingerprints** → substructure descriptors
- **first-order sparse matrix** → algebraic operators (⊙ sparse matrices)
- **first Zagreb index** → Zagreb indices
- **FLAP fingerprints** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **flash point** → physico-chemical properties
- **flexibility index based on path length** → flexibility indices

■ flexibility indices

These are molecular descriptors proposed for the quantification of **molecular flexibility** (or, dually, **molecular rigidity**) and **bond flexibility** (or, dually, **bond rigidity**) [von der Lieth, Stumpf-Nothof *et al.*, 1996].

The concept of molecular flexibility is of primary importance in chemistry since molecular flexibility influences the chemical and biological properties of compounds as well as their interactions with other molecules.

Only a few attempts to quantify this concept can be found in the literature. Molecular dynamics and $\rightarrow$ *grid-based QSAR techniques* have been developed to account for conformational flexibility of compounds [Clark, Willett *et al.*, 1992, 1993; Hahn, 1997; Godha, Mori *et al.*, 2000].

In most cases, a local description of the flexibility is also required to distinguish between flexible and rigid parts of molecules. Therefore, both bond flexibility indices and molecular flexibility quantification were proposed.

Usually, the flexibility of a bond within a molecule is related to its $\rightarrow$ *bond order*, the nature of the atoms incident to the bond, bond participation in one or more cyclic structures, and the branching of adjacent atoms. A single acyclic bond formed by two C_{sp^3} atoms is regarded as freely rotatable, while bonds in polyaromatic ring systems, double and triple bonds are usually regarded as rigid. Analogously, the structural features decreasing molecular flexibility (or increasing molecular rigidity) are few atoms and the presence of rings and branching. It is usually assumed that a completely flexible molecule has an endless chain of C_{sp^3} atoms.

Most of the flexibility indices are derived from the $\rightarrow$ *H-depleted molecular graph*.

The most popular flexibility indices are listed below.

• Rotatable Bond Number (RBN)

This is the number of bonds that allow free rotation around themselves; these bonds are any single bond, not in a ring, bound to a nonterminal heavy atom. In other words, the rotatable bond number is the count of $C_{sp^3}-C_{sp^3}$ and $C_{sp^3}-C_{sp^2}$ bonds in the molecule, often excluding potentially rotatable bonds such as $-OH$ and $-CH_3$ [Bath, Poirrette *et al.*, 1995; Godha, Mori *et al.*, 2000]. It has also been suggested to exclude from the count amide $C-N$ bonds because of their high rotational energy barrier [Veber, Johnson *et al.*, 2002].

As a general systematic rule, all single bonds that satisfy the following criteria are identified as rotatable bonds: (a) the heavy atoms (i.e., nonhydrogen atoms) $A-B$, connected by a single bond, must be connected to a second atom (C, D) as the following $C-A-B-D$. This second atom may be a hydrogen atom. (b) The external bond $C-A$ or $B-D$ must not be a triple bond unless the triple bonded atom is connected to another atom. (c) The bond $A-B$ must not be part of a ring [Munk, Jørgensen and Pedersen, 2001].

Moreover, a general formula for the calculation of the number of rotatable bonds is [Head, Smythe *et al.*, 1996]

$$RBN = N_{nt} + \sum_r (n_r - 4)$$

where N_{nt} is the number of nonterminal freely rotatable bonds, the summation goes over the rings in a molecule, and n_r is the number of single bonds in any nonaromatic ring.

Another formula for the rotatable bond number was proposed by [Oprea, 2000], taking explicitly into account the role of rings in a molecule:

$$RBN = N_{nt} + \sum_r (n_r - 4 - RGB_r - R_B)$$

where N_{nt} is the number of nonterminal freely rotatable bonds (but single bonds observed in groups such as sulfonamides (N–S) or esters (C–O) are excluded); the summation goes over the rings in a molecule, where n_r is the number of single bonds in the r th nonaromatic ring with six or more bonds, RGB_r is the number of rigid bonds in the r th ring, R_B is the number of bonds shared by the r th ring with any other ring, that is, the number of $\rightarrow$ ring bridges.

The complementary quantity to the rotatable bond number is the **rigid bond number**, denoted by RGB , which is defined as the difference between the total number of bonds in a molecule and the total number of rotatable bonds (including terminal single bonds) [Oprea, 2000; Zheng, Luo *et al.*, 2005].

The **rotatable bond fraction**, denoted by RBF , is the fraction of rotatable bonds over the total number of bonds:

$$RBF = \frac{RBN}{B}$$

where B is the total number of bonds in a molecule.

- **flexibility index based on path length (F_K)**

This is a topological index encoding information about molecular flexibility defined as a function of the length L of the longest chain in a molecule and the count ${}^3P$ of paths of length three [Kier and Hall, 1983c]:

$$F_K = \frac{L}{1 - 1/{}^3P}$$

The F_K index increases with increasing chaining and decreases with increasing branching.

F_K is set at zero by definition if no three-bond path is present, that is, for all compounds with ${}^3P = 0$.

- **Kier molecular flexibility index (Φ)**

This is a measure of molecular flexibility derived from the $\rightarrow$ Kier alpha-modified shape descriptors ${}^1\kappa_\alpha$ and ${}^2\kappa_\alpha$:

$$\Phi = \frac{{}^1\kappa_\alpha \cdot {}^2\kappa_\alpha}{A}$$

where A is the total number of atoms in a molecule. The Kier shape indices calculated from the $\rightarrow$ H-depleted molecular graph depend on the heteroatoms by the parameter α [Kier, 1989; Kier and Hall, 1999d]; ${}^1\kappa_\alpha$ encodes information about the count of atoms and relative cyclicity of molecules, whereas ${}^2\kappa_\alpha$ encodes information about branching or relative spatial density of molecules. The atom count A allows comparisons among isomers.

- **global flexibility index (GS)**

This is a measure of molecular flexibility derived from additive contributions of path flexibilities as [von der Lieth, Stumpf-Nothof *et al.*, 1996]

$$GS = \frac{2}{A \cdot (A-1)} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A LS_{ij}$$

where A is the number of vertices in the H-depleted molecular graph. LS_{ij} is the **local simple flexibility index** relative to the path connecting vertices i and j , defined as

$$LS_{ij} = (d_{ij} + 1) - \left(\frac{NRB + 0.75 \cdot {}^4F + 0.50 \cdot {}^3F}{2} \right)_{ij}$$

where d_{ij} is the $\rightarrow$ *topological distance* between vertices i and j , which is the length of the shortest path between the two vertices; NRB is the number of nonrotatable bonds in the path p_{ij} ; ${}^4F$ and ${}^3F$ are the number of branching atoms with $\rightarrow$ *vertex degree* equal to four and three, that is, with four and three adjacent vertices, respectively, in the path p_{ij} .

An acyclic nonbranched chain of C_{sp^3} atoms is regarded to as completely flexible, and LS simply equals the number of atoms involved in the considered path.

• bond flexibility index (Φ_{BD})

This is an index encoding information about the flexibility of a bond and calculated as the mean of the atom flexibility indices Φ^a of the two atoms forming the bond [von der Lieth, Stumpf-Nothof *et al.*, 1996]:

$$\Phi_{BD} = \frac{\Phi_i^a + \Phi_j^a}{2}$$

where i and j are the atoms incident to the considered bond. Φ_{BD} is between 0 (not flexible) and 10 (completely flexible).

Flexibility indices Φ^a are defined for each molecule atom, provided it belongs to one of the defined molecular substructures: aromatic rings, multiple bonds, conjugated systems, simple rings, chains, and bridges. For all the atoms belonging to double bonds, triple bonds, or aromatic rings, $\Phi^a = 0.5$. For all atoms belonging to simple rings, Φ^a is equal to the $\rightarrow$ *Kier molecular flexibility index* Φ ; for atoms in condensed ring systems, Kier molecular flexibility index Φ is calculated for both the simple rings and the condensed system, the smallest value is taken as the Φ^a value for each atom in the ring. Moreover, for each substitution group in the ring or condensed ring, a value of 0.2 is subtracted from the Φ^a value. The C_{sp^3} atoms in the chains and bridges are assigned a $\Phi^a = 10$ and this value is corrected by subtracting the sum of the atom masses of all the atoms in the next four adjacent shells, divided by 100.

The **bond rigidity index** ρ_{BD} , obtained from the bond flexibility index Φ_{BD} in the same range but with opposite meaning, is defined as

$$\rho_{BD} = 10 - \Phi_{BD}$$

• Kier bond rigidity index (ρ_{KB})

This is a measure of the rigidity/flexibility of a bond within a molecule, derived from the $\rightarrow$ *Kier molecular flexibility index* Φ . This index is defined as [von der Lieth, Stumpf-Nothof *et al.*, 1996]:

$$\rho_{KB} = \left(\sum_k \Phi_k^f \right) - \Phi + 1$$

where Φ^f is the Kier flexibility index calculated for a single fragment. The summation goes over all molecule fragments obtained by breaking the bond of interest. It is a measure of bond rigidity because it represents the increase in flexibility of the fragments with respect to the parent molecule; this difference increases as the rigidity of the broken bond increases.

• **molecular flexibility number (ϕ)**

This is an index of molecule flexibility defined as [Dannenfelser and Yalkowsky, 1996; Jain, Yang *et al.*, 2004a, 2004b]

$$\phi = 2.85^\tau$$

where the value 2.85 is proportional to the difference in energy (kJ/mol) between trans and gauche conformations and τ is the number of torsional angles. The most common expression to derive the number of torsional angles τ is

$$\tau = N_{sp^3} + 0.5 \cdot N_{sp^2} + 0.5 \cdot NRG - 1$$

where N_{sp^3} , N_{sp^2} , and NRG are the number of sp^3 -hybridized chain atoms, sp^2 -hybridized chain atoms, and the number of fused ring systems, respectively.

📖 [Luisi, 1977; Fisanick, Cross *et al.*, 1993; Bradbury, Mekenyan *et al.*, 1996; Bayada, Hemersma *et al.*, 1999; Oprea and Gottfries, 2001a; Martinek, Ötvös *et al.*, 2005]

- **flexible descriptors** $\equiv$ *variable descriptors*
- **Flexsim-R fingerprints** $\rightarrow$ affinity fingerprints
- **Flexsim-S fingerprints** $\rightarrow$ affinity fingerprints
- **Flexsim-X fingerprints** $\rightarrow$ affinity fingerprints
- **F matrix** $\rightarrow$ layer matrices ($\odot$ cardinality layer matrix)
- **F-measure** $\rightarrow$ classification parameters
- **FM method** $\equiv$ *Dewar–Grisdale approach* $\rightarrow$ electronic substituent constants ($\odot$ field/resonance effect separation)
- **FMMF method** $\equiv$ *Dewar–Golden–Harris approach* $\rightarrow$ electronic substituent constants ($\odot$ field/resonance effect separation)
- **folding degree index** $\rightarrow$ spectral indices
- **folding profile** $\rightarrow$ spectral indices ($\odot$ folding degree index)
- **Forbes–Mozley similarity coefficient** $\rightarrow$ similarity/diversity ($\odot$ Table S9)
- **forest** $\rightarrow$ graph
- **formal degree** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **formal oxidation number** $\rightarrow$ multiple bond descriptors
- **forward Fukui function** $\rightarrow$ quantum-chemical descriptors ($\odot$ Fukui functions)
- **Fossum similarity coefficient** $\rightarrow$ similarity/diversity ($\odot$ Table S9)
- **Fourier analysis** $\rightarrow$ spectra descriptors

■ **fractals**

Fractals are geometric structures of fractional dimension; their theoretical fundamentals and physical applications were studied by Mandelbrot [Mandelbrot, 1982]. By definition, any structure possessing a self-similarity or a repeating motif invariant under a transformation of scale is called *fractal* and may be represented by a fractal dimension. Mathematically, the fractal dimension D_f of a set is defined through the relation

$$N(\epsilon) \propto \epsilon^{-D_f}$$

where $N(\epsilon)$ denotes the number of spheres of radius ϵ needed to cover the whole set. The fractal dimension of a set can be interpreted as the amount of information needed to fully specify a point of the set.

The concept of fractal dimension has been applied, for example, to represent tertiary structure of protein surface [Isogai and Itoh, 1984; Wagner, Colvin *et al.*, 1985; Åqvist and Tapia, 1987; Wang, Shi *et al.*, 1990; Poirrette, Artymiuk *et al.*, 1997; Tominaga and Fujiwara, 1997a; Tominaga, 1998a; Torrens, 2002], flexibility of alkanes [Rouvray and Pandey, 1986; Rouvray and Kumazaki, 1991], molecular shape [Mayer, Farin *et al.*, 1986], chromatographic profiles [Yiyu, Minjun *et al.*, 2003]; moreover fractals have been used in data structure comparison [Tominaga and Fujiwara, 1997a; Tominaga, 1998a] and in $\rightarrow$ *cell-based methods* [Agrafiotis and Rassokhin, 2002].

- **fractional bond order** $\rightarrow$ bond order indices
- **fractional charged partial negative surface areas** $\rightarrow$ charged partial surface area descriptors
- **fractional charged partial positive surface areas** $\rightarrow$ charged partial surface area descriptors
- **fragmental adjacency matrix** $\rightarrow$ adjacency matrix
- **fragmental connectivity index** $\rightarrow$ adjacency matrix
- **fragmental constants** $\equiv$ *group contributions* $\rightarrow$ group contribution methods
- **fragmental degree** $\rightarrow$ adjacency matrix
- **fragment-based descriptors** $\equiv$ *substructure descriptors*
- **fragment count** $\rightarrow$ count descriptors
- **fragment ID numbers** $\rightarrow$ ID numbers
- **Fragment Molecular Connectivity indices** $\rightarrow$ connectivity indices

■ fragment topological indices (FTI)

Derived from $\rightarrow$ *topological indices* calculated by graph dissections, fragment topological indices were proposed to reflect the interactions between the excised fragment and the remainder of the molecule [Mekenyan, Bonchev *et al.*, 1988a].

Let G be a $\rightarrow$ *H-depleted molecular graph* with A vertices and G' a subgraph (i.e., a fragment F) of G with A' vertices, with $A' < A$ by definition. Topological indices (TI) that consider only vertices and edges belonging to the subgraph are called **internal fragment topological indices** and denoted by IFTI. The two following requirements were proposed for IFTI:

$$0 \leq \text{IFTI}(F) < \text{TI}(G) \quad \text{and} \quad 0 \leq \text{IFTI}(G-F) < \text{TI}(G)$$

where TI denotes one of the common defined topological indices, restricted to those increasing with the increase in the number of graph vertices. $\text{IFTI}(G-F)$ is the topological index calculated on the complementary subgraph.

Indices that describe a fragment in connection with the remainder of the graph are called **external fragment topological indices** and are denoted by EFTI. They were proposed as the difference in value between the topological index $\text{TI}(G)$ for the whole graph and the internal fragment indices for both the fragment $\text{IFTI}(F)$ and the remainder of the molecule $\text{IFTI}(G-F)$:

$$\text{EFTI}(F) = \text{TI}(G) - \left[\text{IFTI}(F) - \sum_k \text{IFTI}(G-F)_k \right]$$

where the summation goes over all the $G-F$ disconnected components; there will be only one component if the subgraph $G-F$ is a connected graph. The following requirement was

proposed for EFTI:

$$\text{EFTI}(F) < \text{TI}(G)$$

The **normalized fragment topological indices** (NIFTI and NEFTI) may be obtained by dividing IFTI and EFTI by the topological index for the whole graph as

$$\text{NIFTI}(F) = \frac{\text{IFTI}(F)}{\text{TI}(G)} \quad \text{and} \quad \text{NEFTI}(F) = \frac{\text{EFTI}(F)}{\text{TI}(G)}$$

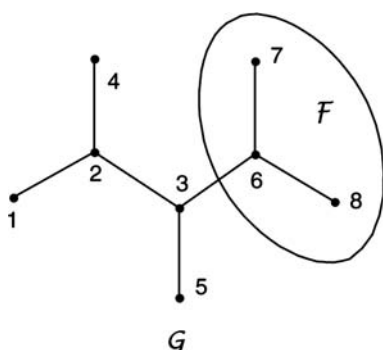
A special case of normalized fragment topological indices NIFTI is the $\rightarrow$ *graphical bond order*.

It has to be noted that $\text{IFTI}(F)$ is a constant for a given fragment of any molecule, whereas $\text{NIFTI}(F)$, $\text{EFTI}(F)$ and $\text{NEFTI}(F)$ depend upon the molecule as a whole. Moreover, the following general relation holds:

$$\text{TI}(G_1) > \text{TI}(G_2) \rightarrow \text{EFTI}(F \subset G_1) > \text{EFTI}(F \subset G_2)$$

Example F1

Calculation of internal and external first Zagreb index. The molecular fragment consists of vertices 6, 7, and 8, and **A** is the adjacency matrix.



A =

Atom	1	2	3	4	5	6	7	8
1	0	1	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0
3	0	1	0	0	1	1	0	0
4	0	1	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0
6	0	0	1	0	0	0	1	1
7	0	0	0	0	0	1	0	0
8	0	0	0	0	0	1	0	0

vertex degrees δ

Atom	G	F	$G-F$
1	1	—	1
2	3	—	3
3	3	—	2
4	1	—	1
5	1	—	1
6	3	2	—
7	1	1	—
8	1	1	—

$$\text{first Zagreb index : } M_1 = \sum_{i=1}^A \delta_i^2$$

$$M_1(G) = 5 \times 1^2 + 3 \times 3^2 = 32$$

$$\text{IFM}_1(F) = 2^2 + 2 \times 1^2 = 6$$

$$\text{IFM}_1(G-F) = 3 \times 1^2 + 2^2 + 3^2 = 16$$

$$\begin{aligned} \text{EFM}_1(F) &= M_1(G) - \text{IFM}_1(F) - \text{IFM}_1(G-F) \\ &= 32 - 6 - 16 = 10 \end{aligned}$$

➤ **F-ratio test** $\rightarrow$ regression parameters

■ FRAU Features (FF)

FRAU (Field characterization for Reaction Analysis and Understanding) features encode information about the reaction field of the atoms in a molecule in the three-dimensional space [Satoh, Itono *et al.*, 1999; Satoh, 2007]. These descriptors are based on steric and electrostatic interactions determined by a pseudoreactant.

To calculate FRAU descriptors, first the frontier surface of each atom is determined by drawing a sphere around each atom with radius equal to or larger than the atomic van der Waals radius. The surface of this sphere is taken as the atomic surface. The part of the atomic surface overlapping with the sphere of another atom is called *interior surface*, and the atomic surface minus the interior surface is called *frontier surface*. Then, a number of points are evenly distributed on the atomic surface.

Only points on the frontier surface are used to evaluate features of the molecule by means of a probe. There are three kinds of FRAU descriptors: the *extent of reaction field* (FF_i^{field}), the *electrostatic feature* ($FF_i^{electro}$), and the *steric feature* (FF_i^{steric}).

For each *i*th atom in the molecule, the *extent of reaction field* FF_i^{field} is defined as the number of points on the surface frontier of the atom.

To evaluate the *electrostatic feature*, a unit charge is taken as the probe placed at every point on the atomic frontier surface. Then, $FF_i^{electro}$ (in kcal/mol) is calculated as

$$FF_i^{electro} = \frac{1}{FF_i^{field}} \cdot \sum_{k=1}^{FF_i^{field}} \sum_{j=1}^A \frac{331.8417 \cdot q_j}{r_{kj}}$$

where *A* is the number of atoms in the molecule, q_j the net charge of the *j*th atom, and r_{kj} the geometric distance between the *k*th surface point and the *j*th atom.

Estimation of the *steric feature* FF_i^{steric} is based on the van der Waals interaction between an atom as the probe (e.g., sp^3 carbon atom) placed at the frontier surface points and every atom in the molecule. The FF_i^{steric} (in kcal/mol) of the *i*th atom is calculated by the MM3 force-field equation as

$$FF_i^{steric} = \frac{1}{FF_i^{field}} \cdot \sum_{k=1}^{FF_i^{field}} \sum_{j=1}^A \left(\sqrt{\eta_i \cdot \eta_j} \right) \\ \times \left[1.84 \times 10^5 \cdot \exp \left(-12.0 \cdot \left(\frac{R_i^{vdw} + R_j^{vdw}}{r_{kj}} \right)^{-1} \right) - 2.25 \cdot \left(\frac{R_i^{vdw} + R_j^{vdw}}{r_{kj}} \right)^6 \right]$$

where *A* is the number of atoms in the molecule, the first summation goes over all the points on the surface frontier of the atom and the second summation on all the atoms of the molecule; η indicates the atomic $\rightarrow$ *hardness*, R^{vdw} the atomic van der Waals radius, and r_{kj} the geometric distance between the probe atom place at the *k*th surface point and any *j*th atom of the molecule.

Finally, to evaluate similarities and differences among the FRAU features of atoms, these are projected on a 2D map by means of the $\rightarrow$ *Self-Organizing Map* approach.

Unlike the common $\rightarrow$ *grid-based QSAR techniques*, FRAU expresses features of molecules having different size and substructures since they do not require alignment of molecules.

FRAU features were applied to classify and predict reagent functions and to distinguish 3D stereochemical environments of atoms.

- **free energy of hydration density tensor** → hydration free energy density
- **free molecular volume** → volume descriptors (⊙ molar volume)
- **free valence index** → quantum-chemical descriptors

■ Free-Wilson analysis ($\equiv$ FW Analysis)

Free-Wilson analysis is a QSAR approach searching for a relationship between a biological response and the presence/absence of substituent groups on a common molecular skeleton [Free and Wilson, 1964; Kubinyi, 1990, 1993b]. The approach, called *de novo approach* when first presented in 1964, is based on the assumption that each substituent gives an additive and constant effect to the biological activity regardless of the other substituents in the rest of the molecule, that is, the substituent effects are considered to be independent of each other. Compounds → *congenericity* is also another basic requirement.

Once a common skeleton for the chemical analogues is defined, regression analysis is performed, considering a number S of substitution sites, denoted as R_s ($s = 1, S$), and, for each site, a number N_s of different substituents. Hydrogen atoms are also considered as substituents if present in a substitution site of some compounds.

The **Free-Wilson descriptors** of the i th compound are → *indicator variables*, denoted by $I_{i,ks}$, where $I_{i,ks} = 1$ if the k th substituent is present in the s th site of the i th molecule, and $I_{i,ks} = 0$ otherwise. These descriptors are usually collected in a table called the **Free-Wilson matrix**, denoted as **FW**, where the rows represent the data set molecules and each column represents a substituent in a specific site (Example F2).

The total number of indicator descriptors (i.e., the Free-Wilson matrix columns) is

$$p = \sum_{s=1}^S N_s$$

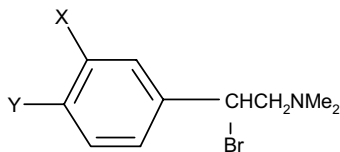
where S is the number of substitution sites and N_s the number of substituents per site.

Given the number of substitution sites S and the number of substituents for each site N_s , the maximum number of structurally diverse compounds that can be studied by a Free-Wilson model is

$$n^{\max} = \prod_{s=1}^S N_s$$

Example F2

Parent molecule and Free-Wilson matrix for 22 derivatives of *N,N*-dimethyl- α -bromophenethylamines; X and Y indicate two substitution sites.



In both X and Y sites, hydrogen, fluorine, chlorine, bromine, iodine, and methyl substituents are allowed ($N_X = N_Y = 6$). The 22 studied derivatives are coded as in the Free-Wilson matrix below. The first row of the Free-Wilson matrix is the H-substituted phenethylamine, used as the reference compound in the Fujita-Ban model and usually excluded in the original Free-Wilson approach.

ID	X	Y	H	m-F	m-Cl	m-Br	m-I	m-Me	H	p-F	p-Cl	p-Br	p-I	p-Me
1	H	H	1	0	0	0	0	0	1	0	0	0	0	0
2	H	F	1	0	0	0	0	0	0	1	0	0	0	0
3	H	Cl	1	0	0	0	0	0	0	0	1	0	0	0
4	H	Br	1	0	0	0	0	0	0	0	0	1	0	0
5	H	I	1	0	0	0	0	0	0	0	0	0	1	0
6	H	Me	1	0	0	0	0	0	0	0	0	0	0	1
7	F	H	0	0	0	0	0	0	1	0	0	0	0	0
8	Cl	H	0	0	0	0	0	0	1	0	0	0	0	0
9	Br	H	0	0	0	1	0	0	1	0	0	0	0	0
10	I	H	0	0	0	0	1	0	1	0	0	0	0	0
11	Me	H	0	0	0	0	0	1	1	0	0	0	0	0
12	Cl	F	0	0	1	0	0	0	0	1	0	0	0	0
13	Br	F	0	0	0	1	0	0	0	1	0	0	0	0
14	Me	F	0	0	0	0	0	1	0	1	0	0	0	0
15	Cl	Cl	0	0	0	0	0	0	0	0	1	0	0	0
16	Br	Cl	0	0	0	1	0	0	0	0	1	0	0	0
17	Me	Cl	0	0	0	0	0	1	0	0	1	0	0	0
18	Cl	Br	0	0	1	0	0	0	0	0	0	1	0	0
19	Br	Br	0	0	0	1	0	0	0	0	0	1	0	0
20	Me	Br	0	0	0	0	0	1	0	0	0	1	0	0
21	Me	Me	0	0	0	0	0	1	0	0	0	0	0	1
22	Br	Me	0	0		1	0	0	0	0	0	0	0	1

$$p = \sum_{s=1}^2 N_s = 6 + 6 = 12 \qquad n^{\max} = \prod_{s=1}^2 N_s = 6 \times 6 = 36$$

The **Free-Wilson model** (also called **additivity model**) is defined as

$$\hat{y}_i = b_0 + \sum_{s=1}^S \sum_{k=1}^{N_s} b_{ks} \cdot I_{i,ks}$$

where b_0 is the intercept of the model corresponding to the theoretical biological activity of a compound without any substituent and b_{ks} are the regression coefficients. The biological response y is usually used in the form $\log(1/C)$, where C is the concentration achieving a fixed effect. The regression coefficients b_{ks} of the Free-Wilson model give the importance of each k th substituent in each s th site in increasing/decreasing the response with respect to the unsubstituted compound.

Fujita-Ban analysis is a **modified Free-Wilson analysis** that accounts for the activity contribution of each substituent relatively to the activity of a $\rightarrow$ *reference compound* [Fujita and Ban, 1971]. Any compound can be chosen as the reference, but usually the H-substituted compound is adopted: the row vector corresponding to the reference compound is characterized by all the descriptor binary values equal to zero. The Free-Wilson matrix in the Fujita-Ban analysis does not contain the descriptors corresponding to the substituents of the reference compound.

The **Fujita-Ban model** is defined as

$$\hat{y}_i = b_0 + \sum_{s=1}^S \sum_{k=1}^{N_s} b_{ks} \cdot I_{i,ks}$$

which differs from the Free-Wilson model because the intercept b_0 corresponds to the estimated biological activity of the reference compound, that is, $b_0 = \hat{Y}_{REF}$, whereas in the Free-Wilson model it corresponds to the theoretical biological activity of a “naked” compound, that is, without any substituent.

The Fujita-Ban model is a linear transformation of the classical Free-Wilson model: indeed, group contributions of the Free-Wilson model can be transformed into Fujita-Ban group contributions by subtracting the group contributions of the corresponding substituents of the reference compound.

The **Cammarata-Yau analysis** is similar to the Fujita-Ban analysis, the only difference being that the intercept is maintained constant and equal to the response of the reference compound [Cammarata and Yau, 1970]; this means that the observed responses are first transformed as

$$Y'_i = Y_i - Y_{REF}$$

where Y_{REF} is the response for the reference compound.

Accordingly, the **Cammarata-Yau model** is a regression through the origin defined as

$$\hat{Y}_i = \sum_{s=1}^S \sum_{k=1}^{N_s} b_{ks} \cdot I_{i,ks}$$

The **Bocek-Kopecky analysis** is another modified Free-Wilson approach proposed to take into account interaction terms, that is, nonlinear effects [Bocek, Kopecky *et al.*, 1964; Kopecky, Bocek *et al.*, 1965]. The **Bocek-Kopecky model** is defined as

$$\hat{Y}_i = b_0 + \sum_{s=1}^S \sum_{k=1}^{N_s} b_{ks} \cdot I_{i,ks} + \sum_{s=1}^{S-1} \sum_{s'=s+1}^S \sum_{k=1}^{N_s} \sum_{k'=1}^{N_{s'}} b_{kk',ss'} \cdot I_{i,ks} \cdot I_{i,k's'}$$

The Fujita-Ban model having the hydrogen substituted compound as the reference compound is related to the $\rightarrow$ *Hansch linear model* by the following relationship:

$$b_{ks} \approx \sum_{j=1}^J b_j \cdot \phi_{ks,j}$$

where b_j are the Hansch regression coefficients, J the number of considered substituent properties (e.g., lipophilic, electronic, and steric properties), and $\phi_{ks,j}$ the j th substituent group constant for the k th substituent in the s th site. This relationship means that the group contribution b_{ks} in the Fujita-Ban model of the k th substituent in the s th site is numerically equivalent to the weighted sum of all the $\rightarrow$ *physico-chemical properties* of that substituent [Singer and Purcell, 1967; Kubinyi and Kehrhaan, 1976; Kubinyi, 1988b].

A great advantage of these approaches is the possibility of a complete $\rightarrow$ *reversible decoding*, that is, the possibility to interpret by the model *how* and *where* the response is increased/decreased.

The main shortcomings of the Free-Wilson related approaches are that (1) structural variation is necessary in at least two different sites; (2) a relatively large number of variables is necessary to describe a relatively small number of compounds; (3) the models can be used to predict a maximum number of compounds equal to $n^{\max} - n$, where n is the number of compounds effectively used in the model; and (4) predictions of substituents not included in the analysis are usually not reliable.

Related to the Free–Wilson analysis is the $\rightarrow$ *DARC/PELCO analysis*, which is an extension of the former to the $\rightarrow$ *hyperstructure* concept [Duperray, Chastrette *et al.*, 1976a].

Moreover, molecular descriptors different from Free–Wilson descriptors were calculated by transformation of the Free–Wilson matrix through $\rightarrow$ *Fourier analysis* [Holik and Halamek, 2002]. In this case, Fourier analysis is used to change site- and substituent-oriented binary variables into a few real numbers [Holik and Halamek, 2002].

To calculate Fourier coefficients, each row of the Free–Wilson matrix, denoted by $\mathbf{FW}(n, p)$, where n is the number of molecules and p the number of site/substituent indicator variables, is transformed into cosine and sine terms according to the following equation:

$$F_{ix}(k) = \sum_{j=1}^p [\mathbf{FW}]_{ij} \cdot \cos(\phi_{kj}) \quad \text{and} \quad F_{iy}(k) = \sum_{j=1}^p [\mathbf{FW}]_{ij} \cdot \sin(\phi_{kj})$$

where F_{ix} and F_{iy} are the two Fourier coefficients of the i th molecule, $[\mathbf{FW}]_{ij}$ indicates the elements of the i th row of the Free–Wilson matrix, and ϕ_{kj} is defined as

$$\phi_{kj} = \frac{2 \cdot \pi \cdot k \cdot (j-1)}{p} \quad k = 1, 2, \dots, L$$

where p is the number of matrix columns and L a user-defined integer parameter. Depending on the size p of the matrix $\mathbf{FW}(n, p)$, a different number of linearly independent Fourier coefficients are obtained: if p is odd, then this number is $(p-1)/2$, if p is even, then there are $(p-2)/2$ independent Fourier coefficients. A data compression is performed if L is chosen to be smaller than the number of linearly independent Fourier coefficients.

Example F3

Calculation of the Fourier coefficients for the second row (compound with $R_1 = \text{H}$ and $R_2 = \text{F}$) of the data set comprised of 22 *N,N*-dimethyl- α -bromo-phenethylamines (Appendix C). The corresponding Free–Wilson matrix ($n=22$ and $p=12$) is reported in Example F2.

$$\mathbf{x} = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]$$

For $k=1$ and $j=1, 2, \dots, 12$, the ϕ terms are

$$[0 \ 0.524 \ 1.047 \ 1.571 \ 2.094 \ 2.618 \ 3.142 \ 3.665 \ 4.189 \ 4.712 \ 5.236 \ 5.760]$$

and the first coefficient F_{2x} for $k=1$ is

$$\begin{aligned} F_{2x}(1) &= 1 \times \cos(0) + 0 \times \cos(0.524) + \dots + 0 \\ &\times \cos(2.618) + 1 \times \cos(3.143) + \dots + 0 \times \cos(5.760) = 0.134 \end{aligned}$$

Then k is increased (until $k=5$) and the other cosine Fourier coefficients are calculated. Finally, the procedure is repeated with sine function. As a result, the 12 original variables are transformed into a vector $\mathbf{f}$ of 10 real-valued variables:

$$\mathbf{f} = [0.134 \ 1.500 \ 1.000 \ 0.500 \ 1.866 \ -0.500 \ 0.866 \ -1.000 \ 0.866 \ -0.500]$$

📖 [Craig, 1972; Cammarata, 1972; Cammarata and Bustard, 1974; Thomas, Berkoff *et al.*, 1975; Kubinyi, 1976a; Hall and Kier, 1978a; Schaad, Hess Jr. *et al.*, 1981; Duewer, 1990; Liwo, Tarnowska *et al.*, 1992; Franke and Buschauer, 1992; Simmons, Dixson *et al.*, 1992; Franke and Buschauer, 1993; Henrie II, Plummer *et al.*, 1993; Norinder, 1993; Singh, Ojha *et al.*, 1993; De Castro and Reissmann, 1995; Hasegawa, Shigyou *et al.*, 1995; Timofei, Kurunczi *et al.*, 1995; Hatrik and Zahradnik, 1996; Hasegawa, Yokoo *et al.*, 1996; Fleischer, Froberg *et al.*, 2000; Tomić, Nilsson *et al.*, 2000; Shi, Qian *et al.*, 2001; Waisser, 2001; Holik and Halamek, 2002; Pimple, Kelkar *et al.*, 2004; Prabhakar, Gupta *et al.*, 2005; Saxty, Woodhead *et al.*, 2007]

- **Free–Wilson descriptors** → Free–Wilson analysis
- **Free–Wilson matrix** → Free–Wilson analysis
- **Free–Wilson model** → Free–Wilson analysis
- **freezing point** → physico-chemical properties (⊙ melting point)
- **Friedman's lack-of-fit function** → variable selection (⊙ Genetic Function Approximation)
- **frontier orbitals** → quantum-chemical descriptors
- **frontier orbital electron densities** → quantum-chemical descriptors
- **F strain** $\equiv$ *substituent front strain* → steric descriptors
- **fugacity** → physico-chemical properties
- **Fujita–Ban analysis** → Free–Wilson analysis
- **Fujita–Ban model** → Free–Wilson analysis
- **Fujita steric constant** → steric descriptors (⊙ Taft steric constant)
- **Fukui functions** → quantum-chemical descriptors
- **Functional Connectivity FingerPrints** → substructure descriptors (⊙ fingerprints)
- **functional group count** → count descriptors
- **functional group filters** → property filters

■ functional coordination index (I_C)

Based on the idea to model the morphology–functionality relationship in organisms, the functional coordination index is defined in terms of the → *molecular surface area* SA , the molecular weight MW , and the standard enthalpy of formation H_f as [Torrens, 2003a]

$$I_C = \frac{H_f}{I_m} = \frac{MW \cdot H_f}{SA}$$

where I_m is the **morphologic index** defined as

$$I_m = \frac{SA}{MW}$$

- **functionality index** → ETA indices
- **FW analysis** $\equiv$ *Free–Wilson analysis*

G

- **Galvez matrix** → topological charge indices
- **ganglia-augmented atom keys** → substructure descriptors
- **GAO descriptors** → Graph of Atomic Orbitals
- **GAO** ≡ *Graph of Atomic Orbitals*
- **GAP** ≡ *HOMO–LUMO energy gap* → quantum-chemical descriptors
- **gas–solvent partition coefficient** → physico-chemical properties (⊙ partition coefficients)
- **Geary coefficient** → autocorrelation descriptors
- **general a_N -index** → determinant-based descriptors
- **general distance–degree matrix** → distance–degree matrices
- **general free valence index** → quantum-chemical descriptors
- **general graph** → graph
- **General Interaction Properties Function approach** ≡ *GIPF approach*
- **generalized average graph energy** → spectral indices
- **generalized centric information indices** → centric indices
- **generalized cluster significance analysis** → variable selection (⊙ cluster significance analysis)
- **generalized complete centric index** → centric indices
- **generalized connectivity indices** → connectivity indices
- **generalized distance code centric index** → centric indices
- **generalized distance–degree centric index** → centric indices
- **generalized distance matrices** → distance matrix
- **generalized edge complete centric index** → centric indices
- **generalized edge distance code centric index** → centric indices
- **generalized edge distance degree centric index** → centric indices
- **generalized edge radial centric information index** → centric indices
- **generalized expanded Wiener numbers** → expanded distance matrices
- **generalized final prediction error criteria** → regression parameters
- **generalized graph center** → center of a graph
- **generalized graph energy** → spectral indices
- **generalized Hosoya indices** → Hosoya Z index
- **generalized Hosoya Z matrix** → Hosoya Z matrix
- **generalized hyper-Wiener indices** → Wiener matrix
- **generalized matrices** → matrices of molecules
- **generalized molecular-graph matrix** → variable descriptors

- **generalized radial centric information index** → centric indices
- **generalized reciprocal distance sum** → distance matrix
- **generalized reciprocal matrices** → matrices of molecules
- **Generalized Topological Distance Indices** → distance matrix
- **generalized topological indices** → variable descriptors
- **generalized vertex degree matrix** → vertex degree
- **generalized Wiener indices** → Wiener index
- **generalized Wiener matrix** → Wiener matrix
- **general solubility equation** → property filters (⊙ drug-like indices)
- **Generating Optimal Linear PLS Estimations** $\equiv$ *GOLPE* → variable selection
- **genetic algorithm – variable subset selection** → variable selection
- **genetic function approximation** → variable selection
- **geodesic** → graph
- **geodesic matrix** → algebraic operators (⊙ sparse matrices)

■ geometrical descriptors

These are molecular descriptors defined in several different ways but always derived from the three-dimensional structure of the molecule [Ivanciuc, 2001a; Todeschini and Consonni, 2003]. In general, geometrical descriptors are calculated either from some optimized → *molecular geometry* obtained by the methods of the → *computational chemistry* or from crystallographic coordinates.

→ *Topographic indices* constitute a special subset of geometrical descriptors, being calculated on the graph representation of molecules but using the geometric distances between atoms instead of the topological distances.

Examples of geometrical descriptors (Figure G1) are the → *quantum-chemical descriptors*, → *moments of inertia*, → *length-to-breadth ratio*, → *surface areas*, → *volume descriptors*, → *CPSA descriptors*, → *EVA descriptors*, → *WHIM descriptors*, → *GETAWAY descriptors*, → *3D-MoRSE descriptors*, → *interaction energy values*, and → *spectrum-like descriptors*.

📖 [Mihalić and Trinajstić, 1991; Tvaruzek and Komenda, 1991; Zhu and Klein, 1996; Basak, Gute *et al.*, 1997; Todeschini and Consonni, 2000]

- **geometrical eccentricity** → shape descriptors
- **geometrical representation** → molecular descriptors
- **geometrical shape coefficient** → shape descriptors (⊙ Petitjean shape indices)
- **geometric atom pair descriptors** → substructure descriptors
- **geometric binding property pairs** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **geometric center** → center of a molecule
- **geometric diameter** → molecular geometry
- **geometric distance** → molecular geometry
- **geometric distance degree** → molecular geometry
- **geometric distance–detour distance combined matrix** → matrices of molecules (⊙ Table M3)
- **geometric distance/detour distance quotient matrix** → matrices of molecules (⊙ Table M2)
- **geometric distance matrix** $\equiv$ *geometry matrix* → molecular geometry

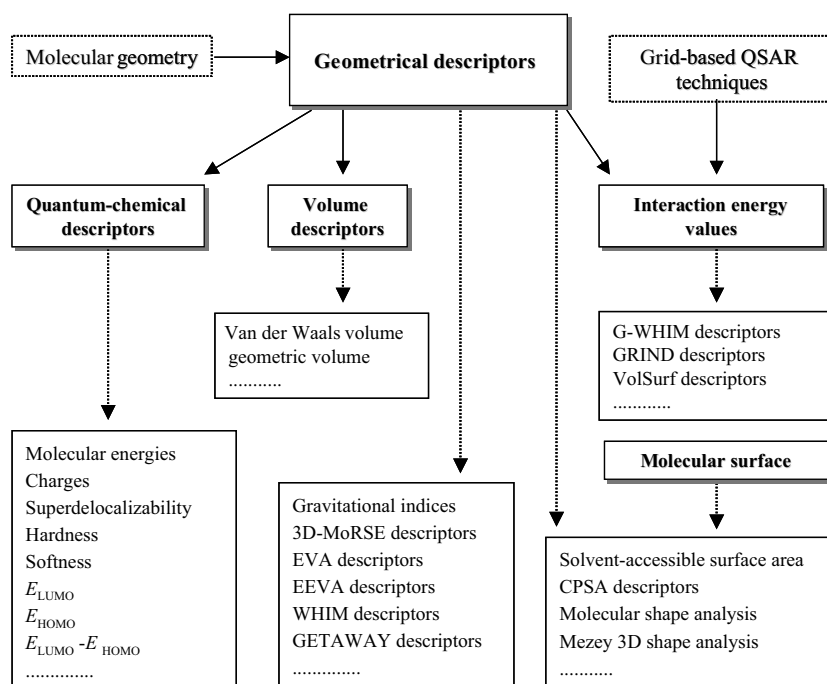


Figure G1 Scheme of the molecular descriptors classified as geometrical descriptors.

- **geometric distance–resistance distance combined matrix** → matrices of molecules (⊙ Table M3)
- **geometric distance/resistance distance quotient matrix** → matrices of molecules (⊙ Table M2)
- **geometric distance–topological distance combined matrix** → molecular geometry
- **geometric distance/topological distance quotient matrix** → molecular geometry
- **geometric eccentricity** → molecular geometry
- **geometric edge distance matrix** → edge distance matrix
- **geometric factors** → weighted matrices (⊙ weighted distance matrices)
- **geometric mean** → statistical indices (⊙ indices of central tendency)
- **geometric mean of the leverage magnitude** → GETAWAY descriptors
- **Geometric Mean Polarizability Effect Index** → electric polarization descriptors (⊙ Polarizability Effect Index)
- **Geometric Mean Polarizability Effect Index of π bond** → electric polarization descriptors (⊙ Polarizability Effect Index)
- **geometric modification number** → weighted matrices (⊙ weighted distance matrices)
- **geometric radius** → molecular geometry
- **geometric sum layer matrix** → layer matrices
- **geometric topological index** → vertex degree
- **geometric volume** → volume descriptors
- **geometry matrix** → molecular geometry
- **George-Foster criterion** → regression parameters (⊙ Table R1)

■ GETAWAY descriptors

GETAWAY (*GEometry, Topology, and Atom-Weights Assembly*) descriptors are derived from the **Molecular Influence Matrix** (MIM), that is, a matrix representation of molecules denoted by **H** and defined as the following [Consonni, Todeschini *et al.*, 2002a, 2002b]:

$$\mathbf{H} = \mathbf{M} \times (\mathbf{M}^T \times \mathbf{M})^{-1} \times \mathbf{M}^T$$

where **M** is the $\rightarrow$ *molecular matrix* consisting of the centered Cartesian coordinates *x*, *y*, *z* of the molecule atoms (hydrogens included) in a chosen conformation. Atomic coordinates are assumed to be calculated with respect to the geometrical center of the molecule to obtain translational invariance. The molecular influence matrix is a symmetric $A \times A$ matrix, where *A* represents the number of atoms, and shows rotational invariance with respect to the molecule coordinates, thus resulting independent of molecule $\rightarrow$ *alignment rules*.

The diagonal elements h_{ii} of the molecular influence matrix, called *leverages* being the elements of the $\rightarrow$ *leverage matrix* defined in statistics, range from 0 to 1 and encode atomic information related to the “influence” of each molecule atom in determining the whole shape of the molecule; in effect, mantle atoms always have higher h_{ii} values than atoms near the molecule center. Moreover, the magnitude of the maximum leverage in a molecule depends on the size and shape of the molecule. As derived from the geometry of the molecule, leverage values are effectively sensitive to significant conformational changes and to the bond lengths that account for atom types and bond multiplicity.

Each off-diagonal element h_{ij} represents the degree of accessibility of the *j*th atom to interaction with the *i*th atom or, in other words, the attitude of the two considered atoms to interact with each other. A negative sign for the off-diagonal elements means that the two atoms occupy opposite molecular regions with respect to the center, hence the degree of their mutual accessibility should be low.

Combining the elements of the molecular influence matrix **H** with those of the $\rightarrow$ *geometry matrix* **G**, which encodes spatial relationships between pairs of atoms, another symmetric $A \times A$ molecular matrix, called **influence/distance matrix** and denoted by **R**, is derived as the following:

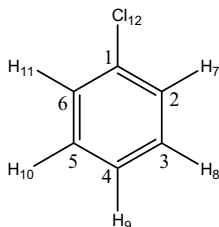
$$[\mathbf{R}]_{ij} \equiv \left[\frac{\sqrt{h_i \cdot h_j}}{r_{ij}} \right]_{ij} \quad i \neq j$$

where h_i and h_j are the leverages of the atoms *i* and *j*, and r_{ij} is their geometric distance. The diagonal elements of the matrix **R** are zero, while each off-diagonal element *i-j*, resembling the single terms in the summation of the $\rightarrow$ *gravitational indices*, is calculated by the ratio of the geometric mean of the corresponding *i*th and *j*th diagonal elements of the matrix **H** over the interatomic distance r_{ij} provided by the geometry matrix **G**.

The square-root product of the leverages of two atoms is divided by their interatomic distance to make less significant contributions from pairs of atoms far apart, according to the basic idea that interaction between atoms in the molecule decreases as their distance increases. Obviously, the largest values of the matrix elements derive from the most external atoms (i.e., those with high leverages) and simultaneously next to each other in the molecular space (i.e., those having small interatomic distances).

Example G1

Hydrogen-filled molecular graph and molecular influence matrix of chlorobenzene, whose three-dimensional structure was optimized by minimizing the conformational energy.

Molecular Influence Matrix **H**

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	H ₇	H ₈	H ₉	H ₁₀	H ₁₁	Cl ₁₂
C ₁	0.065	0.031	-0.036	-0.070	-0.036	0.031	0.057	-0.063	-0.123	-0.063	0.057	0.148
C ₂	0.031	0.075	0.042	-0.034	-0.077	-0.044	0.134	0.076	-0.059	-0.136	-0.079	0.071
C ₃	-0.036	0.042	0.079	0.039	-0.039	-0.077	0.075	0.141	0.068	-0.071	-0.138	-0.082
C ₄	-0.070	-0.034	0.039	0.075	0.039	-0.034	-0.061	0.067	0.132	0.067	-0.061	-0.159
C ₅	-0.036	-0.077	-0.039	0.039	0.079	0.042	-0.138	-0.071	0.068	0.141	0.075	-0.082
C ₆	0.031	-0.044	-0.077	-0.034	0.042	0.075	-0.079	-0.136	-0.059	0.076	0.134	0.071
H ₇	0.057	0.134	0.075	-0.061	-0.138	-0.079	0.242	0.135	-0.108	-0.246	-0.141	0.130
H ₈	-0.063	0.076	0.141	0.067	-0.071	-0.136	0.135	0.250	0.118	-0.129	-0.246	-0.143
H ₉	-0.123	-0.059	0.068	0.132	0.068	-0.059	-0.108	0.118	0.232	0.118	-0.108	-0.280
H ₁₀	-0.063	-0.136	-0.071	0.067	0.141	0.076	-0.246	-0.129	0.118	0.250	0.135	-0.143
H ₁₁	0.057	-0.079	-0.138	-0.061	0.075	0.134	-0.141	-0.246	-0.108	0.135	0.242	0.130
Cl ₁₂	0.148	0.071	-0.082	-0.159	-0.082	0.071	0.130	-0.143	-0.280	-0.143	0.130	0.337

It can be noted that the outer atoms (Cl and hydrogens) have larger leverage values (0.337, 0.242, 0.250, 0.232) than the carbon atoms of the aromatic ring (0.065, 0.075, 0.079). Then, among the outer atoms, the chlorine atom has the largest value (0.337), with its bond length larger than the bond distances of hydrogens. It must also be noted that equal leverage values are obtained for symmetric atoms, such as (C₂, C₆), (C₃, C₅), (H₇, H₁₁), and (H₈, H₁₀). Moreover, the off-diagonal terms give, to some extent, information on the relative spatial position of pairs of atoms. For instance, atoms C₁, C₂, C₆, H₇ and H₁₁ have positive off-diagonal values with respect to the chlorine atom and, among these, C₁ has the largest value being the nearest one.

A set of the GETAWAY descriptors (H_{GM} , I_{TH} , I_{SH} , HIC , $RARS$, $RCON$, and $REIG$) was derived by applying some traditional matrix operators and concepts of information theory both to the molecular influence matrix $\mathbf{H}$ and to the influence/distance matrix $\mathbf{R}$. Most of these descriptors are simply calculated only by the leverages used as the atomic weightings.

The **geometric mean of the leverage magnitude** (H_{GM}) is defined as

$$H_{GM} = 100 \cdot \left(\prod_{i=1}^A h_i \right)^{1/A}$$

where A is the number of atoms and h_i the leverage of the i th atom. It was proposed to encompass information related to molecular shape. It has, in effect, been found that in an isomeric series of hydrocarbons, the H_{GM} index is sensitive to the molecular shape increasing from linear to more branched molecules; it is also inversely related to molecular size, decreasing as the number of atoms in the molecule increases.

The **total information content on the leverage equality** (I_{TH}) and **standardized information content on the leverage equality** (I_{SH}) are defined as

$$I_{TH} = A_0 \cdot \log_2 A_0 - \sum_{g=1}^G N_g \cdot \log_2 N_g \quad I_{SH} = \frac{I_{TH}}{A_0 \cdot \log_2 A_0} = 1 - \frac{\sum_{g=1}^G N_g \cdot \log_2 N_g}{A_0 \cdot \log_2 A_0}$$

where A_0 is the number of nonhydrogen atoms, N_g the number of atoms with equal leverage value (within a certain tolerance), and G is the number of equivalence classes.

These descriptors mainly encode information on molecular symmetry; if all the atoms have different leverage values, that is, the molecule does not show any element of symmetry, $I_{TH} = A_0 \log A_0$ and $I_{SH} = 1$; otherwise, if all the atoms have equal leverage values (a perfectly symmetric theoretical case), $I_{TH} = 0$ and $I_{SH} = 0$. The total information content on the leverage equality I_{TH} is more discriminating than I_{SH} because of its dependence on molecular size, and thus it could be thought of as a measure of $\rightarrow$ *molecular complexity*. These indices were demonstrated to be useful in modeling physico-chemical properties related to entropy and symmetry [Consonni, Todeschini *et al.*, 2002b].

The **mean information content on the leverage magnitude** (HIC) is defined as

$$HIC \equiv \bar{I}_H = - \sum_{i=1}^A \frac{h_i}{M} \cdot \log_2 \frac{h_i}{M}$$

where M is a constant equal to 1 for linear, 2 for planar, and 3 for nonplanar molecules. This descriptor seems to encompass more information related to molecular complexity than the total and standardized information content on the leverage equality. Unlike I_{TH} and I_{SH} , HIC can, for example, recognize the different substituents in a series of monosubstituted benzenes. It is also sensitive to the presence of multiple bonds.

The **average row sum of the influence/distance matrix** ($RARS$) and **R-connectivity index** ($RCON$) are defined as

$$RARS = \frac{1}{A} \cdot \sum_{i=1}^A \sum_{j=1}^A \frac{\sqrt{h_i \cdot h_j}}{r_{ij}} = \frac{1}{A} \cdot \sum_{i=1}^A VS_i(\mathbf{R})$$

$$RCON = \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot (VS_i(\mathbf{R}) \cdot VS_j(\mathbf{R}))^{1/2}$$

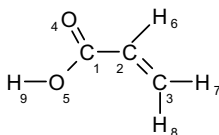
where VS_i is the i th row sum of the influence/distance matrix $\mathbf{R}$, A the number of atoms, and elements a_{ij} in $RCON$ are equal to 1 for pairs of bonded atoms and zero otherwise. The row sums VS encode useful information that may be related to the presence of significant substituents or fragments in the molecule. It was, in effect, observed that larger row sums correspond to terminal atoms that are located very next to other terminal atoms such as those in substituents on a parent structure. Moreover, the $RCON$ index is very sensitive to the molecular size as well as to conformational changes and cyclicity.

The **R -matrix leading eigenvalue** ($REIG$), in analogy with the $\rightarrow$ *Lovasz–Pelikan index*, that is, an index of molecular branching calculated as the first eigenvalue of the $\rightarrow$ *adjacency matrix*, is the largest eigenvalue of the influence/distance matrix $\mathbf{R}$.

$RARS$ and $REIG$ indices are closely related; their values decrease as the molecular size increases and seem to be a little more sensitive to molecular branching than to cyclicity and conformational changes.

Example G2

Hydrogen-filled molecular graph, molecular influence matrix $\mathbf{H}$, and influence/distance matrix $\mathbf{R}$ for acrylic acid. The matrices were calculated from the x, y, z coordinates of the atoms in the minimum energy conformation optimized by AM1 semiempirical method. Calculation of H_{GM} , I_{TH} , I_{SH} , HIC , $RARS$, $RCON$, and $REIG$ indices for acrylic acid is here exemplified. VS_i indicates the matrix row sums.



Molecular influence matrix $\mathbf{H}$

	C1	C2	C3	O4	O5	H6	H7	H8	H9
C1	0.056	0.004	-0.076	0.130	0.017	0.037	-0.114	-0.110	0.056
C2	0.004	0.054	0.009	0.040	-0.096	0.134	0.049	-0.071	-0.122
C3	-0.076	0.009	0.109	-0.171	-0.048	-0.018	0.170	0.135	-0.109
O4	0.130	0.040	-0.171	0.321	-0.017	0.163	-0.233	-0.293	0.059
O5	0.017	-0.096	-0.048	-0.017	0.179	-0.225	-0.136	0.082	0.243
H6	0.037	0.134	-0.018	0.163	-0.225	0.347	0.061	-0.230	-0.270
H7	-0.114	0.049	0.170	-0.233	-0.136	0.061	0.291	0.157	-0.247
H8	-0.110	-0.071	0.135	-0.293	0.082	-0.230	0.157	0.292	0.038
H9	0.056	-0.122	-0.109	0.059	0.243	-0.270	-0.247	0.038	0.351

Influence/distance matrix **R**

	C1	C2	C3	O4	O5	H6	H7	H8	H9	VS _i
C1	0	0.037	0.031	0.108	0.073	0.064	0.037	0.046	0.073	0.469
C2	0.037	0	0.058	0.054	0.041	0.124	0.059	0.059	0.043	0.475
C3	0.031	0.058	0	0.052	0.050	0.091	0.162	0.162	0.052	0.658
O4	0.108	0.054	0.052	0	0.109	0.125	0.067	0.077	0.150	0.742
O5	0.073	0.041	0.050	0.109	0	0.074	0.059	0.091	0.258	0.755
H6	0.064	0.124	0.091	0.125	0.074	0	0.126	0.102	0.086	0.792
H7	0.037	0.059	0.162	0.067	0.059	0.126	0	0.157	0.066	0.733
H8	0.046	0.059	0.162	0.077	0.091	0.102	0.157	0	0.092	0.786
H9	0.073	0.043	0.052	0.150	0.258	0.086	0.066	0.092	0	0.820

$$H_{GM} = 100 \times \left(\prod_{i=1}^9 h_i \right)^{1/9} = 100 \times (0.059 \times 0.054 \times 0.109 \times 0.321 \times 0.179 \times 0.347 \times 0.291 \times 0.292 \times 0.351)^{1/9} = 179.8$$

$$I_{TH} = 5 \times \log_2 5 - \sum_{g=1}^5 N_g \times \log_2 N_g = 11.61 - 5 \times (1 \times \log_2 1) = 11.61$$

$$I_{SH} = \frac{I_{TH}}{5 \times \log_2 5} = \frac{11.61}{11.61} = 1$$

$$HIC = \bar{I}_H = - \sum_{i=1}^9 \frac{h_i}{2} \times \log_2 \frac{h_i}{2} = - \frac{0.056}{2} \times \log_2 \frac{0.056}{2} - \frac{0.054}{2} \times \log_2 \frac{0.054}{2} - \frac{0.109}{2} \times \log_2 \frac{0.109}{2} - \frac{0.321}{2} \times \log_2 \frac{0.321}{2} - \frac{0.179}{2} \times \log_2 \frac{0.179}{2} - \frac{0.347}{2} \times \log_2 \frac{0.347}{2} - \frac{0.291}{2} \times \log_2 \frac{0.291}{2} - \frac{0.292}{2} \times \log_2 \frac{0.292}{2} - \frac{0.351}{2} \times \log_2 \frac{0.351}{2} = 2.938$$

$$RARS = \frac{1}{9} \times (0.469 + 0.475 + 0.658 + 0.742 + 0.755 + 0.792 + 0.733 + 0.786 + 0.820) = 0.692$$

$$RCON = (0.469 \times 0.475 + 0.469 \times 0.742 + 0.469 \times 0.755 + 0.475 \times 0.658 + 0.475 \times 0.792 + 0.658 \times 0.733 + 0.658 \times 0.786 + 0.755 \times 0.820)^{1/2} = 5.028$$

The set of the eigenvalues of the influence/distance matrix **R** is 0.713, 0.159, 0.022, -0.037, -0.103, -0.149, -0.166, -0.177, -0.263. Therefore, REIG = 0.713.

The other set of GETAWAY descriptors consists of autocorrelation vectors obtained by double-weighting the molecule atoms in such way as to account for atomic mass, polarizability, van der Waals volume, and electronegativity together with 3D information encoded by the elements of the molecular influence matrix **H** and influence/distance matrix **R**.

HATS indices are defined by analogy with the $\rightarrow$ Moreau–Broto autocorrelation descriptors ATS, weighting each atom of the molecule by its physico-chemical properties combined with the

diagonal elements of the molecular influence matrix **H**, thus also accounting for the 3D features of the molecules:

$$HATS_k(w) = \sum_{i=1}^A \sum_{j \geq i}^A (w_i \cdot h_i) \cdot (w_j \cdot h_j) \cdot \delta(d_{ij}; k) \quad \text{for } k = 0, 1, 2, \dots, D$$

where w is an atomic weighting scheme and $\delta(d_{ij}; k)$ a Dirac delta function equal to 1 when the topological distance d_{ij} between atoms i and j is equal to k and zero otherwise. D is the molecule $\rightarrow$ topological diameter, that is, the maximum topological distance in the molecule.

The **HATS total index** (*HATS*) is defined as the sum of all the *HATS* indices as

$$HATS(w) = HATS_0(w) + 2 \cdot \sum_{k=1}^D HATS_k(w)$$

Example G3

Calculation of *HATS*(m) indices for acrylic acid. This is based on the atomic mass weighting scheme scaled on the Carbon atom: $m(\text{C}) = 1$, $m(\text{H}) = 0.084$, $m(\text{O}) = 1.332$. The molecular influence matrix **H** is in Example G2. Because the topological diameter D is equal to 5, six *HATS* indices ($k = 0, 5$) can be derived. Examples of calculation for $k = 0$ and $k = 3$ are reported. For $k = 0$, the summation goes over the single atoms, then:

$$\begin{aligned} HATS_0(m) &= \sum_{i=1}^9 (m_i \cdot h_i)^2 = 0.003 + 0.003 + 0.012 + 0.183 \\ &\quad + 0.057 + 0.001 + 0.001 + 0.001 + 0.001 = 0.262 \end{aligned}$$

For $k = 3$, the summation goes over all of the atom pairs at topological distance 3:

$$\begin{aligned} HATS_3(m) &= (m_1 \cdot h_1) \cdot (m_7 \cdot h_7) + (m_1 \cdot h_1) \cdot (m_8 \cdot h_8) + (m_2 \cdot h_2) \cdot (m_9 \cdot h_9) + (m_3 \cdot h_3) \cdot (m_4 \cdot h_4) \\ &\quad + (m_3 \cdot h_3) \cdot (m_5 \cdot h_5) + (m_4 \cdot h_4) \cdot (m_9 \cdot h_9) + (m_4 \cdot h_4) \cdot (m_6 \cdot h_6) \\ &\quad + (m_5 \cdot h_5) \cdot (m_6 \cdot h_6) + (m_6 \cdot h_6) \cdot (m_7 \cdot h_7) + (m_6 \cdot h_6) \cdot (m_8 \cdot h_8) \\ &= 0.001 + 0.001 + 0.002 + 0.047 \\ &\quad + 0.026 + 0.013 + 0.012 + 0.007 + 0.001 + 0.001 = 0.110 \end{aligned}$$

H indices are filtered autocorrelation descriptors defined as

$$H_k(w) = \sum_{i=1}^A \sum_{j \geq i}^A h_{ij} \cdot w_i \cdot w_j \cdot \delta(d_{ij}; h_{ij}; k) \quad \text{for } k = 0, 1, 2, \dots, D$$

where h_{ij} are the off-diagonal elements of the molecular influence matrix **H** and the Dirac delta function $\delta(d_{ij}; h_{ij}; k)$ here is defined as

$$\delta(d_{ij}; h_{ij}; k) = \begin{cases} 1 & \text{if } d_{ij} = k \text{ and } h_{ij} > 0 \\ 0 & \text{if } d_{ij} \neq k \text{ or } h_{ij} \leq 0 \end{cases}$$

While the *HATS* indices make use of the diagonal elements of the matrix **H**, the *H* indices exploit the off-diagonal elements, which can be either positive or negative. To emphasize interactions between spatially near atoms, only off-diagonal positive h values are used. In effect,

for a given lag (i.e., topological distance), the product of the atom properties is multiplied by the corresponding h_{ij} value and only those contributions with a positive h_{ij} value are considered. This means that, for a given atom i , only those atoms j at topological distance d_{ij} with a positive h_{ij} value are considered because they may have the chance to interact with the i th atom.

The **H total index (HT)** is defined as the sum of all the H indices:

$$HT(w) = H_0(w) + 2 \cdot \sum_{k=1}^D H_k(w)$$

Example G4

Calculation of $H(m)$ indices for acrylic acid. Calculation is based on the atomic mass weighting scheme scaled on the Carbon atom: $m(\text{C}) = 1$, $m(\text{H}) = 0.084$, $m(\text{O}) = 1.332$. The molecular influence matrix **H** is in Example G2. Because the topological diameter D is equal to 5, six H indices ($k = 0, 5$) can be derived. Examples of calculations for $k = 0$ and $k = 3$ are reported.

For $k = 0$, the summation goes over the single atoms, then:

$$\begin{aligned} H_0(m) &= \sum_{i=1}^9 h_i m_i^2 = 0.056 + 0.054 + 0.109 + 0.570 + 0.318 \\ &\quad + 0.002 + 0.002 + 0.002 + 0.002 = 1.115 \end{aligned}$$

For $k = 3$, the summation goes over the atom pairs at topological distance 3, which have a positive h_{ij} value:

$$\begin{aligned} H_3(m) &= (m_4 \times h_4)(m_9 \times h_9) + (m_4 \times h_4) \times (m_6 \times h_6) + (m_6 \times h_6) \times (m_7 \times h_7) \\ &= 0.0182 + 0.0066 + 0.0004 = 0.025 \end{aligned}$$

R indices are defined in the same way as the H indices, by using the off-diagonal elements of the influence/distance matrix **R** instead of the elements of the matrix **H**:

$$R_k(w) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{\sqrt{h_i \cdot h_j}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(d_{ij}; k) \quad \text{for } k = 1, 2, \dots, D$$

In this case, no filtering is applied, because geometrical distances r_{ij} act as a smoothing function.

The **R total index (RT)** is defined as twice the sum of the R indices:

$$RT(w) = 2 \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{\sqrt{h_i \cdot h_j}}{r_{ij}} \cdot w_i \cdot w_j = 2 \cdot \sum_{k=1}^D R_k(w)$$

where w is the atomic property and D is the topological diameter. In the case of unitary weights, that is, $w = 1$, the R total index is twice the $\rightarrow$ *Wiener-type index* derived from the influence-distance matrix as the half sum of all the matrix elements. Moreover, it is strictly related to the $\rightarrow$ *gravitational index G1*.

To take into account local aspects of the molecule and allow $\rightarrow$ *reversible decoding*, the **maximal R indices (R^+)** were also proposed as

$$R_k^+(w) = \max_{ij} \left(\frac{\sqrt{h_i \cdot h_j}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(d_{ij}; k) \right) \quad i \neq j; k = 1, 2, \dots, D$$

where only the maximum property product between atom pairs at a given topological distance (lag) is retained.

The maximum value among the k th order maximal indices $R_k^+(w)$ is called **maximal R total index** (RT^+) and defined as

$$RT^+(w) = \max_k(R_k^+(w))$$

Example G5

Calculation of $R(m)$ and $R^+(m)$ indices for acrylic acid. Calculation is based on the atomic mass weighting scheme scaled on the Carbon atom: $m(C) = 1$, $m(H) = 0.084$, $m(O) = 1.332$. The influence/distance matrix R is given in Example G2. Because the topological diameter D is equal to 5, five R indices ($k = 1, 5$) can be derived. Example of calculations for $k = 3$ is reported. In this case, the summation goes over all the atom pairs at topological distance 3:

$$\begin{aligned} R_3(m) &= [R]_{1,7} \cdot m_1 \cdot m_7 + [R]_{1,8} \cdot m_1 \cdot m_8 + [R]_{2,9} \cdot m_2 \cdot m_9 + [R]_{3,4} \cdot m_3 \cdot m_4 + [R]_{3,5} \cdot m_3 \cdot m_5 \\ &\quad + [R]_{4,9} \cdot m_4 \cdot m_9 + [R]_{4,6} \cdot m_4 \cdot m_6 + [R]_{5,6} \cdot m_5 \cdot m_6 + [R]_{6,7} \cdot m_6 \cdot m_7 + [R]_{6,8} \cdot m_6 \cdot m_8 \\ &= 0.003 + 0.004 + 0.004 + 0.069 + 0.067 \\ &\quad + 0.017 + 0.014 + 0.008 + 0.001 + 0.001 = 0.188 \end{aligned}$$

$$R_3^+(m) = \max(0.003; 0.004; 0.004; 0.069; 0.067; 0.017; 0.014; 0.008; 0.001; 0.001) = 0.069$$

Note that the $R_3^+(m)$ index identifies the structural fragment $C_3 = C_2 - C_1 = O_4$.

The atomic weighting schemes applied for GETAWAY descriptor calculation are those proposed for the $\rightarrow$ WHIM descriptors, that is, atomic mass (m), $\rightarrow$ atomic polarizability (p), Sanderson $\rightarrow$ atomic electronegativity (e), atomic $\rightarrow$ van der Waals volume (v), and the unit weighting scheme (u).

HATS, H , R , and maximal R indices are $\rightarrow$ vectorial descriptors for structure–property correlations, but they can also be used as molecular profiles suitable for $\rightarrow$ similarity/diversity analysis studies. These descriptors, based on spatial autocorrelation, encode information on structural fragments and therefore seem to be particularly suitable for describing differences in congeneric series of molecules. Unlike the Moreau–Broto autocorrelations, GETAWAYs are geometrical descriptors encoding information on the effective position of substituents and fragments in the molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties.

A joint use of GETAWAY and WHIM descriptors is advised, exploiting both local information of the former and holistic information of the latter set of descriptors. The GETAWAY descriptors have been used for modeling several data sets of pharmacological and environmental interest [Consonni, Todeschini *et al.*, 2002b; Fedorowicz, Singh *et al.*, 2005; Pérez González, Terán *et al.*, 2005b; Saiz-Urra, Pérez González *et al.*, 2007].

Table G1 Some GETAWAY descriptors for the data set comprised of 22 *N,N*-dimethyl- α -bromo-phenethylamines (Appendix C – Set 2).

Mol.	X	Y	I_{TH}	I_{SH}	HIC	H_{GM}	RCON	RARS	REIG	HT(u)	HT(m)	RT(u)	RT(m)
1	H	H	43.020	1.000	4.428	8.933	17.445	0.744	0.796	17.233	18.411	19.341	8.499
2	H	F	44.106	0.917	4.387	8.935	17.180	0.733	0.780	17.237	27.588	19.071	11.590
3	H	Cl	46.106	0.958	4.384	8.915	16.997	0.726	0.773	17.220	29.869	18.875	13.031
4	H	Br	46.106	0.958	4.381	8.897	16.903	0.722	0.769	17.192	40.542	18.781	17.296
5	H	I	46.106	0.958	4.389	8.936	16.945	0.722	0.768	17.269	58.506	18.783	21.530
6	H	Me	44.106	0.917	4.562	8.199	18.570	0.708	0.752	19.608	24.297	20.542	9.316
7	F	H	46.106	0.958	4.380	8.905	17.123	0.731	0.778	17.179	27.692	19.009	11.965
8	Cl	H	46.106	0.958	4.371	8.846	16.880	0.721	0.768	17.141	30.052	18.739	13.709
9	Br	H	46.106	0.958	4.367	8.825	16.793	0.717	0.763	17.126	42.815	18.636	18.862
10	I	H	46.106	0.958	4.361	8.801	16.701	0.712	0.759	17.095	66.534	18.525	24.200
11	Me	H	48.106	1.000	4.546	8.081	18.754	0.707	0.755	19.323	23.817	20.502	9.407
12	Cl	F	51.303	0.962	4.372	8.836	16.788	0.717	0.764	17.150	33.358	18.648	16.166
13	Br	F	51.303	0.962	4.364	8.797	16.670	0.712	0.760	17.094	47.747	18.523	22.247
14	Me	F	47.303	0.887	4.539	7.975	18.541	0.701	0.751	19.305	24.883	20.331	11.064
15	Cl	Cl	53.303	1.000	4.371	8.822	16.619	0.710	0.757	17.160	36.601	18.468	18.041
16	Br	Cl	53.303	1.000	4.359	8.768	16.462	0.704	0.752	17.040	52.386	18.310	24.842
17	Me	Cl	47.303	0.887	4.533	7.953	18.326	0.694	0.743	19.219	27.406	20.126	12.496
18	Cl	Br	51.303	0.962	4.364	8.792	16.501	0.706	0.753	17.083	49.660	18.354	23.683
19	Br	Br	53.303	1.000	4.355	8.744	16.365	0.701	0.748	17.008	68.303	18.215	32.193
20	Me	Br	49.303	0.925	4.544	7.979	18.348	0.694	0.743	19.321	35.356	20.133	16.136
21	Me	Me	53.303	1.000	4.711	7.439	20.224	0.691	0.736	21.603	22.027	22.096	9.134
22	Br	Me	51.303	0.962	4.555	8.155	18.266	0.696	0.739	19.438	42.620	20.173	17.704

X and Y refer to molecule substituents.

📖 [Consonni and Todeschini, 2001; Grodnitzky and Coats, 2002; Gramatica, Pilutti *et al.*, 2003a, 2004b; Kiralj, Takahata *et al.*, 2003; Pérez González and Helguera, 2003; Farkas, Héberger *et al.*, 2004; Fedorowicz, Zheng *et al.*, 2004; Garkani-Nejad, Karlovits *et al.*, 2004; Gramatica, Battaini *et al.*, 2004; Guha, Serra *et al.*, 2004; Jelcic, 2004; Marrero-Ponce, 2004a; Pérez González, Helguera Morales *et al.*, 2004; Pérez González, Helguera *et al.*, 2004; Pérez González and Moldes Teran, 2004; Schefzik, Kibbey *et al.*, 2004; Kabankin and Gabrielyan, 2005; Kovatcheva, Golbraikh *et al.*, 2004; Deconinck, Hancock *et al.*, 2005; Fedorowicz *et al.*, 2005; Panek, Jezierska *et al.*, 2005; Papa, Battaini *et al.*, 2005; Papa, Villa *et al.*, 2005; Pérez González, Terán *et al.*, 2005a; 2006; Caballero and Fernández, 2006; Li *et al.*, 2006; Li, Maldonado, Doucet *et al.*, 2006; Pis Diez, Duchowicz *et al.*, 2006; Yap, Li *et al.*, 2006; Carlucci, D'Archivio *et al.*, 2007; Cruz-Monteagudo, Borges *et al.*, 2007; Deconinck, Ates *et al.*, 2007; Duchowicz, Pérez González *et al.*, 2007; Zheng, Zheng *et al.*, 2007]

- **Ghose–Crippen descriptors** → lipophilicity descriptors (☉ Ghose–Crippen hydrophobic atomic constants)
- **Ghose–Crippen hydrophobic atomic constants** → lipophilicity descriptors
- **Gini concentration index** → statistical indices (☉ concentration indices)
- **Gini index** → information content

■ **GIPF approach** ($\equiv$ *General Interaction Properties Function approach*)

This is a general method, proposed by Politzer and coworkers, to estimate physico-chemical properties depending on noncovalent interactions [Brinck, Murray *et al.*, 1993; Murray, Brinck *et al.*, 1993; Politzer, Murray *et al.*, 1993; Murray, Brinck *et al.*, 1994]. This approach is based on molecular surface area in conjunction with some statistically based quantities related to the $\rightarrow$ *molecular electrostatic potential* (MEP) at the $\rightarrow$ *molecular surface*. The $\rightarrow$ *electron isodensity contour surface* (0.001 a.u. contour of $\rho(\mathbf{r})$) is taken as the molecular surface model.

The general GIPF model for a physico-chemical property Φ is

$$\Phi = f(SA, \Pi, \sigma_{tot}^2, \nu)$$

where SA is the surface area and Π is the $\rightarrow$ *local polarity index*. The other two molecular surface indices are defined as the following:

$$\sigma_{tot}^2 = \sigma_+^2 + \sigma_-^2 = \frac{1}{n^+} \cdot \sum_{i=1}^{n^+} [V^+(\mathbf{r}_i) - \bar{V}^+]^2 + \frac{1}{n^-} \cdot \sum_{i=1}^{n^-} [V^-(\mathbf{r}_i) - \bar{V}^-]^2 \quad \text{and}$$

$$\nu = \frac{\sigma_+^2 \cdot \sigma_-^2}{(\sigma_{tot}^2)^2}$$

where σ_{tot}^2 is the **surface electrostatic potential variance**, which measures the electrostatic interaction tendency of the molecule, σ_+^2 and σ_-^2 are the variances of positive and negative regions of the molecular surface potential, V^+ and V^- are the positive and negative values of the MEP at a grid point $\mathbf{r}$ on the molecular surface, $\bar{V}^+$ and $\bar{V}^-$ are their average values, and n^+ and n^- are the numbers of grid points with positive and negative values. ν is the **electrostatic balance term** that reaches a maximum value of 0.25 when σ_+^2 and σ_-^2 are equal.

Site-specific molecular quantities can be added to the global molecular descriptors in the GIPF model depending on the physico-chemical property to be estimated. Some of these site-specific descriptors are defined below:

$\bar{I}_{S,min}$ is the lowest value of the $\rightarrow$ *average local ionization energy* found on the molecular surface; this reflects the tendency for charge transfer and polarization at any particular molecular site [Haeberlein and Brinck, 1997].

$V_{S,min}$ and $V_{S,max}$ are the most negative and positive values of the molecular electrostatic potential on the molecular surface; the maximum reflects the tendency for long-range attraction of nucleophiles at a specific site, whereas the minimum reflects the tendency for long-range attraction of electrophiles at a specific site. $V_{S,min}$ and $V_{S,max}$ for a large variety of molecules correlate with hydrogen bond basicity and acidity, respectively [Murray and Politzer, 1998].

SA^+ and SA^- are the portions of the surface area over which $V(\mathbf{r})$ is positive and negative, respectively.

Several properties have been estimated by the GIPF approach such as heat of vaporization, sublimation [Politzer, Murray *et al.*, 1997] and fusion [Murray, Brinck *et al.*, 1996], boiling point and critical constants [Murray, Lane *et al.*, 1993a], surface tension, liquid and solid density [Murray, Brinck *et al.*, 1996], crystal lattice energy [Politzer and Murray, 1998], impact sensitivity [Murray, Lane *et al.*, 1998], diffusion coefficient [Politzer, Murray *et al.*, 1996],

solubility [Politzer, Lane *et al.*, 1992; Murray, Gagarin *et al.*, 1995], aqueous solvation free energy [Murray, Abu-Awwad *et al.*, 1999; Politzer, Murray *et al.*, 2000], → *hydrogen-bonding parameters* [Lowrey, Cramer *et al.*, 1995], and → *lipophilicity*.

📖 [Murray, Ranganathan *et al.*, 1991; Brinck, Murray *et al.*, 1993; Murray, Lane *et al.*, 1993b; Politzer and Murray, 1994; Beck, Horn *et al.*, 1998]

- **girth** → graph
- **glass transition temperature** → technological properties
- **GLI index** ≡ *Global Leachability Index* → environmental descriptors (⊙ leaching indices)
- **global cyclicity indices** → resistance matrix
- **global flexibility index** → flexibility indices
- **Global Leachability Index** → environmental indices (⊙ leaching indices)
- **global site-property analysis** → Hansch analysis
- **global synthetic invariant** → iterated line graph sequence
- **global topological charge index** → topological charge indices
- **Global Warming Potential** → environmental indices
- **global weighted walk numbers** → walk matrices
- **global WHIM descriptors** → WHIM descriptors
- **globularity** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **globularity factor** → shape descriptors (⊙ ovality index)
- **Gombar hydrophobic model** ≡ *VLOGP* → lipophilicity descriptors
- **GMPEI** ≡ *Geometric Molecular Polarizability Effect Index* → electric polarization descriptors (⊙ polarizability effect index)
- **Golbraikh–Tropsha statistics** → regression parameters
- **GOLPE** → variable selection
- **goodness of fit** → regression parameters
- **goodness of prediction** → regression parameters
- **Gordon–Scantlebury index** ≡ *connection number* → edge adjacency matrix
- **Gordy's bond order** → bond order indices (⊙ bond order–bond length relationships)

■ graph

A graph is a mathematical object defined within the *graph theory* [Harary, 1964, 1969a, 1969b; Rouvray, 1971, 1990a; Wilson, 1972; Rouvray and Balaban, 1979; Balaban and Harary, 1976; Bonchev and Rouvray, 1991, 1998; Trinajstić, 1992; Ivanciuc and Balaban, 1999c; Marks *et al.*, 2002; Ivanciuc, 2003c; Kruja, Randić, 2003b; Gutman, 2006].

Note. **Graph theory** is a branch of mathematics that studies the structure of graphs and networks. Graph theory started in 1736 when Euler solved the problem known as the Königsberg bridges problem [Euler, 1741], which was reduced by him to a graph-theoretical problem.

Although the term “graph” was first introduced into literature by mathematician Sylvester [Sylvester, 1877, 1878], who derived it from the contemporary chemical term “graphical notation,” used to denote the chemical structure of a molecule, the research field that is nowadays called *chemical graph theory* started some years before when the British mathematician Arthur Cayley published his works about *trees* [Cayley, 1857, 1859] and then the paper “*On the mathematical theory of isomers*” [Cayley, 1874].

The first chemical application of graph theory dates back to 1875 when William Clifford proposed the solution for the counting of alkane isomers. The modern chemical graph theory started with the works of Henze and Blair in 1931 [Henze and Blair, 1931a, 1931b, 1933, 1934] and Pólya in 1936 [Pólya, 1936, 1937a, 1937b; Pólya and Read, 1987].

A graph is usually denoted as $G = (V, E)$, where V is a set of **vertices** and E is a set of elements representing the binary relationship between pairs of vertices; unordered vertex pairs are called **edges**, ordered vertex pairs are called **arcs**, and elements of E that relate a vertex with itself are called **loops**. A graph is described by either **adjacencies**, which refer to adjacent vertex–vertex or edge–edge pairs or **incidences** that refer to adjacent vertex–edge pairs or, more generally, to pairs of mathematical objects of two different kinds.

An edge is a **cut edge** if its removal produces a disconnected graph; it cannot be a part of a cycle; similarly, a vertex is a **cut vertex** if its removal produces a disconnected graph. Of course, each vertex incident to a cut edge is a cut vertex, but a cut vertex can also be a part of a cycle.

If two vertices occur as an unordered pair more than once, they define a **multiple edge**; if two vertices occur as an ordered pair more than once, they define a **multiple arc**. Two edges in a graph G are said to be **independent edges** if they have no common vertex. A collection of k mutually (i.e., pairwise) independent edges in a graph G ($k \geq 2$) is called a **k -matching** of G .

G and G' are called **isomorphic graphs** if a bijective mapping of the vertex and the edge sets exists, that is,

$$V(G) \leftrightarrow V(G') \quad \text{and} \quad E(G) \leftrightarrow E(G')$$

or, in other words, if there exists a one-to-one correspondence between the vertices and the edges, such that adjacency is preserved. A **graph automorphism** is an isomorphic mapping of a graph G onto itself, that is, it is a bijective mapping of the vertex and edge sets onto themselves, which preserves the number of links joining any two vertices:

$$V(G) \leftrightarrow V(G) \quad \text{and} \quad E(G) \leftrightarrow E(G)$$

The set of all automorphisms of a graph forms an **automorphism group**. The occurrence of automorphism depends on the symmetry of the graph; in particular, it depends on the presence of equivalent vertices, which can be mapped automorphically onto each other, that is, they can interchange preserving the adjacency of the graph. The cardinality of the automorphism group of a graph is called **symmetry number** and is considered among the $\rightarrow$ *symmetry descriptors*.

Topologically equivalent vertices constitute disjoint subsets of vertices called **orbits**.

A graph G' is a **subgraph** of the graph G if the following relationships hold:

$$V(G') \subseteq V(G) \quad \text{and} \quad E(G') \subseteq E(G)$$

Graph components are connected subgraphs or vertices that are not connected to each other.

The $\rightarrow$ *vertex degree* is the number of edges incident to a given vertex. If two vertices are connected by an arc, two degrees are assigned to each vertex; the **indegree** counts the arcs ending on the vertex, the **outdegree** counts the arcs starting from the vertex. **Terminal vertices** are the vertices of a graph with degree equal to 1; **terminal edges** are the edges incident to terminal vertices. **Central vertices** and **central edges** are the vertices (edges) belonging to the $\rightarrow$ *graph center*. All vertices with vertex degree equal to zero are called **isolated vertices**. *Branching* of a graph is a fuzzy concept that can be based on the presence in the graph of vertices with degrees equal to 3 or higher. It plays a basic role in assessing the $\rightarrow$ *molecular complexity*.

A **walk** (or **random walk**) in G is a sequence of vertices $w = (v_1, v_2, \dots, v_k)$ such that $(v_i, v_{i+1}) \in E$ for each $i = 1, k-1$, that is, a sequence of pairwise adjacent edges leading from vertex v_1 to vertex v_k ; any vertex or edge can be traversed several times. The **walk length** is the number of edges traversed by the walk.

A **path** (or **self-avoiding walk**) is a walk without any repeated vertices. The **path length** is the number of edges associated with the path. The smallest path between two vertices considered is called **geodesic** and its length corresponds to the $\rightarrow$ *topological distance*; **elongation** is the longest path between two vertices considered and its length corresponds to the $\rightarrow$ *detour distance* between the vertices.

A walk closed in itself is called **self-returning walk**, that is, a walk starting and ending on the same vertex.

A self-returning path is called **cyclic path** (or **cycle** or **circuit**), that is, a cycle is a walk with no repeated vertices (i.e., a path) other than its first and last ones ($v_1 = v_k$). The number of independent cycles (or rings) in a graph is the $\rightarrow$ *cyclomatic number*. **Cyclicity** C^+ is the number of all possible cycles in a graph. A **girth** is the length of the shortest cycle (if any) in a graph G . Acyclic graphs are considered to have infinite girth.

A **trail** is a walk in which vertices can be revisited but edges can be traversed only once; an **Eulerian walk** is a trail in which all vertices of the graph must be encountered.

A **Hamiltonian path** is a path in which all vertices of the graph must be visited once and the beginning and the end are different. A **Hamiltonian circuit** is a path in which all vertices of the graph must be visited once, starting and ending on the same vertex.

A **dissection of a graph** G is a collection of subgraphs obtained by erasing one vertex at a time from a graph G and all its so-obtained subgraphs G^* generated from G , which neither represent isolated vertices nor isolated bonds [Randić, Guo *et al.*, 2000]. The dissection of a graph is determined by a stepwise procedure in which one removes one vertex at a time from the graph considered and continues to do so on all subgraphs that are neither isolated vertices nor isolated bonds. The total number a of subgraphs, which are isolated vertices, and the total number b of subgraphs, which are isolated edges, obtained by the whole graph dissection, can be considered two simple topological descriptors.

A list of graphs of practical interest follows.

- **simple graph** ($\equiv$ *normal graph*, *schlicht graph*)

Graph having no arcs, no multiple edges or loops.

- **planar graph**

Graph that can be drawn so that no edge-crossing appears.

- **cyclic graph**

Graph containing at least one cycle. Each cycle is usually denoted as C_m ($m \geq 3$), where m is the number of vertices in the cycle.

- **digraph** ($\equiv$ *directed graph*)

Graph in which all vertex pairs are arcs. If any vertex pair is associated with only one arc, the graph is called **oriented graph**.

- **multigraph** ($\equiv$ *multiple graph*)

Graph having no arcs or loops, but including multiple edges between at least a pair of vertices.

- **general graph** ($\equiv$ *nonsimple graph*)

Graph containing multiple edges and loops.

- **pseudograph** ($\equiv$ *loop-multigraph*)

Graph having no arcs or multiple edges but containing loops.

- **connected graph**

Graph in which for each pair of vertices $\{i, j\} \in V(G)$ at least one path exists. Otherwise G is called **disconnected graph**. The simplest disconnected graph is a graph with an isolated vertex and the vertices not joined by a path belong to different components of the graph. The number of components of a graph is denoted $k(G)$.

- **regular graph**

If all vertices in a graph have the same degree, then the graph is called regular graph, otherwise **irregular graph**.

- **tree** ($\equiv$ *acyclic graph*)

Connected graph without cycles, usually denoted as T_A , where A is the number of vertices in the graph. The number of edges B and the number of vertices A are related by the condition $B = A - 1$. A **rooted tree** is a tree having one vertex distinguished from the others; if this vertex is an end point, the graph is called **planted tree**. In chemistry, rooted and planted trees can be used to represent molecular substituents.

- **linear graph** ($\equiv$ *path graph*)

A tree without branching; there are exactly two terminal vertices of degree 1 and $A - 2$ vertices of degree two.

- **star graph**

A maximally branched tree, that is, a set of vertices joined by a common vertex; there are $A - 1$ terminal vertices of degree 1 and one vertex of degree $A - 1$. It is usually denoted as S_A , where A is the number of vertices in the graph.

- **complete graph**

Graph in which all vertices and edges are mutually adjacent, that is, all vertices have degree $A - 1$. The maximal number of edges in a graph is

$$B = \binom{A}{2} = \frac{A \cdot (A-1)}{2}$$

A complete graph contains the maximal number of cycles and is denoted as K_A , A being the number of vertices.

- **clique**

A maximal complete subgraph in which every vertex is connected to every other vertex and which is not contained in any other larger subgraph with this property.

- **forest**

A set of disjoint trees $\mathcal{F} = \{(V_1, \mathcal{E}_1), \dots, (V_k, \mathcal{E}_k)\}$; a forest does not contain cycles.

- **spanning tree**

A connected acyclic subgraph containing all the vertices of G .

- **minimal spanning tree**

A spanning tree in which the number of edges is minimal.

- **indexed graph**

A graph G associated to a mapping ϕ such as

$$\phi : V(G) \rightarrow \{1, 2, \dots, A\}$$

where $V(G)$ is the set of A vertices of a graph and the indexing function ϕ assigns an integer number to each vertex of the graph. Univocally defined indexed graphs are obtained by $\rightarrow$ *canonical numbering* of graph vertices.

- **Sachs graph** ($\equiv$ *basic graph, mutation graph, characteristic graph*)

A graph defined as a subgraph of G whose components are K_2 (complete graphs) or C_m (cycle graphs) or combinations between a K_2 components and b C_m components, under the constraint:

$$a \times 2 + b \times m = A$$

where A is the number of vertices in the Sachs graph.

- **signed graph**

A signed graph is a graph with a sign attached to each edge.

- **weighted graph**

A graph G in which a weight $w_{ij} \geq 0$ is assigned to each edge $\{i, j\} \in \mathcal{E}(G)$ or a weight w_i is assigned to each vertex $i \in V(G)$.

- **line graph**

A line graph $L(G)$ is a graph obtained by representing the edges of the graph G by points and then by joining two such points with a line if the edges they represent are adjacent in the original graph; the following relation holds:

$$|V(L(G))| = |\mathcal{E}(G)|$$

Multiple edges are considered as independent vertices in the line graph.

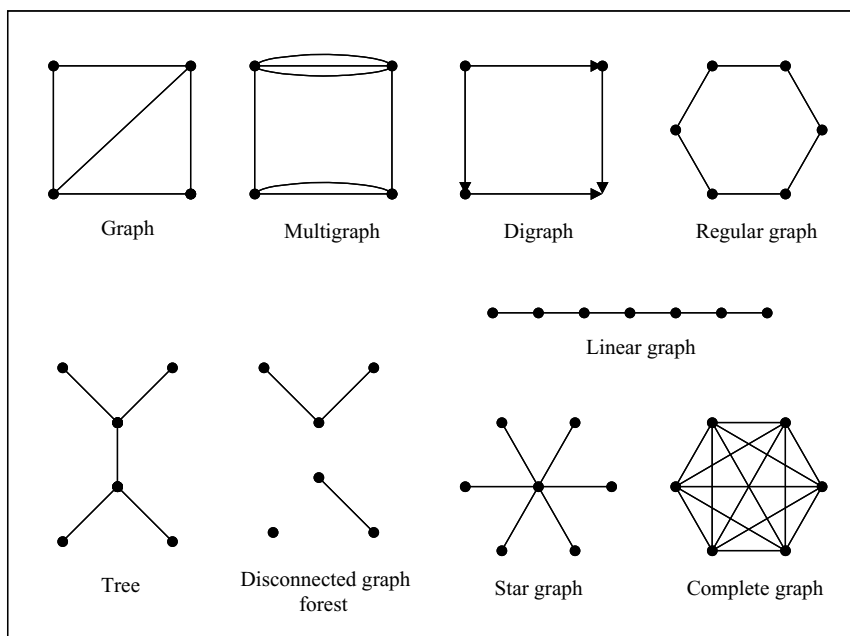


Figure G2 Examples of graphs.

- **chromatic graph**

Graph whose vertices or edges are symbolically differentiated by assigning a minimal number of different colors to the vertices (or edges) of G such that no two adjacent vertices (or edges) have the same color ($\rightarrow$ *chromatic decomposition*).

- **isocodal graphs**

Graphs with identical atomic codes for all the vertices, that is, there exists a one-to-one correspondence between the atomic codes of all vertices. The most known atomic codes are $\rightarrow$ *walk count atomic code*, $\rightarrow$ *self-returning walk count atomic code*, and $\rightarrow$ *atomic path code*.

- **isospectral graphs** ($\equiv$ *cospectral graphs*)

Nonisomorphic graphs having the same $\rightarrow$ *characteristic polynomial*.

- **subspectral graphs**

Graphs whose eigenvalues of the characteristic polynomial are contained in the spectrum of another graph.

- **homeomorphic graphs**

Graphs obtained from the same graph by a sequence of line subdivisions.

- **chemical graph**

A chemically interpreted graph is called chemical graph, that is, graph representing a chemical system such as molecules, reactions, crystals, polymers, and orbitals. The common feature of chemical systems is the presence of sites (atoms, electrons, molecules, molecular fragments, etc.)

and connections (bonds, reaction steps, van der Waals forces, etc.) between them. In the graph representation of a chemical system, sites are replaced by vertices and connections by edges. The most common chemical graphs are molecular graphs and reaction graphs. The former correspond to specific chemical structures, whereas the latter to sets of chemical reactions. A topological representation of a molecule is given by a $\rightarrow$ *molecular graph*.

📖 [Sachs, 1964; Graovac, Gutman *et al.*, 1972; Hosoya, 1972a; Balaban, 1976d, 1993e; Read and Corneil, 1977; Quintas and Slater, 1981; von Knop, Müller *et al.*, 1981; King, 1983; Randić, Woodworth *et al.*, 1983; Grossman, 1985; Balaban, Kennedy *et al.*, 1988; Balaban and Tomescu, 1988; Hansen and Jurs, 1988a; Gutman, 1991a; Liu and Klein, 1991; Polansky, 1991; Rouvray, 1991; Rücker and Rücker, 1991a, 1992; Bangov, 1992; Bonchev and Rouvray, 1992; Figueras, 1992; Gautzsch and Zinn, 1992a, 1992b, 1994; Ivanciuc and Balaban, 1992a, 1999c; Müller, Szymanski *et al.*, 1995; Lukovits, 1996a; Lepovic and Gutman, 1998; Balinska, Gargano *et al.*, 2001; Vukicević and Graovac, 2004b; Vukicević and Graovac, 2005]

- **graph automorphism** $\rightarrow$ graph
- **graph characteristic polynomial** $\rightarrow$ characteristic polynomial-based descriptors
- **graph center** $\equiv$ center of a graph
- **graph coloring** $\rightarrow$ chromatic decomposition
- **graph distance code** $\rightarrow$ distance matrix
- **graph distance complexity** $\rightarrow$ topological information indices
- **graph distance count** $\rightarrow$ distance matrix
- **graph distance index** $\rightarrow$ distance matrix
- **graph eigenvalues** $\rightarrow$ characteristic polynomial-based descriptors
- **graph energy** $\rightarrow$ spectral indices

■ graph entropy

The graph entropy approach is based on the idea to catch the structural $\rightarrow$ *mean information content* in a graph by means of an information functional f [Dehmer, 2008a, 2008b; Dehmer and Emmert-Streib, 2008].

An **information functional** is a positive and monotonous function that captures structural information in a graph by defining the probability value for each graph vertex. The probability of the vertex v_i is defined as

$$p(v_i) = \frac{f(v_i)}{\sum_{j=1}^A f(v_j)}$$

where f represents an arbitrary information functional and A the total number of graph vertices. Because, by definition, it holds the equation:

$$p(v_1) + p(v_2) + p(v_3) + \dots + p(v_A) = 1$$

the quantities $p(v_i)$ can be interpreted as the vertex probabilities.

Therefore, the graph entropy, denoted as I_f is the structural mean information content defined as

$$I_f = - \sum_{i=1}^A \frac{f(v_i)}{\sum_{j=1}^A f(v_j)} \cdot \log_2 \frac{f(v_i)}{\sum_{j=1}^A f(v_j)}$$

The information functional is defined as

$$f(v_i) = \alpha^{[c_1 \cdot {}^1f_i + c_2 \cdot {}^2f_i + \dots + c_D \cdot {}^Df_i]} \quad c_k > 0; \alpha > 0$$

where c_k are arbitrary real positive coefficients, D is the $\rightarrow$ *topological diameter*, and ${}^kf_i$ is the $\rightarrow$ *vertex distance count*, that is, the number of vertices at a $\rightarrow$ *topological distance* k from vertex v_i .

- **graphical bond order** $\rightarrow$ bond order indices
- **graphical bond order descriptors** $\rightarrow$ bond order indices ($\odot$ graphical bond order)
- **graphical matrices** $\rightarrow$ double invariants

■ **graph invariants** ($\equiv$ *graph-theoretical invariants*)

These are $\rightarrow$ *molecular descriptors* derived from a graph representation of the molecule and representing graph-theoretical properties that are preserved by isomorphism, that is, properties with identical values for $\rightarrow$ *isomorphic graphs*. A graph invariant may be a $\rightarrow$ *characteristic polynomial*, a sequence of numbers ($\rightarrow$ *vectorial descriptors*) or a single numerical index obtained by the application of $\rightarrow$ *algebraic operators* to $\rightarrow$ *graph-theoretical matrices* and whose values are independent of vertex numbering or labelling [Kier and Hall, 1976a, 1986; Bonchev and Trinajstić, 1977; Bonchev, Mekenyan *et al.*, 1979; Balaban, Motoc *et al.*, 1983; Rouvray, 1983, 1995, 1989a; Basak, Magnuson *et al.*, 1987; Hansen and Jurs, 1988a; Basak, Niemi *et al.*, 1990c; Trinajstić, 1992; Randić, 1993a, 1998c, 2003b; Balaban, 1997a, 1998; Basak, Grunwald *et al.*, 1997; Diudea and Gutman, 1998; Balaban and Ivanciuc, 1999; Devillers and Balaban, 1999; Ivanciuc and Balaban, 1999c; Bonchev and Rouvray, 2000; Bonchev, 2003a; Ivanciuc, 2003c; Kerber, Laue *et al.*, 2004].

Single indices that are a numerical representation of the molecular structure derived from a $\rightarrow$ *molecular graph* are called **topological indices** (TIs) or **molecular topological indices** (MTIs). These are numerical quantifiers of molecular topology that are mathematically derived in a direct and unambiguous manner from the structural graph of a molecule, usually a $\rightarrow$ *H-depleted molecular graph*. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching, and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. In fact, topological indices were proposed to be divided into two categories: *topostructural* and *topochemical indices* [Basak, Gute *et al.*, 1997; Gute, Grunwald *et al.*, 1999]. **Topostructural indices** encode only information about the adjacency and distances between atoms in the molecular structure; **topochemical indices** quantify information about not only topology but also specific chemical properties of atoms such as their chemical identity and hybridization state (Figure G3).

$\rightarrow$ *Topological information indices* are graph invariants, based on information theory and calculated as $\rightarrow$ *information content* of specified equivalence relationships on the molecular graph.

Topological indices are mainly based on distances between atoms calculated by the number of intervening bonds and are thus considered *through-bond* indices; they differ from $\rightarrow$ *topographic indices* and $\rightarrow$ *geometrical descriptors* that are, instead, considered *through-space* indices because they are based on interatomic $\rightarrow$ *geometric distances* [Diudea, Horvath *et al.*, 1995b; Balaban, 1997a].

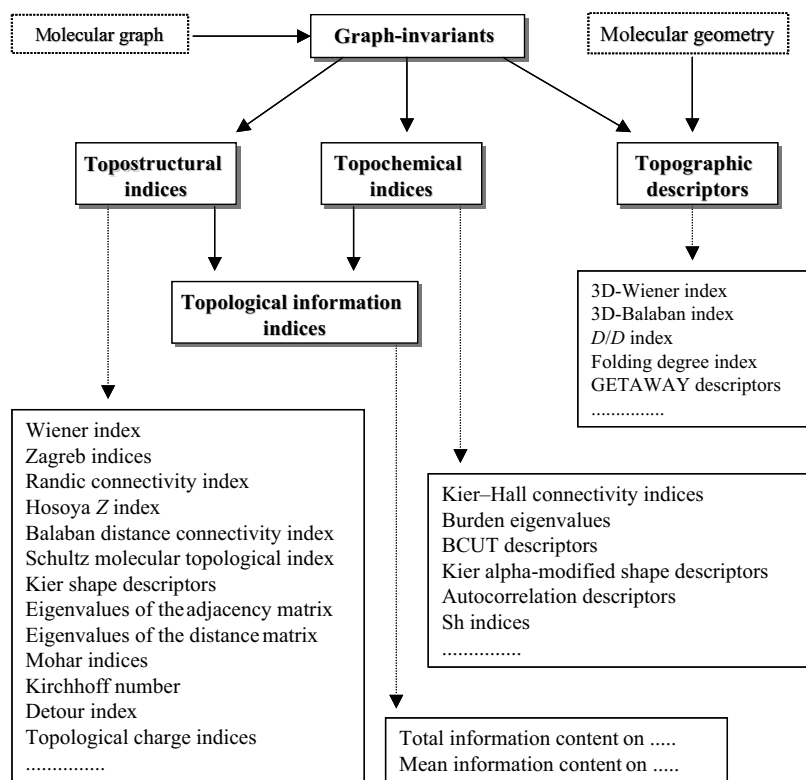


Figure G3 Different classes of graph invariants.

In general, TIs do not uniquely characterize molecular topology, different structures may have some of the same TIs. A consequence of topological indices' nonuniqueness is that they do not, in general, allow reconstructing molecule. Therefore, suitably defined ordered sequences of TIs can be used to characterize molecules with higher discrimination.

There are several ways to obtain topological descriptors. Simple topological indices consist in the counting of some specific graph elements; for examples, the $\rightarrow$ *Hosoya Z index*, $\rightarrow$ *path counts*, *walk counts*, $\rightarrow$ *self-returning walk counts*, $\rightarrow$ *Kier shape descriptors*, and $\rightarrow$ *path/walk shape indices*. However, the most common TIs are derived by applying some algebraic operators (e.g., the $\rightarrow$ *Wiener operator*) to $\rightarrow$ *graph-theoretical matrices*, such as $\rightarrow$ *adjacency matrix A*, $\rightarrow$ *distance matrix D*, $\rightarrow$ *detour matrix Δ* , $\rightarrow$ *Szeged matrices SZ*, $\rightarrow$ *Cluj matrices CJ*, $\rightarrow$ *layer matrices LM*, and $\rightarrow$ *walk matrices W*; among them there are the $\rightarrow$ *Wiener index*, $\rightarrow$ *Randić connectivity index* and related indices, $\rightarrow$ *Balaban distance connectivity index*, $\rightarrow$ *Schultz molecular topological index*, $\rightarrow$ *hyper-Wiener index*, $\rightarrow$ *quasi-Wiener index*, $\rightarrow$ *spectral indices*, $\rightarrow$ *determinant-based descriptors*, and $\rightarrow$ *Harary indices*. The most common functions to derive graph invariants from graph-theoretical matrices are listed in Table G2. Note that in functions $\mathcal{D}_1$ and $\mathcal{D}_2$, the most common parameter values are $\alpha = 1/2$ and $\lambda = 1$. Function $\mathcal{D}_3$ is used to generate descriptors derived from the matrix determinant and function $\mathcal{D}_4$ descriptors that

are linear combinations of the coefficients of the characteristic polynomial of a graph-theoretical matrix, such as the $\rightarrow$ *Hosoya-type indices*. Function $\mathcal{D}_5$ is based on the eigenvalues calculated from graph-theoretical matrices, and the related molecular descriptors are the so-called $\rightarrow$ *spectral indices*. Function $\mathcal{D}_6$ makes use of the matrix row sums VS_i ($\rightarrow$ *row sum operator*) as the $\rightarrow$ *local vertex invariants* and, then, adds up the contributions from different graph fragments (e. g., edges), each weighted by the product of the local invariants of all the vertices contained in the fragment; $\rightarrow$ *connectivity-like indices*, such as the $\rightarrow$ *Randić connectivity index* and $\rightarrow$ *Balaban-like indices*, are calculated according to this function. Function $\mathcal{D}_7$ for $\alpha = 1/2$ and $\lambda = 2$ generates the $\rightarrow$ *hyper-Wiener-type indices*.

Table G2 Classical functions to derive molecular descriptors from graph-theoretical matrices.

ID	Function	ID	Function
1.	$\mathcal{D}_1(\mathbf{M}; \alpha, \lambda) = \alpha \cdot \sum_{i=1}^n \sum_{j=1}^n [\mathbf{M}]_{ij}^\lambda$	5.	$\mathcal{D}_5(\mathbf{M}) = f(\mathbf{\Lambda}(\mathbf{M}))$
2.	$\mathcal{D}_2(\mathbf{M}; \alpha, \lambda) = \alpha \cdot \sum_{i=1}^n \sum_{j=1}^n a_{ij} [\mathbf{M}]_{ij}^\lambda$	6.	$\mathcal{D}_6(\mathbf{M}; \alpha, \lambda) = \alpha \cdot \sum_{k=1}^K \left(\prod_{i=1}^{n_k} VS_i(\mathbf{M}) \right)_k^\lambda$
3.	$\mathcal{D}_3(\mathbf{M}; \alpha) = \alpha \cdot \det(\mathbf{M})$	7.	$\mathcal{D}_7(\mathbf{M}; \alpha, \lambda) = \alpha \cdot \sum_{i=1}^n \sum_{j=1}^n ([\mathbf{M}]_{ij}^\lambda + [\mathbf{M}]_{ji}^\lambda)$
4.	$\mathcal{D}_4(\mathbf{M}; \alpha, \lambda) = \alpha \cdot \sum_{i=0}^n c(Ch(\mathbf{M}; \mathbf{x}))_i ^\lambda$	8.	$\mathcal{D}_8(\mathbf{M}; \alpha, \lambda) = \alpha \cdot \max_{ij} ([\mathbf{M}]_{ij}^\lambda)$

$\mathbf{M}$ is a graph-theoretical matrix, n the matrix dimension, $c(Ch(\mathbf{M}; \mathbf{x}))_i$ the i th coefficient of the characteristic polynomial of $\mathbf{M}$, $\mathbf{\Lambda}(\mathbf{M})$ indicates the graph spectrum (i.e., the set of eigenvalues of $\mathbf{M}$), and α and λ are real parameters. In function $\mathcal{D}_6$, $VS_i(\mathbf{M})$ is the i th matrix row sum, K the total number of selected graph fragments, and n_k the number of vertices in the k th fragment. a_{ij} indicates the elements of the adjacency matrix that are equal to 1 for pairs of adjacent vertices and zero otherwise.

Other topological indices can be obtained by using suitable functions applied to $\rightarrow$ *local vertex invariants*; the most common functions are atom and/or bond additive, resulting into descriptors, which correlate well physico-chemical properties, that are atom and/or bond additive themselves. $\rightarrow$ *Zagreb indices* and $\rightarrow$ *ID numbers* are derived according to this approach.

Some functions to derive molecular descriptors $\mathcal{D}$ from local vertex invariants, denoted by $\mathcal{L}$, are listed in Table G3. It should be noted that function $\mathcal{D}_4$, that is, the well-known Randić-type formula for $\alpha = 1$ and $\lambda = -1/2$, is restricted to pairs of adjacent vertices, a_{ij} being the elements of the $\rightarrow$ *adjacency matrix*, which are equal to 1 only for pairs of adjacent vertices and zero otherwise. Function $\mathcal{D}_6$ is an extension of function $\mathcal{D}_4$ to any type of graph fragments as in the $\rightarrow$ *Kier–Hall connectivity indices*. Function $\mathcal{D}_7$ gives $\rightarrow$ *autocorrelation descriptors*, while function $\mathcal{D}_8$ gives $\rightarrow$ *maximum auto-crosscorrelation descriptors*. Moreover, similar functions can be applied to $\rightarrow$ *local edge invariants* $\mathcal{L}_{ij}$ in place of local vertex invariants $\mathcal{L}_i$ so that other sets of molecular descriptors can be generated.

Table G3 Classical functions to derive molecular descriptors from local vertex invariants.

ID	Function	ID	Function
1.	$\mathcal{D}_1(\mathcal{L}; \alpha, \lambda) = \alpha \cdot \sum_{i=1}^A \mathcal{L}_i^\lambda$	5.	$\mathcal{D}_5(\mathcal{L}; \alpha, \lambda) = \alpha \cdot \sum_{i=1}^A \sum_{j=1}^A (\mathcal{L}_i \cdot \mathcal{L}_j)^\lambda \quad j \neq i$
2.	$\mathcal{D}_2(\mathcal{L}; \alpha, \lambda) = \alpha \cdot \left(\prod_{i=1}^A \mathcal{L}_i \right)^\lambda$	6.	$\mathcal{D}_6(\mathcal{L}; \alpha, \lambda) = \alpha \cdot \sum_{k=1}^K \left(\prod_{i=1}^{n_k} \mathcal{L}_i \right)_k^\lambda$
3.	$\mathcal{D}_3(\mathcal{L}; \alpha) = \alpha \cdot \max_{i \in V} (\mathcal{L}_i)$	7.	$\mathcal{D}_7(\mathcal{L}; \alpha, \lambda, k) = \alpha \cdot \sum_{i=1}^A \sum_{j=1}^A (\mathcal{L}_i \cdot \mathcal{L}_j)^\lambda \cdot \delta(d_{ij}; k)$
4.	$\mathcal{D}_4(\mathcal{L}; \alpha, \lambda) = \alpha \cdot \sum_{i=1}^A \sum_{j=1}^A a_{ij} (\mathcal{L}_i \cdot \mathcal{L}_j)^\lambda$	8.	$\mathcal{D}_8(\mathcal{L}; \alpha, \lambda, k) = \alpha \cdot \max_{i,j \in V} [(\mathcal{L}_i \mathcal{L}_j)^\lambda \cdot \delta(d_{ij}; k)]$

$\mathcal{L}_i$ and $\mathcal{L}_j$ are local invariants associated with the vertices v_i and v_j , respectively. A is the number of graph vertices, V denotes the set of graph vertices, and $\delta(d_{ij}; k)$ is a Dirac delta function equal to 1 for pairs of vertices at topological distance d_{ij} equal to k and zero otherwise. In function $\mathcal{D}_4$, a_{ij} indicates the elements of the adjacency matrix, which are equal to 1 for pairs of adjacent vertices and zero otherwise. In function $\mathcal{D}_6$, the summation goes over fragments of a given type, K is the total number of selected graph fragments, and n_k is the number of vertices in the k th fragment.

Another way to derive topological indices is by generalizing the existing indices or graph-theoretical matrices. $\rightarrow$ *Kier–Hall connectivity indices*, $\rightarrow$ *higher order Wiener numbers*, $\rightarrow$ *generalized Wiener indices*, $\rightarrow$ *variable Zagreb indices*, $\rightarrow$ *generalized expanded Wiener numbers*, and $\rightarrow$ *generalized Hosoya indices* are all examples of the generalization of existing indices, while $\rightarrow$ *generalized distance matrix*, $\rightarrow$ *expanded distance matrices*, and $\rightarrow$ *graphical matrices* are examples of generalized matrices.

Several $\rightarrow$ *fragment topological indices* can be derived by any topological index calculated for molecular subgraphs.

Particular topological indices are derived from weighted molecular graphs where vertices and/or edges are weighted by quantities representing some 3D features of the molecule, like those obtained by $\rightarrow$ *computational chemistry*. The graph invariants obtained in this way encode both information on molecular topology and $\rightarrow$ *molecular geometry*. Examples of these topological descriptors are $\rightarrow$ *BCUT descriptors*, $\rightarrow$ *electronic-topological descriptors*, $\rightarrow$ *electron charge density connectivity index*, and several descriptors obtained from $\rightarrow$ *weighted matrices*.

$\rightarrow$ *Triplet topological indices* were proposed based on a general matrix–vector multiplication approach and several $\rightarrow$ *combined descriptors* are combinations of existing descriptors.

Several graph invariants can also be derived by the **vector–matrix–vector multiplication approach** (or **VMV approach**) proposed by Estrada [Estrada, Rodríguez *et al.*, 1997; Estrada and Rodríguez, 1997; Estrada, 2001; Estrada and Gutierrez, 2001]. This approach allows to generate graph invariants $\mathcal{D}$ according to the following equation:

$$\mathcal{D}(\mathbf{M}, \mathbf{v}_1, \mathbf{v}_2; \alpha, \lambda) = \alpha \cdot (\mathbf{v}_1^T \cdot \mathbf{M}^\lambda \cdot \mathbf{v}_2)$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are two column vectors collecting atomic properties or local vertex invariants, $\mathbf{M}$ is a graph-theoretical matrix, and α and λ are two real parameters. Examples of well-known molecular descriptors derived from the VMV approach are reported in Table G4; moreover, all the $\rightarrow$ *TOMOCOMD descriptors* are calculated by this approach.

Table G4 Examples of molecular descriptors derived from VMV approach.

M	v_1	v_2	λ	α	Descriptor
D	1	1	1	1/2	Wiener index, W
I	δ	δ	1	1	First Zagreb index, M_1
A	δ	δ	1	1	Second Zagreb index, M_2
D	1	1	-1	1/2	Harary index, H
A	$\delta^{-0.5}$	$\delta^{-0.5}$	1	1/2	Randić connectivity index, ${}^1\chi$
A	$\sigma^{-0.5}$	$\sigma^{-0.5}$	1	$(1/2) [B/(C + 1)]$	Balaban distance connectivity index, J

D is the $\rightarrow$ distance matrix, I the $\rightarrow$ identity matrix, and A the $\rightarrow$ adjacency matrix. δ indicates the $\rightarrow$ vertex degree and σ the $\rightarrow$ distance degree; B is the number of graph edges and C the number of rings.

Another general procedure to generate graph invariants is that used to calculate the so-called **Molecular Descriptor Family (MDF)** [Jäntschi, 2004a, 2004b, 2005; Bolboacă and Jäntschi, 2005b, 2005, 2006, 2007]. This procedure utilizes both topological and geometrical distances between atoms, 6 atomic properties as the weighting scheme for graph vertices, 24 formulas for interaction descriptors, 6 overlapping interaction models, 4 fragmentation criteria, and 19 fragmental property selector functions. To all 131 328 resulted values, 6 linearization operations are applied, and finally it results in a number of 787968 MDF values for a given molecule.

Graph invariants have been successfully applied in characterizing the structural similarity/dissimilarity of molecules and in QSAR/QSPR modeling.

Due to the large proliferation of graph invariants, the result of many authors following the procedures outlined above and other general schemes, some rules are needed to critically analyze such invariants, paying particular attention to their effective role in correlating physico-chemical properties, biological and other experimental responses, and their chemical meaning. In this respect, a list of desirable attributes for topological indices was suggested by Randić [Randić, 1991b] (see Table M11).

Theoretical studies on variability (e.g., covariance) and correlation of graph invariants were presented by Hollas [Hollas, 2002, 2003, 2005a, 2005b, 2005c, 2006; Hollas, Gutman *et al.*, 2005].

 Additional references are listed in the thematic bibliography (see Introduction).

➤ **graph kernel** $\equiv$ pseudocenter $\rightarrow$ center of a graph

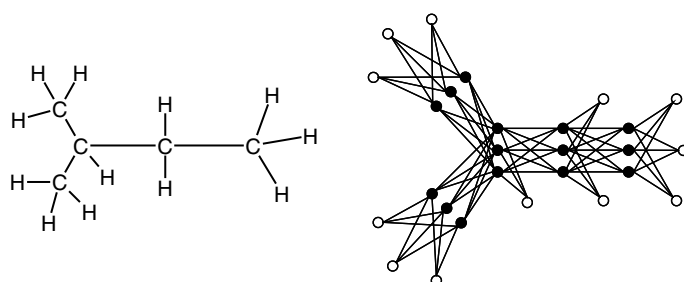
■ Graph of Atomic Orbitals (GAO)

GAO is a molecular representation that accounts for the electron configuration of different atoms in the molecule. It is defined as a molecular graph where each vertex represents a group of atomic orbitals of the respective atom [Mercader, Castro *et al.*, 2000; Toropov and Toropova, 2000a, 2000b, 2001b; Toropova and Toropov, 2000].

Let G be the $\rightarrow$ H -filled molecular graph of a molecule and $V(G) = \{v_1, v_2, \dots, v_A\}$ be the set of vertices in G , then the graph of atomic orbitals (GAO) is obtained from G by replacing each of its vertex v_i with a set of n_i distinct vertices, the value n_i depending on the type of atom (Table G5). Two vertices in the GAO are adjacent if and only if they correspond to two different and adjacent atoms. Consequently, two vertices in the GAO, representing different groups of orbitals of the same atom, are not adjacent (Figure G4).

Table G5 Groups of atomic orbital for the most frequently occurring atoms in organic molecules.

Atom	Groups of atomic orbitals	<i>n</i>
H	1s ¹	1
C	1s ¹ 2s ² 2p ²	3
N	1s ¹ 2s ² 2p ³	3
O	1s ¹ 2s ² 2p ⁴	3
F	1s ¹ 2s ² 2p ⁵	3
S	1s ¹ 2s ² 2p ⁶ 3s ² 3p ⁴	5
Cl	1s ¹ 2s ² 2p ⁶ 3s ² 3p ⁵	5
Br	1s ¹ 2s ² 2p ⁶ 3s ² 3p ⁶ 3d ¹⁰ 4s ² 4p ⁵	8
I	1s ¹ 2s ² 2p ⁶ 3s ² 3p ⁶ 3d ¹⁰ 4s ² 4p ⁶ 4d ¹⁰ 5s ² 5p ⁵	11

**Figure G4** Graph of atomic orbitals (GAO) of 2-methylbutane.

GAO is an orbital-based graph-theoretical representation of molecules from which the common $\rightarrow$ *graph invariants*, such as connectivity, Zagreb, and Wiener indices, can be calculated, and thus it represents a source of orbital-based molecular descriptors, which can be generally called **GAO descriptors**. Note that $\rightarrow$ *orbital interaction graph of linked atoms* is another representation of molecules, which accounts for atom orbitals.

📖 [Toropov and Toropova, 2001b, 2003; Toropov, Toropova *et al.*, 2003, 2004]

- **graph potentials** $\rightarrow$ MPR approach
- **graph radius** $\rightarrow$ biodescriptors ($\odot$ DNA sequences)
- **graph-theoretical invariants** $\equiv$ *graph invariants*
- **graph-theoretical matrices** $\rightarrow$ matrices of molecules
- **graph-theoretical shape coefficient** $\rightarrow$ shape descriptors ($\odot$ Petitjean shape indices)
- **graph theory** $\rightarrow$ graph
- **graph valence shells** $\equiv$ *valence shell counts* $\rightarrow$ path counts
- **graph vertex complexity** $\rightarrow$ topological information indices
- **graph walk count** $\equiv$ *molecular walk count* $\rightarrow$ walk counts
- **gravitational indices** $\rightarrow$ size descriptors
- **Green resonance energy** $\rightarrow$ delocalization degree indices
- **grid** $\rightarrow$ grid-based QSAR techniques
- **GRID/GOLPE method** $\rightarrow$ grid-based QSAR techniques ($\odot$ GRID method)

- **GRID method** → grid-based QSAR techniques
- **grid-based QSAR model** → grid-based QSAR techniques

■ grid-based QSAR techniques

These are QSAR techniques based on molecular descriptors calculated by embedding compounds into a fixed grid and encoding information about → *molecular interaction fields* (MIF) or physico-chemical properties related to ligand–receptor binding interactions [Esposito, Hopfinger *et al.*, 2003; Kubinyi, 2003a; Goodford, 2006].

Grid-based techniques could be used to model a variety of biological and physico-chemical properties; their most common application has, by far, been focused on ligand-target binding properties; moreover, grid-based descriptors can be compared to measure → *similarity/diversity* of molecules. Grid-based descriptors mainly characterize molecular shape and charge distribution in the 3D space responsible for nonbonding interactions involved in ligand–receptor binding. Therefore, they give the possibility of representing the molecular interaction regions of interest graphically, a big advantage in pharmacological studies.

The basic starting steps in grid-based techniques are the generation of three-dimensional structures of molecules, conformational search, and, in most of the cases, alignment of all the data set molecules according to some → *alignment rules*; alignment can be among the molecules themselves or with respect to a reference compound or a → *pharmacophore*.

Once the molecules have been aligned, a rigid orthogonal grid of regularly spaced points representing an approximation of the binding site cavity space is established around each compound. A **grid** is a regular 3D array of $N_x \times N_y \times N_z$ points (N_G), that is, a lattice of grid points with N_x points along the X-axis, N_y points along the Y-axis, and N_z points along the Z-axis, each point **p** being characterized by the Cartesian coordinates (x, y, z) in the 3D space. The grid can be chosen to embed all the atoms of all compounds of the data set or else cover common particular regions of interest in the compounds. The density of the grid must be such as to sample the potential energies of the theoretically continuous scalar field reliably. A density sampling about 0.25–0.50 Å for sharp fields like molecular electrostatic potential seems to preserve field invariance [Todeschini, Moro *et al.*, 1997]. Steps of 2 Å are the most commonly used; however, a finer grid was suggested to obtain better predictive models [Liljefors, 1998].

Then, for each molecule embedded in the grid, specific values are calculated at every grid point; these, usually, are molecular **interaction energy values** or some function of them, taken as molecular descriptors within the framework of grid-based QSAR techniques. Usually, a reasonable selection of interaction energy values is performed based on energy cut-off values, selected molecular regions, or other specific criteria, depending on the method.

Finally, the classical **grid-based QSAR model** is estimated from interaction energy values. Linear models are the most common and take the following general form:

$$\hat{y}_i = b_0 + \sum_{j=1}^F \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} \sum_{z=1}^{N_z} b_{j,xyz} \cdot E_{ij,xyz} = \sum_{j=1}^F \sum_{k=1}^{N_G} b_{jk} \cdot E_{ijk}$$

where F is the number of fields used in the analysis, that is, the number of molecular interaction fields; N_x , N_y and N_z are the number of grid points along the X-axis, Y-axis, and Z-axis, respectively; $N_G = N_x \times N_y \times N_z$ is the total number of grid points; and $E_{ij,xyz}$ is the potential interaction energy of the i th compound for the j th field in the grid coordinate (x, y, z). The k index of the last summation runs over the grid points in a vectorial representation (Figure G5).

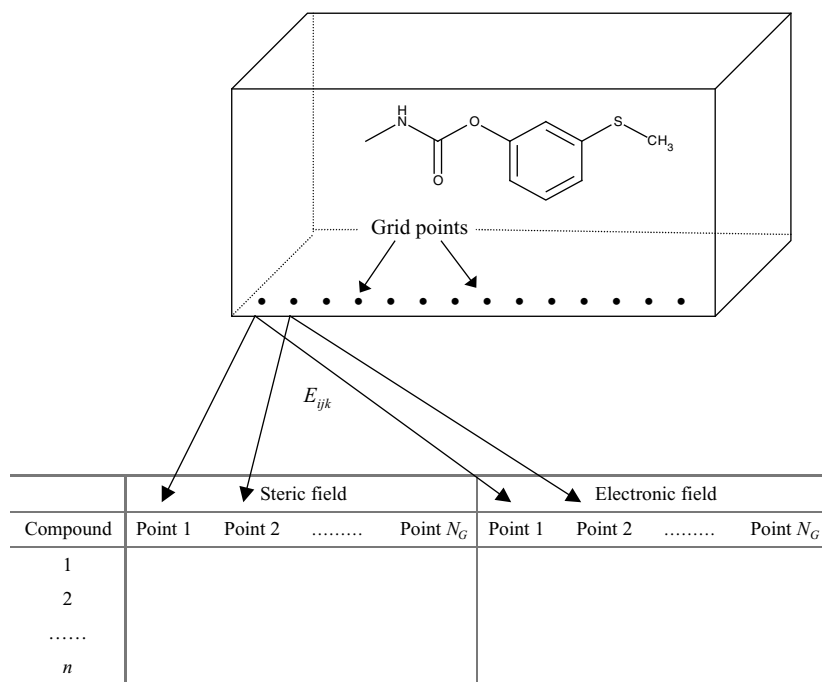


Figure G5 Construction of the data set matrix in grid-based QSAR techniques.

Due to the large number of descriptors (commonly 15 000–20 000 for each field), multivariate regression analysis is usually performed by partial least squares regression (PLS), with or without $\rightarrow$ *variable selection*. Alternatively to grid-based QSAR models, $\rightarrow$ *similarity/diversity* between molecules can be measured by comparing their interaction fields.

The most popular grid-based QSAR techniques are *GRID* and *CoMFA*, based on molecular interaction fields derived from different probes; these methods together with a number of related techniques are discussed below; other related techniques are $\rightarrow$ *hydration free energy density* and $\rightarrow$ *Comparative Molecular Surface Analysis*.

A critical point of most of the grid-based techniques is the alignment of molecules, which determines to what extent the descriptors differ from one molecule to the next. Consequently, it substantially influences the results of the analysis. Hence, significant and reliable results can only be expected if the alignment was carried out properly and unambiguously. Often, the need for an alignment limits the application of these grid-based descriptors to homogeneous data sets, and even, then, the alignment is not always easily performed. To overcome this drawback, different research groups started to develop alignment-independent molecular descriptors. The first alignment-independent descriptors derived from scalar fields are *G-WHIM descriptors* [Todeschini, Moro *et al.*, 1997], based on the theoretical principles of the $\rightarrow$ *WHIM descriptors* but applied to $\rightarrow$ *molecular interaction fields*. *VolSurf* [Cruciani, Pastor *et al.*, 2000] and *GRIND* [Pastor, Cruciani *et al.*, 2000] descriptors are other grid-based descriptors independent of any previous alignment of the molecules.

• GRID method

This method is a computational procedure for detecting favorable binding sites on a molecule of known structure [Goodford, 1985, 1995]. A small molecule, such as water (the $\rightarrow$ *probe*) is used to generate $\rightarrow$ *interaction energy values* at all the grid points. The probe is rotated at the grid point until it makes the most favorable energetic interactions with the target. The final 3D array of energies constitutes the *GRID Map* for that particular probe. These energies are calculated by a sophisticated Empirical Force-Field method (*GRID Force Field*). Typically, grid spacing of 0.5 Å is used.

The main advantage of the GRID method is the great variety of available probes, represented by several functional groups such as water, methyl, ammonium, carboxylate, and benzene; in particular, among them there are probes that can both accept and donate a hydrogen bond (e.g., water), probes that cannot turn around (e.g., carbonyl probe), and a hydrophobic probe, named DRY.

The GRID method, unlike other grid-based techniques, explicitly takes the flexibility of the molecule into account in ligand–receptor interactions. The conformational flexibility of compounds is studied, allowing them to be attracted or repelled by the probe as the probe is moving around [Liljefors, 1998]. This algorithm works by dividing the target molecule into a flexible core and flexible side chain on an atom basis.

The **GRID/GOLPE method** is a QSAR methodology that combines interaction fields calculated by GRID with statistical analysis implemented in the program $\rightarrow$ *GOLPE* [Baroni, Clementi *et al.*, 1992; Baroni, Costantino *et al.*, 1993a, 1993b; GOLPE – Multivariate Infometric Analysis s.r.l., 2007]. Applications of GRID/GOLPE method are [Cruciani, Clementi *et al.*, 1993, 1994, 1998; Cruciani and Watson, 1994; Norinder, 1996b; Cruciani, Pastor *et al.*, 1997; Pastor, Cruciani *et al.*, 1997; Cinone, Hölte *et al.*, 2000; Fichera, Cruciani *et al.*, 2000; Lozano, Pastor *et al.*, 2000; Sippl, 2006].

 Additional references are listed in the thematic bibliography (see Introduction).

• Comparative Molecular Field Analysis ($\equiv$ CoMFA)

This is the most popular QSAR approach among the grid-based QSAR techniques. CoMFA compares the molecular potential energy fields of a set of molecules and searches for differences and similarities that can be correlated with differences and similarities in the property values considered [Cramer III, Patterson *et al.*, 1988; Marshall and Cramer III, 1988; Cramer III, DePriest *et al.*, 1993; Folkers, Merz *et al.*, 1993a, 1993b; Kim, 1995a; Oprea and Waller, 1997; Martin, 1998; Norinder, 1998; Kubinyi, 2003a].

The first step of the CoMFA approach consists in the selection of a group of compounds having a common $\rightarrow$ *pharmacophore*, in the generation of three-dimensional structures of reasonable conformation and in their alignment.

The grid established around each compound is referred to the **CoMFA lattice**; the grid point distance is arbitrarily chosen (2 Å by default), bearing in mind that even small desirable distances lead to too great a number of grid points; the walls of the grid usually extend at least 4 Å beyond the union volume of the superimposed molecules. The rigidity of receptor walls derived from the use of a rigid grid is a basic assumption and approximation in the CoMFA method.

In the original CoMFA method only two fields of noncovalent ligand–receptor interactions were calculated: the steric field that is a $\rightarrow$ *Lennard-Jones 6–12 potential function* and the electrostatic field that is a $\rightarrow$ *Coulomb potential energy function*. Usually, the two fields are kept separate to facilitate the interpretation of the final results. As steric and electrostatic

fields account only for enthalpic contributions to free binding energy, other fields that account for solvation and entropic terms should be added. For example, hydrophobic interactions, which are entropic properties, are accounted for by the use of $\rightarrow$ *HINT* and $\rightarrow$ *molecular lipophilicity potential*. Since the Lennard-Jones potential is characterized by very steep increases in energy at short distances from the molecular surface, it was proposed to use the van der Waals volume intersections between probe and ligand molecule for steric field calculation [Sulea, Oprea *et al.*, 1997]; this molecular interaction field was called intersection volume field (INVOL).

Interaction energy values at the grid points are the **CoMFA descriptors** and are collected into a QSAR matrix whose rows represent the molecules and columns the grid points for each field considered.

Unreasonably large positive energy values, that is, grid points inside the molecules, are set constant at chosen cutoff value. They mainly derive from the large values of van der Waals repulsion caused by even a small overlap of ligand atoms and probe atoms. Moreover, grid points without variance, that is, within the volume shared by all molecules, or with small variance, that is, far away from the van der Waals surface of molecules, are discarded. Moreover, other parameters such as $\rightarrow$ *log P* or $\rightarrow$ *quantum-chemical descriptors* can be added as variables to the QSAR matrix after properly scaling. The combination of global parameters and CoMFA fields leads to a **mixed CoMFA approach** [Kubinyi, 1993b].

The **mixed CoMFA model** is defined as

$$\hat{y}_i = b_0 + \sum_{j=1}^F \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} \sum_{z=1}^{N_z} b_{j,xyz} \cdot E_{ij,xyz} + \sum_{j=1}^J b_j \cdot \Phi_{ij} = \sum_{j=1}^F \sum_{k=1}^{N_G} b_{jk} \cdot E_{ijk} + \sum_{j=1}^J b_j \cdot \Phi_{ij}$$

where Φ_{ij} is any global molecular property and J the total number of molecular properties considered.

Partial least squares (PLS) regression is usually performed to search a correlation between the thousands of CoMFA descriptors and biological responses. However, usually after $\rightarrow$ *variable selection*, the PLS model is transformed into and presented as a multiple regression equation to allow comparison with classical QSAR models.

Developments of the CoMFA approach have been also proposed based on a selection of molecule regions of interest for binding interactions [Cho and Tropsha, 1995; Cruciani, Clementi *et al.*, 1998; Tropsha and Cho, 1998].

A critical review of CoMFA applications is given in [Kim, Greco *et al.*, 1998] and a complete list of references 1993–1997 in [Kim, 1998].

📖 Additional references are listed in the thematic bibliography (see Introduction).

• Comparative Molecular Similarity Indices Analysis ($\equiv$ CoMSIA)

CoMSIA is a method of measuring the similarity of molecules on the basis of their physico-chemical properties. It implements the steric, electrostatic, hydrophobic, and hydrogen-bonding $\rightarrow$ *similarity indices* utilized in the molecular alignment program SEAL [Abraham *et al.*, 1994; Diudea, 1997d; Klebe, 1998; Klebe and Abraham, 1999].

Using the $\rightarrow$ *similarity score* A_F based on the weighted combination of steric, electrostatic, and hydrophobic properties, molecule alignment is performed starting from a random orientation of two molecules relative to each other; the best alignment is achieved with the maximum similarity score.

Moreover, $\rightarrow$ *molecular interaction fields* are calculated for each molecule in terms of similarity indices instead of the usual interaction potential functions, such as Lennard-Jones and Coulomb potential functions. Similarity fields are calculated representing the similarity between molecules and different probe atoms. In particular, the similarity values at the intersections of the regularly spaced grid (1.1 and 2.0 Å) relative to the j th physico-chemical property between the i th compound and a probe atom is calculated as

$$(A_F)_{ik,j} = \sum_t w_{probe,j} \cdot w_{tj} \cdot e^{-\alpha \cdot r_{ik}^2}$$

where the summation goes over all atoms of the molecule, $w_{probe,j}$ and w_{tj} are, respectively, the actual value of the j th property of the probe and the t th atom of the target molecule, α is an attenuation factor, and r_{ik} is the geometric distance between the probe atom at the k th grid point and the i th atom of the molecule. Large values of α give rise to a strong distance-dependent attenuation of the similarity measures, that is, only local similarities are considered; otherwise, for small α values, global molecular features are of greater importance. The probe interaction with the molecule is, then, calculated for each grid point, including those inside the molecule atomic van der Waals radius, avoiding the need for cutoff as in CoMFA.

The studied properties are electrostatic, steric, hydrophobic, hydrogen-acceptor, and hydrogen-donor abilities; for electrostatic properties, the probe assumes charge +1, for steric properties radius 1 Å, for hydrophobicity and hydrogen-bonding abilities a value of +1.

These indices replace the distance functions used in the standard Lennard-Jones and Coulomb potential functions, which generate unrealistically extreme values as the surface of the considered molecule is approached.

CoMSIA results show regions of the compound that prefer or dislike the presence of a group with a specific physico-chemical property.

A molecular $\rightarrow$ *similarity matrix* can be obtained both from the similarity scores between pairs of molecule and any distance function applied to similarity fields [Klebe, Abraham *et al.*, 1994; Diudea, 1997d; Klebe, 1998; Kubinyi, Hamprecht *et al.*, 1998].

📖 [Hou, Li *et al.*, 2000; Anzini, Cappelli *et al.*, 2001; Ducrot, Andrianjara *et al.*, 2001; Makhija and Kulkarni, 2001b, 2002a; Zhu, Hou *et al.*, 2001; Buolamwini and Assefa, 2002; Doytchinova and Flower, 2002; Schleifer and Tot, 2002; Sreenivasa and Kulkarni, 2002; Wellsow, Machulla *et al.*, 2002; Xu, Zhang *et al.*, 2002; Assefa, Kamath *et al.*, 2003; Boström, Böhm *et al.*, 2003; Bringmann and Rummey, 2003; Liu, Yang *et al.*, 2003; Raichurkar and Kulkarni, 2003; Chen, Yao *et al.*, 2004; Kelkar, Pednekar *et al.*, 2004; Medina-Franco, Rodríguez-Morales *et al.*, 2004; Sutherland and Weaver, 2004; Jójárt, Martinek *et al.*, 2005; Zhao, Yu *et al.*, 2005]

• **G-WHIM descriptors** ($\equiv$ *Grid-Weighted Holistic Invariant Molecular descriptors*)

Based on a similar approach to that used to define $\rightarrow$ *WHIM descriptors*, G-WHIM descriptors are global molecular descriptors of $\rightarrow$ *molecular interaction fields* [Todeschini, Moro *et al.*, 1997; Todeschini and Gramatica, 1998].

Once the optimal choices for the grid have been made, the G-WHIM descriptors are used to condense the whole information contained in the scalar field constituted by the calculated $\rightarrow$ *interaction energy values* into a few global parameters, whose values are independent of the molecular orientation within the grid.

For each molecule, the G-WHIM descriptors are calculated by the following steps:

- The molecule is freely and separately embedded in the center of the grid.
- The interaction energy values are calculated at all the grid points by using the selected probe.
- The interaction energy values are used as weighting scheme for the grid points. This is the main difference between G-WHIM and WHIM descriptors, for which the defined atomic properties are used to weight molecule atoms.
- Finally, the G-WHIM descriptors are calculated in the same way as the WHIM descriptors, that is, by the calculation of a weighted covariance matrix, principal component analysis, and the calculation of statistical parameters on the projected points along each principal component (i.e., on the score values) (Figure G6).

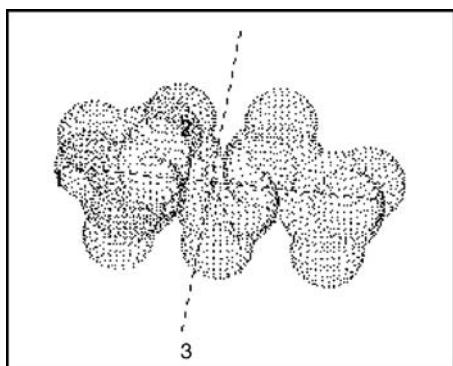


Figure G6 Principal axes of the grid interaction energy values of the 2-methylpentane.

It should be noted that only points with nonzero interaction energy are effective in the computation of the descriptors. Second, when the calculated interactions give both positive and negative values, interaction energy values cannot be used directly in this form as statistical weights, which must always be defined semipositive. In this case, the scalar field values are divided into two blocks: a grid-negative (positive) block containing the grid coordinates associated with negative (positive) interaction values, getting their absolute values and setting the positive (negative) values to zero.

This assumption leads to two sets of G-WHIM descriptors: one describing the positive part of the molecular field and the other describing the negative part.

Thus, for each region (positive (+) and negative (-)), the G-WHIM descriptors consist of 8 directional and 5 nondirectional molecular descriptors (26 for a complete description of each interaction field), calculated from each molecule:

$$\lambda_1(\pm) \quad \lambda_2(\pm) \quad \lambda_3(\pm) \quad \vartheta_1(\pm) \quad \vartheta_2(\pm) \quad \eta_1(\pm) \quad \eta_2(\pm) \quad \eta_3(\pm)$$

and

$$T(\pm), \quad A(\pm), \quad V(\pm), \quad K(\pm), \quad D(\pm)$$

The directional γ and the nondirectional G parameters, defined for $\rightarrow$ WHIM descriptors and containing information about the molecular symmetry, are not considered in the frame of the G-

WHIM approach as their meaning becomes doubtful, depending heavily upon the point sampling. However, information regarding molecular symmetry can be obtained by directly using the WHIM symmetry parameters.

The meaning of the G-WHIM descriptors is that previously defined for WHIM descriptors, but now the descriptors refer to the interaction molecular field instead of the molecule. For example, the eigenvalues λ_1 , λ_2 , and λ_3 relate to the interaction field size; the eigenvalue proportions ϑ_1 and ϑ_2 relate to the interaction field shape; the group of descriptors constituted by the inverse function of the kurtosis (k), that is, $\eta_m = 1/\kappa_m$ relate to the interaction field density along each axis. Moreover, global information about the interaction field is obtained as for $\rightarrow$ *global WHIM descriptors* (T, A, V, K, D), with the same meaning.

First [Todeschini, Moro *et al.*, 1997], the acentric factor was used as the shape descriptor, defined as $\omega = \vartheta_1 - \vartheta_3$, ranging from zero (spherical interaction field) to one (linear interaction field); this shape descriptor was later substituted by the global K shape descriptor.

It must be noted that the invariance to rotation of G-WHIM descriptors, that is, the independence of any molecular $\rightarrow$ *alignment rule*, is obtained if the grid points are dense enough. A too sparse distribution of grid points represents an inadequate sampling of the ideal scalar field and is not able to guarantee that the calculated scalar field is representative of the ideal scalar field in such a way as to preserve rotational invariance. If a molecule is placed into an infinite, isotropic, and even very dense grid, the scalar field F calculated at the grid points must contain the same information, independent of the molecule's orientation and depending only on the potential energy of the selected probe and the mathematical function representing the interaction. Thus F contains the whole information about the interaction properties of the molecule.

In practice, this ideal situation cannot be achieved, but it can be simulated by plunging the molecule into a finite grid: the aim is to represent the theoretical scalar field F by a finite sampling of this field.

G-WHIM descriptors can be calculated for any selected region of the field. To avoid irrelevant or unreliable chemical information due to long-range interactions, an energy cutoff criterion, which takes into account only interaction energy values relevant to the considered interaction (e.g., long-range or chemical interaction) is used. In this way, points far from the molecule and not contributing to the interaction are not included in the calculations, for example, the field values inside the van der Waals surface of the molecule are not considered. Moreover, specific chemical information is gained by using different energy cutoff values to select the regions, keeping in mind that the higher the cutoff value the smaller the considered region around the molecule (i.e., the total number of nonzero weighted grid points). For instance, surface points at a given cutoff value, that is, points on an isopotential energy surface, can be selected. G-WHIM descriptors calculated on the $\rightarrow$ *Connolly surface area* were called **MS-WHIM descriptors** [Bravi, Gancía *et al.*, 1997].

The ability to take the individual parameters provided by different cutoff values into account when defining different molecular interaction regions could possibly lead to a deeper chemical insight into molecular interactions and properties.

The G-WHIM approach integrates the information contained in $\rightarrow$ *WHIM descriptors* and overcomes any problem due to the alignment of the different molecules and the explosion of variables arising from traditional grid-based QSAR techniques, such as GRID and CoMFA.

📖 [Ekins, Bravi *et al.*, 1999a, 1999b, 1999c; Zaliani and Gancia, 1999; Bravi and Wikel, 2000a, 2000b; Cosentino, Moro *et al.*, 2000; Gancia, Bravi *et al.*, 2000; Ekins, Durst *et al.*, 2001; Baumann, 2002b; Snyder, Sangar *et al.*, 2002]

• **Self-Organizing Molecular Field Analysis ($\equiv$ SOMFA)**

SOMFA is a grid-based approach that does not use a probe to determine interaction energies. Instead, each grid point is assigned the shape or $\rightarrow$ molecular electrostatic potential (MEP) value: (a) shape is represented by binary values equal to 1 for points inside the van der Waals envelope and zero otherwise; (b) electrostatic potential values at grid points are calculated from partial charges distributed across the atom centers [Robinson, Winn *et al.*, 1999].

Crucial to SOMFA is the concept of *mean centered activity* (A_i^c), which is the activity of a molecule obtained by subtracting the mean activity of the training set molecules. Therefore, the molecule activity has a scale such that the most active compounds have positive values and the least active ones have negative values. The mean centered activity is used as the weighting scheme for the grid points: the value of the shape or electrostatic potential at every grid point for a given molecule is multiplied by the mean centered activity. This procedure allows grid points to filter in such a way as to highlight the features that differentiate high-affinity and low-affinity compounds.

In general, a SOMFA master grid can be trained on any calculable molecular property that can be distributed in a grid by using all the training compounds. This grid is constructed by overlaying all the individual molecular grids; the value at the grid point (x, y, z) of the master grid is defined as

$$SOMFA(x, y, z) = \sum_{i=1}^n P_i(x, y, z) \cdot A_i^c$$

where n is the number of training compounds, P_i is the property of the i th compound at grid point (x, y, z), and A_i^c is its mean centered activity. Maximum and minimum grid values of the master grid can be displayed to highlight regions favorable or unfavorable to activity.

Finally, for a given property P , a SOMFA molecular descriptor can be calculated for each molecule as

$$SOMFA_i(P) = \sum_x \sum_y \sum_z P_i(x, y, z) \cdot SOMFA(x, y, z)$$

where $P_i(x, y, z)$ is the property value for the i th molecule at point (x, y, z) and $SOMFA(x, y, z)$ is the value of the master grid at the same point.

A linear combination of the activities calculated from shape and MEP properties was also proposed as a model for the better prediction of activity:

$$\hat{A}_i = \alpha \cdot A_i^{SH} + (1-\alpha) \cdot A_i^{MEP}$$

where α is a parameter that can be optimized to maximize the predictive power of the model.

SOMFA, like other grid-based techniques, needs a reliable procedure for molecule alignment and suffers from the need for the bioactive conformations.

📖 [Amat, Besalú *et al.*, 2001; Liu, Yin *et al.*, 2001b; Klein, Kaiblinger *et al.*, 2002; Klocker, Wailzer *et al.*, 2002a; Smith, Sorich *et al.*, 2003; Tuppurainen, Viisas *et al.*, 2004; Martinek, Ötvös *et al.*, 2005]

• **Voronoi Field Analysis** ($\equiv$ VFA)

Among the grid-based QSAR techniques, Voronoi Field Analysis was proposed with the aim of reducing the very large number of potential $\rightarrow$ *interaction energy values* assigned to the grid points of $\rightarrow$ *molecular interaction fields* [Chuman, Karasawa *et al.*, 1998]. Interaction energy values are assigned to each of the **Voronoi polyhedra** into which the superimposed molecular space is decomposed using an approach similar to that used in the $\rightarrow$ *Voronoi binding site models*.

The molecules in the data set are superimposed considering their conformational flexibility and the total volume of the superimposed molecular space is calculated, after expansion with a 4.0 Å of thick shell outside the surface. The total volume is divided into Voronoi polyhedra each including an atom as a reference point. The outer boundaries of the most outer subspaces are not bisecting planes between two reference points.

The reference points are assigned by selecting as the template the simplest molecule or the unsubstituted one, and the position in the template of the atoms including the hydrogens are automatically defined as reference points. As the second step, the largest compound in terms of number of atoms is selected and the atomic positions of this compound are compared with the previous ones: new reference points are defined if no reference point within 1.0 Å of each atom of this molecule is found. The remaining molecules are then selected in order of decreasing size and atomic positions are compared with reference points as in the previous step. As the final step, a Voronoi polyhedron is assigned to each reference point with its own molecular space; the *Voronoi polyhedron* is a region delimited by a set of planes, each of which bisects as well as is perpendicular to the line connecting the reference point with each of the neighboring reference points of the adjacent regions. In other words, each polyhedron is a set of points closer to the reference point than to any other.

After obtaining the decomposition of the superimposed molecular space into Voronoi polyhedra, a grid exactly containing the expanded molecular surface is defined, with grid points spaced at 0.3 Å, and potential energy values are calculated at each of the lattice points located inside the surface. The steric and electrostatic potential energy values calculated at each k th grid point are transformed into the corresponding Voronoi potential energy values VE_g by summing all the contributions of the grid points belonging to the g th Voronoi polyhedron VP_g :

$$VE_g^T = \sum_k E_k^T \quad k \in VP_g$$

where superscript “ T ” denotes any kind of potential energy value (steric, electrostatic, etc.).

 [Aurenhammer, 1991]

• **GRIND descriptors** ($\equiv$ GRId INdependent Descriptors)

These are molecular descriptors derived from $\rightarrow$ *molecular interaction fields* (MIF) calculated by using different $\rightarrow$ *probes* and representing the geometrical relationships among MIF regions [ALMOND – Multivariate Infometric Analysis s.r.l., 2007; Pastor, Cruciani *et al.*, 2000; Cruciani, Pastor *et al.*, 2001b; Gratteri, Cruciani *et al.*, 2001].

The procedure for obtaining GRIND descriptors is threefold: (a) the molecular interaction field (MIF) is computed, (b) the MIF is filtered, considering only the regions in which the intensity of the field is maximum at relative distances, and (c) the GRIND descriptors are calculated on the basis of the maximum value of the property products obtained at different distances.

Molecular interaction fields are calculated according to the → *GRID method* and considering three different probes: the DRY probe for representing hydrophobic interactions, the O probe (carbonyl oxygen) that represents hydrogen-bonding donor groups, and the N1 probe (amide nitrogen) to represent hydrogen-bonding acceptor groups. By default, a grid-spacing of 0.5 Å is used with the grid extending 5 Å beyond a molecule.

In the second step, the most interesting regions are selected as those characterized by favorable interaction (negative) energies. Therefore, the method extracts from each MIF a number of grid points (about 100) that represent favorable probe–ligand interaction regions by using two optimality criteria: the intensity of the field at a grid point and the mutual distances between the chosen grid points.

In the third step, GRIND descriptors are calculated according to the distance between the grid points, basically using auto- (same probe) and cross-correlation (combinations of pairs of different probes) transforms (Table G6).

Table G6 Correlogram types used to generate GRIND descriptors.

Correlogram	Probe 1	Probe 2	Interaction
1	DRY	DRY	Hydrophobic
2	O	O	Hydrogen bond donor
3	N1	N1	Hydrogen bond acceptor
4	DRY	O	Hydrophobic and hydrogen bond donor
5	DRY	N1	Hydrophobic and hydrogen bond acceptor
6	O	N1	Hydrogen bond donor and acceptor

Unlike the classical → *autocorrelation descriptors*, only the highest product of interaction energies per distance bin is stored as GRIND descriptor (**MACC-2 transform**). This difference is responsible for the reversibility of GRIND descriptors. Unlike most of the grid-based methods, GRIND descriptors are also independent of the molecule alignment.

📖 [Afzelius, Masimirembwa *et al.*, 2002; Cruciani, Pastor *et al.*, 2002; Fontaine, Pastor *et al.*, 2003, 2004; Lapinsh, Prusis *et al.*, 2003; Crivori, Zamora *et al.*, 2004; Gratteri, Romanelli *et al.*, 2004; Aureli, Cruciani *et al.*, 2005; Sciabola, Alex *et al.*, 2005; Gedeck, Rohde *et al.*, 2006; Pastor, 2006; Urbano-Cuadrado, Carbó *et al.*, 2007]

• VolSurf descriptors

VolSurf descriptors, such as G-WHIM and GRIND, encode information present in → *molecular interaction fields* (MIF) calculated by the → *GRID* force field parametrization [Crivori, Cruciani *et al.*, 2000; Cruciani, Crivori *et al.*, 2000; Cruciani, Pastor *et al.*, 2000; Mannhold, Berellini *et al.*, 2006].

VolSurf descriptors were designed to compress relevant MIF information into a few alignment-independent descriptors encoding information about molecular size and shape, the overall distribution of hydrophobic and hydrophilic regions and the balance between them (Table G7).

Interaction fields obtained with different probes (H₂O, DRY, O) are analyzed and the volume and surface of the regions that encompass interaction energy values under certain cutoff limits,

together with some additional variables that express their geometrical spatial distribution, are calculated.

From each MIF, a unique framework (a volume and/or a surface) related to specific molecular properties is constructed. Similar to 2D images, each 3D molecular field map is made of a regular lattice of boxes called *voxels*, which represent attractive and repulsive forces between a probe and a molecule. Each voxel is defined by a volume, by a surface, and by an interaction energy value. By contouring the voxels at different energy levels, different images can be obtained. The images are then used to compute the volumes and the surfaces related to the contouring method selected. In the building phase of volumes, the voxels are grouped by a shape function that assigns the value of 1 to voxels, inside an energy range, and 0 to all other voxels. Subsequently, a simple summation over the selected voxels yields back the overall volume for the considered property. For example, when computing a molecular volume, under standard conditions, all voxels with an energy interaction greater than $+0.2$ kcal/mol are marked as 1 and then counted. As a defined volume is associated with each voxel, the total volume is obtained by multiplying the number of selected voxels by their volume. Conversely, when a hydrophilic volume is computed, only the voxels with interaction energy below -0.2 kcal/mol are marked as 1 and then counted.

Table G7 List of VolSurf descriptors [Zamora, Oprea *et al.*, 2001; Mannhold, Berellini *et al.*, 2006].

Symbol	Probe	Meaning
V	H ₂ O	Water-excluded volume (in Å ³) at $+0.2$ kcal/mol energy
S	H ₂ O	Accessible surface of the water interaction field at $+0.2$ kcal/mol energy
R	H ₂ O	Rugosity, defined as the ratio of volume (V) to surface (S)
G	H ₂ O	Globularity, defined as the ratio of surface (S) to the surface area of a sphere with the same volume (V)
W1–W8	H ₂ O	Volume of the hydrophilic interactions at eight different energy levels: -0.2 , -0.5 , -1.0 , -2.0 , -3.0 , -4.0 , -5.0 , -6.0 kcal/mol
IW1–IW8	H ₂ O	Integy moments at the same energy levels as W1–W8
CW1–CW8	H ₂ O	Capacity factors, defined as the ratio of the hydrophilic surface to the total molecular surface (S), at the same energy levels as W1–W8
D1–D8	DRY	Volume of the hydrophobic interactions at eight energy levels: -0.2 , -0.4 , -0.6 , -0.8 , -1.0 , -1.2 , -1.4 , -1.6 kcal/mol
ID1–ID8	DRY	Integy moments at the same energy levels as D1–D8
A	DRY	Strength of the amphiphilic moment
POL	DRY	Polarizability
BV1–BV3	DRY/H ₂ O	Best volumes calculated at the three maximum hydrophobic/hydrophilic regions
Emin1–Emin3	DRY/H ₂ O	Energy values of the three lowest energy minima
D12, D13, D23	DRY/H ₂ O	Distances between the three energy minima
HL	DRY/H ₂ O	Hydrophilic–lipophilic balance, defined as the ratio of the volume of hydrophobic regions at -4 kcal/mol to the volume of hydrophobic regions at -0.8 kcal/mol
CPP	DRY/H ₂ O	Critical packing parameter, defined as the ratio of surface of the hydrophobic regions at -0.6 kcal/mol to the surface of the hydrophilic regions at -3 kcal/mol

(Continued)

Table G7 (Continued)

Symbol	Probe	Meaning
Wp1–Wp8	O	Volume of the interactions with the probe O at eight different energy levels
HB1–HB8	O	Hydrogen-bond donor capacity of the target, defined as the difference between the volume of the hydrophilic interactions (W1–W8) and volume of the O probe interactions (Wp1–Wp8)
E	–	Elongation
EEFR	–	Elongation/elongation-fixed ratio
MW	–	Molecular weight
log <i>P</i>	DRY/H ₂ O	octanol–water partition coefficient
D	H ₂ O	Diffusivity
α	–	polarizability
HBP	Polar	Hydrogen bonding parameter

VolSurf descriptors related to molecular size and shape are: the *molecular volume* represents the water-excluded volume (in Å³), calculated as the volume enclosed by the water-accessible surface computed at a repulsive value of +0.20 kcal/mol; the *molecular surface* represents the accessible surface (in Å²) traced out by a water probe interacting at +0.20 kcal/mol when it rolls over the target molecule; the **rugosity** is a measure of a molecular wrinkled surface defined as the ratio of volume/surface, the smaller the ratio the larger the rugosity; the molecular **globularity** is defined as the ratio of the molecular surface over the surface area of a sphere of the same volume *V*. Globularity is 1.0 for perfect spherical molecules. It assumes values greater than 1.0 for real spheroidal molecules. Globularity is also related to molecular flexibility. **Elongation** represents the maximum extension a molecule could reach if properly stretched. **Elongation/Elongation-Fixed Ratio**, denoted as EEFR, represents the portion of the extension given by the rigid part of the molecule; within each molecule a fixed part is considered as the rigid core, and EEFR is defined as the ratio of the elongation to the elongation of the rigid core.

Another set of VolSurf descriptors consists of descriptors of hydrophilic regions. These are hydrophilic descriptors defined as the volume of the molecular envelope, which is accessible to and attractively interacts with water molecules. The volume of this envelope varies with the level of interaction energies. In general, hydrophilic descriptors computed from molecular fields of –0.2 to –1.0 kcal/mol account for polarizability and dispersion forces, whereas descriptors computed from molecular fields of –1.0 to –6.0 kcal/mol account for polar and H-bond donor–acceptor regions. Moreover, **best hydrophilic volumes** are six molecular descriptors that refer to the best three hydrophilic interactions generated by a water molecule; these are the first, second, and third volumes calculated in separate regions of maximum hydrophilicity. The best volumes are measured at –1.0 and –3.0 kcal/mol. **Capacity factors** are defined as the ratio of the hydrophilic surface over the total molecular surface. Capacity factors are calculated at eight different energy levels, the same levels used to compute the hydrophilic descriptors.

VolSurf descriptors of hydrophobic regions are: *hydrophobic descriptors* defined in terms of the volume of the molecular envelope generating attractive hydrophobic interactions and **best hydrophobic volumes** that represent the best three hydrophobic interactions generated by the DRY probe and measured at –0.6 and –1.0 kcal/mol. VolSurf computes hydrophobic

descriptors at eight different energy levels adapted to the usual energy range of hydrophobic interactions (i.e., from -0.2 to -1.6 kcal/mol).

Integy moments (*INT*eraction enerGY moments), like dipole moments, express the unbalance between the barycenter of a molecule and the center of its hydrophilic or hydrophobic regions. When referring to hydrophilic regions, integy moments are vectors pointing from the center of mass to the center of the hydrophilic regions; when the integy moment is high, there is a clear concentration of hydrated regions in only one part of the molecular surface. Small moments indicate that the polar moieties are either close to the center of mass or they balance at opposite ends of the molecule, so that their resulting barycenter is close to the center of the molecule. When referring to hydrophobic regions, integy moments measure the unbalance between the center of mass of a molecule and the center of the hydrophobic regions. All the integy moments can be visualized in the real 3D molecular space.

Local interaction energy minima are VolSurf descriptors representing the energy of interaction (in kcal/mol) of the best three local energy minima between the water probe and the target molecule. Alternatively, the minima can refer to the three deepest local minima in the $\rightarrow$ *molecular electrostatic potential* (MEP). They are generated both for probes H₂O and DRY. The *energy minima distances* are VolSurf descriptors that represent the distances between all combinations between the best three local energy minima of a molecular interaction field. They are generated both for probes H₂O and DRY. **Hydrophilic–lipophilic balance** is defined as the ratio of the volume of hydrophilic regions measured at -4.0 kcal/mol over the volume of hydrophobic regions measured at -0.8 kcal/mol. The balance describes which effect dominates in the molecule or if they are roughly equally balanced. The $\rightarrow$ *amphiphilic moment* is defined as a vector pointing from the center of the hydrophobic domain to the center of the hydrophilic domain. The vector length is proportional to the strength of the amphiphilic moment, and it may determine the ability of a compound to permeate a membrane. In contrast to the hydrophilic–lipophilic balance, the **critical packing parameter** refers just to molecular shape, being defined as

$$\text{CPP} = \frac{\text{volume}(\text{hydrophobic regions})}{\text{surface}(\text{hydrophilic regions}) \cdot \text{length}(\text{hydrophobic regions})}$$

Lipophilic and hydrophilic calculations are performed at -0.6 and -3.0 kcal/mol, respectively. Critical packing is a good parameter to predict molecular packing such as in micelle formation and may be relevant in solubility studies in which melting point plays an important role.

Other VolSurf descriptors are: the $\rightarrow$ *molecular weight*; the $\rightarrow$ $\log P$, computed via a linear model derived by fitting VolSurf descriptors to experimental data; the $\rightarrow$ *hydrogen bonding parameter* used to describe the H-bonding capacity of a molecular target, as obtained with a polar probe (e.g., the water probe); the **diffusivity**, which controls the dispersion of chemical in water fluid and is calculated according to a modified Stokes–Einstein equation; the $\rightarrow$ *polarizability*, defined as an estimate of the average molecular polarizability and calculated from the structure of a compound according to Cruciani [Cruciani, Crivori *et al.*, 2000] and the additive method of Miller [Miller, 1990b].

📖 [Alifrangis, Christensen *et al.*, 2000; Testa and Bojarski, 2000; Ekins, Durst *et al.*, 2001; Filipponi, Cruciani *et al.*, 2001; Zamora, Oprea *et al.*, 2001, 2003; Cruciani, Pastor *et al.*, 2002; Oprea, 2002b; Oprea, Zamora *et al.*, 2002; Cruciani, Meniconi *et al.*, 2003; Fontaine, Pastor *et al.*, 2003; Menezes, Lopes *et al.*, 2003; Cianchetta, Mannhold *et al.*, 2004; Crivori, Zamora

et al., 2004; Jacobs, 2004; Kovatcheva, Golbraikh *et al.*, 2004; Schefzik, Kibbey *et al.*, 2004; Stærk, Skole *et al.*, 2004; Aureli, Cruciani *et al.*, 2005; Caron and Ermondi, 2005; de Cerqueira Lima, Golbraikh *et al.*, 2006; Lamanna, Catalano *et al.*, 2007]

• 4D-QSAR analysis

4D-QSAR analysis is a grid-based technique that explicitly accounts for ligand conformational flexibility and explores different alignments of compounds [Hopfinger, Wang *et al.*, 1997; Albuquerque, Hopfinger *et al.*, 1998]. With respect to 3D-QSAR analysis, the fourth dimension of 4D-QSAR analysis is the ensemble sampling.

Unlike the other grid-based techniques, molecular descriptors are not derived from → *molecular interaction fields* but from the partition of molecules into different parts that are expected to have different types of interactions with receptor sites.

To generate 4D-QSAR analysis descriptors, the following procedure is used. First, 3D molecular structures are constructed and the conformers of the minimum energy state are used as the initial structures of conformational search. The reference cell grid is constructed to arrange the largest compound in the data set and usually has a grid spacing of 1.0 Å. In the second step, atoms of each molecule are classified into eight types of → *Interaction Pharmacophore Elements* (IPEs) that are the generic type (i.e., any type of atom), nonpolar atom, positively charged atom, negatively charged atom, hydrogen-bond acceptor, hydrogen-bond donor, aromatic atom, and nonhydrogen atom (see Table 4-1 in → *4D-Molecular Similarity Analysis*).

The third step is to estimate the → *Conformational Ensemble Profile* (CEP) for each compound by molecular dynamic simulation; this profile encodes those conformations selected on the basis of the Boltzmann distribution. Then, different alignments are selected to compare the molecules of the training set. In the following step, each conformation of a molecule is placed in the reference grid space on the basis of the alignment scheme being explored and the thermodynamic probability of each grid cell occupied by each IPE type is computed.

The normalized occupancy of each grid cell by each IPE type over the CEP of each molecule, for a given alignment, constitutes a unique set of molecular descriptors referred to as **Grid Cell Occupancy Descriptors** (GCODs). These descriptors were used directly to estimate QSAR models and indirectly in 4D-Molecular Similarity Analysis to generate a set of → *spectral indices*.

📖 [Ravi, Hopfinger *et al.*, 2001; Santos-Filho and Hopfinger, 2001, 2002; Esposito, Hopfinger *et al.*, 2003; Hong and Hopfinger, 2003; Romeiro, Albuquerque *et al.*, 2005]

- **Grid Cell Occupancy Descriptors** → grid-based QSAR techniques (⊙ 4D-QSAR analysis)
- **grid region selection methods** → variable selection
- **GRID electrostatic energy function** → molecular interaction fields (⊙ electrostatic interaction fields)
- **Grid-Weighted Holistic Invariant Molecular descriptors** ≡ *G-WHIM descriptors* → grid-based QSAR techniques
- **GRIND descriptors** → grid-based QSAR techniques
- **Grob inductive constant** → electronic substituent constants (⊙ inductive electronic constants)
- **group charge transfer** ≡ *charge transfer constant* → electronic substituent constants

■ group contribution methods (GCM)

Group contribution methods search for relationships between structural properties and a physico-chemical or biological response based on the following general models:

$$\gamma_i = f(G_1, G_2, \dots, G_m; n_1, n_2, \dots, n_m)$$

where the experimental property γ_i for the i th compound is a function of m group contributions G_j and their occurrences n_j [Reinhard and Drefahl, 1999]. The **group contributions**, also known as **fragmental constants**, are numerical quantities associated with substructures of the molecule, such as single atoms, atom pairs, atom-centered substructures, molecular fragments, functional groups, and so on. The specification of the structural groups depends on the particular GCM scheme adopted. → *Cluster expansion of chemical graphs* is an example of a group contribution method based on all the connected subgraphs of the molecular graph.

Generally, the application of GCM to a molecule requires the following steps:

- 1 Identification of all groups in the molecule applicable to the particular GCM scheme. An automated search for substructures of interest for a given property is performed by the *CASE approach*.
- 2 Calculation of fragmental constants measuring contributions to the molecular property of the considered fragments by employing the function associated with the particular GCM.
- 3 Evaluation of some correction factors that should account for interactions among molecular groups.

Linear GCM models are defined as the following:

$$\gamma_i = k_0 + \sum_{j=1}^m G_j \cdot I_{ij} \quad \text{or} \quad \gamma_i = k_0 + \sum_{j=1}^m G_j \cdot n_{ij}$$

where k_0 is a model-specified constant, j runs over the m group contributions defined within the GCM scheme, G_j is the contribution of the j th group. I_{ij} and n_{ij} are → *substructure descriptors*, and in particular, I_{ij} is a binary variable taking a value equal to 1 if the j th group is present in the i th molecule, zero otherwise, while n_{ij} is the number of times the j th group occurs in the i th molecule.

Nonlinear GCM models are usually defined as

$$\gamma_i = k_0 + \sum_{j=1}^m G_j \cdot n_{ij} - \left(\sum_{j=1}^m G_j \cdot n_{ij} \right)^2$$

Moreover, mixed GCM models are defined by adding, usually, one or more molecular descriptors to the group contributions:

$$\gamma_i = k_0 + \sum_{j=1}^m G_j \cdot n_{ij} + \sum_{j'=1}^p \mathcal{D}_{ij'}$$

where the second summation runs over the p molecular descriptors defined in the GCM scheme and $\mathcal{D}_{ij'}$ is the j' th molecular descriptor value of the i th molecule.

The group contributions G are usually estimated by multivariate regression analysis, but they can also be experimental, theoretical, or user-defined quantities. For example, in the latter case,

the molecular weight can be viewed as a simple linear atom contribution model, where the group contributions are atomic masses. In the first case, large training sets are used to obtain reliable estimates of the group contributions. Usually a battery of group contributions (a field of scalar parameters) is defined taking into account several structural characteristics of the molecules, also sometimes adding extra terms (correction factors) referring to special substructures. If correction factors are considered, the GCM models are usually called **additive-constitutive models**.

Group contribution models were proposed for several molecular property estimations [Zhao, Abraham *et al.*, 2003b], such as boiling and melting points [Wang, Milne *et al.*, 1994; Krzyzaniak, Myrdal *et al.*, 1995; Le and Weers, 1995], $\rightarrow$ molar refractivity [Huggins, 1956; Ghose and Crippen, 1987], pK_a [Perrin, Dempsey *et al.*, 1981; Hilal, Karichoff *et al.*, 1995], critical temperatures, solubilities [Hine and Mookerjee, 1975; Klopman, Wang *et al.*, 1992; Myrdal, Ward *et al.*, 1993; Thomsen, Rasmussen *et al.*, 1999], soil sorption coefficients [Tao, Piao *et al.*, 1999], and several thermodynamic properties [Thinh and Trong, 1976; Yoneda, 1979; Reid, Prausnitz *et al.*, 1988; Suzuki, 2001; Béliveau, Tardif *et al.*, 2003; Béliveau, Lipscomb *et al.*, 2005]. The Rekker method [Nys and Rekker, 1973; Rekker, 1977a] is an example of group contribution method applied to the estimation of $\rightarrow \log P$. Another well-known group contribution model is that proposed by Atkinson for the evaluation of reaction rate constants with hydroxyl radicals of organic compounds [Atkinson, 1987, 1988].

📖 [Smolenskii, 1964; Essam, Kennedy *et al.*, 1977; Ghose and Crippen, 1986; Klein, 1986; Elbro, Fredeslund *et al.*, 1991; Gao, Govind *et al.*, 1992; Drefahl and Reinhard, 1993; Bhattacharjee, 1994; Klopman, Li *et al.*, 1994; Yalkowsky, Dannenfelser *et al.*, 1994; Yalkowsky, Myrdal *et al.*, 1994; Meylan and Howard, 1995; Klein, Schmalz *et al.*, 1999; Platts, Butina *et al.*, 1999; Viswanadhan, Ghose *et al.*, 1999; Wildman and Crippen, 1999; Platts, Abraham *et al.*, 2000]

- **group contributions** $\rightarrow$ group contribution methods
- **group electronegativity** $\rightarrow$ atomic electronegativity
- **group molar refractivity** $\rightarrow$ physico-chemical properties ($\odot$ molar refractivity)
- **GTI** $\equiv$ *Estrada Generalized Topological Indices* $\rightarrow$ variable descriptors
- **GUS index** $\rightarrow$ environmental indices ($\odot$ leaching indices)
- **Gutman index** $\equiv$ *first Zagreb index* $\rightarrow$ Zagreb indices
- **Gutman molecular topological index** $\rightarrow$ Schultz molecular topological index
- **Gutmann's acceptor number** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- **Gutmann's donor number** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- **GVW drug-like indices** $\rightarrow$ property filters ($\odot$ drug-like indices)
- **G-WHIM descriptors** $\rightarrow$ grid-based QSAR techniques

H

- **HACA index** → charged partial surface area descriptors
- **Hadamard matrix product** → algebraic operators
- **hafnian** → algebraic operators (⊙ determinant)
- **half-life time** → environmental indices
- **HLOGP** → lipophilicity descriptors
- **Hamann similarity coefficient** → similarity/diversity (⊙ Table S9)
- **Hamiltonian circuit** → graph
- **Hamiltonian path** → graph
- **Hammett electronic constant** → Hammett equation

■ Hammett equation

Proposed by Hammett in 1937 [Hammett, 1937, 1938; Johnson, 1973], the Hammett equation is defined for the rate constants k and equilibrium constants K of reactions of *meta*- and *para*-substituted benzoic acid derivatives:

$$\rho \cdot \sigma = \log \left(\frac{k_X}{k_H} \right) \quad \text{and} \quad \rho \cdot \sigma = \log \left(\frac{K_X}{K_H} \right)$$

where the constants k_H and K_H refer to an unsubstituted compound (i.e., with hydrogen in the substitution site), while k_X and K_X refer to a *meta*- or *para*-X-substituted compound. *Ortho*-substituents are less used as the electronic effect could be complicated by steric interactions.

The substituent constant σ , called **Hammett electronic constant**, depends only on the nature and position of the substituent and is related to the electronic effect the substituent has on the rate or equilibrium of the reaction, relative to the unsubstituted compound.

The **reaction constant** ρ depends upon the reaction, the conditions under which it is studied and the nature of the reaction series. The magnitude of ρ gives the susceptibility of a given reaction to polar substituents. Large positive values are obtained from all base-catalyzed reactions, while for acid-catalyzed reactions ρ values are of variable sign but in all cases quite small.

Therefore, σ is an electronic descriptor of the substituent estimated by measured rate or equilibrium constants of a reaction, under the control parameter ρ .

From the Hammett equation, several $\sigma \rightarrow$ *electronic substituent constants* are derived from different reactions and different experimental conditions; a modification of the Hammett equation was defined as the $\rightarrow$ *Yukawa–Tsumo equation*.

☞ [Hammett, 1935; Jaffé, 1953; Yamamoto and Otsu, 1967; Shorter, 1978; Roberts, 1995; Suresh and Gadre, 1998; Popelier, 1999; Drmanić, Jovanović *et al.*, 2000; Lin, Yin *et al.*, 2003; Verma, Kapur *et al.*, 2003; Liu, Fu *et al.*, 2004; Simón-Manso, 2005; Smith and Popelier, 2005]

- **Hammett substituent constants** $\equiv$ *electronic substituent constants*
- **Hamming distance** $\rightarrow$ similarity/diversity (☉ Table S10)
- **Hamming similarity coefficient** $\rightarrow$ similarity/diversity (☉ Table S9)
- **Hancock steric constant** $\equiv$ *corrected Taft steric constant* $\rightarrow$ steric descriptors (☉ Taft steric constant)
- **Hannan–Quinn ϕ -criterion** $\rightarrow$ regression parameters (☉ Table R1)

■ Hansch analysis

Derived from physical organic chemistry and the $\rightarrow$ *Hammett equation*, it can be considered the first approach to modern QSAR studies. Proposed by Hansch and coworkers in the early 1960s [Hansch, Maloney *et al.*, 1962; Hansch, Muir *et al.*, 1963; Hansch and Fujita, 1964; Hansch, Deutsch *et al.*, 1965; Hansch and Anderson, 1967; Hansch, 1969, 1971, 1978], it is the investigation of the quantitative relationships between the biological activity of a series of compounds and their physico-chemical parameters representing hydrophobic, electronic, steric, and other effects using multivariate regression methods [Kubinyi, 1993b].

Hansch analysis assumes that variations in the magnitude of a certain biological activity exhibited by a series of bioactive compounds can be correlated to variations in different physico-chemical factors associated with their structure. Therefore, the basic QSAR equation in the Hansch analysis is defined as

$$\text{biological activity} = f(\Phi_1, \Phi_2, \dots, \Phi_J)$$

where Φ are $\rightarrow$ *physico-chemical properties* of congeneric compounds having a common skeleton but varying substituents. Together with the most significant parameters, factors for hydrogen bonding, van der Waals and charge-transfer forces, etc. can be used, depending on the situation.

The biological activity is usually defined as $\log(1/C)$, where C is the molar concentration of a compound producing a fixed effect.

Hansch analysis tries to correlate biological activity with physico-chemical properties by linear and nonlinear regression analysis, finding property-activity relationship models. A **Craig plot** is a plot of two substituent parameters (e.g., Hansch–Fujita π and Hammett σ values).

The simplest Hansch analysis is based on the **Hansch linear model** [Kubinyi, 1988b], defined as

$$\hat{y}_i = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij}$$

where Φ_{ij} represents the j th physico-chemical property of the i th compound, b_j the regression coefficients, and J the total number of considered properties. The intercept b_0 corresponds to a theoretical biological activity of a compound whose all the property values are zero. This condition is approximately fulfilled for a hydrogen substituted reference compound as several property values are normalized to zero. Depending on the regression coefficient significance, some factors can result not relevant.

For example, a typical Hansch linear equation for monosubstituted derivatives is as the following:

$$\log(1/C) = b_0 + b_1 \cdot \pi + b_2 \cdot \sigma + b_3 \cdot E_s$$

where π indicates $\rightarrow$ *Hansch–Fujita hydrophobic substituent constants*, σ the $\rightarrow$ *electronic substituent constants*, and E_s the $\rightarrow$ *Taft steric constant*.

The j th molecular property of the i th compound Φ_{ij} can be defined as the sum of the values of the substituent constant ϕ of type j over all substituents of that compound:

$$\Phi_{ij} = \sum_{s=1}^S \sum_{k=1}^{N_s} \phi_{ks,j} \cdot I_{i,ks}$$

where $I_{i,ks}$ are $\rightarrow$ *indicator variables* (such as in $\rightarrow$ *Free–Wilson analysis*) indicating the presence, that is, $I_{i,ks} = 1$, and absence, that is, $I_{i,ks} = 0$, of the k th substituent in the s th site for the i th compound; $\phi_{ks,j}$ the j th property of the k th substituent in the s th site; S the number of substitution sites; and N_s the number of group substituents in the s th site. As in each site only one substituent is present for a given compound, S is the total number of contributions to the considered molecular property. Alternatively, the j th molecular property Φ_{ij} of the i th compound is defined as the j th global molecular property such as $\rightarrow$ *log P* or some global $\rightarrow$ *steric descriptors*.

Both the molecular physico-chemical properties Φ obtained by the above relationship and the $\rightarrow$ *substituent constants* ϕ are usually known as **Hansch descriptors** [Hansch, Leo *et al.*, 1995].

Also to take into account some nonlinear contributions of the properties [Hansch and Clayton, 1973; Kubinyi, 1993b], a **Hansch parabolic model** is defined as

$$\hat{y}_i = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij} + \sum_{j=1}^J b''_j \cdot \Phi_{ij}^2$$

Among the quadratic terms, usually the most used is $(\log P)^2$, to mimic the nonlinear behavior of the interchange between a two-phase system (e.g., aqueous/organic system), that is, too low or too high lipophilicity values act as limiting factor. The most common parabolic model is specifically defined as

$$\log(1/C) = b_0 - b_1 \cdot (\log P)^2 + b_2 \cdot \log P + b_3 \cdot \sigma$$

The general **Hansch nonlinear model** is defined as

$$\hat{y}_i = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij} + \sum_{j=1}^J \sum_{j'=j}^J b''_{jj'} \cdot \Phi_{ij} \cdot \Phi_{ij'}$$

also taking into account the combined effects of the properties, even if not usually considered.

Besides the nonlinear models and, specifically, the parabolic model, other models were proposed for nonlinear dependence of the biological response on hydrophobic interactions. Among them, the most important are the **Hansch bilinear models** [Kubinyi, 1977, 1979] such as

$$\log(1/C) = b_0 + b_1 \cdot \log P - b_2 \cdot \log(\beta \cdot P + 1)$$

Special cases of such bilinear models are the **McFarland model** [McFarland, 1970], derived for $b_2 = 2 \times b_1$ and $\beta = 1$ and the **Higuchi–Davis model** [Higuchi and Davis, 1970], for $b_2 = 1$ and

$\beta = V_{\text{lip}}/V_{\text{aq}}$, which is the ratio of the volume of the lipid phase V_{lip} over the volume of the aqueous phase V_{aq} .

The Hansch linear model is related to the $\rightarrow$ *Fujita–Ban model* when, in both models, the hydrogen substituted compound is taken as the reference compound; each Fujita–Ban regression coefficient b_{ks} corresponds to the Hansch equation for a single substituent:

$$b_{ks} \approx \sum_{j=1}^J b_j \cdot \phi_{ks,j}$$

where J is the number of considered properties (e.g., lipophilic, electronic, and steric properties) and $\phi_{ks,j}$ is the j th substituent group property for the k th substituent in the s th site. This relationship means that the group contribution b_{ks} in the Fujita–Ban model of the k th substituent in the s th site is numerically equivalent to the weighted sum of all the physico-chemical properties of that substituent [Kubinyi and Kehrnhahn, 1976]. Substituting the previous relationship in the Fujita–Ban model, it can be observed that the two models are closely related:

$$\begin{aligned} \hat{y}_i &= b_0 + \sum_{s=1}^S \sum_{k=1}^{N_s} b_{ks} \cdot I_{i,ks} = b_0 + \sum_{s=1}^S \sum_{k=1}^{N_s} \sum_{j=1}^J b_j \cdot \phi_{ks,j} \cdot I_{i,ks} \\ &= b_0 + \sum_{j=1}^J b_j \cdot \sum_{s=1}^S \sum_{k=1}^{N_s} \phi_{ks,j} \cdot I_{i,ks} = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij} \end{aligned}$$

In particular, the Fujita–Ban group contributions implicitly contain all the possible physico-chemical contributions of a substituent; as a consequence, the Fujita–Ban models always give an upper limit of correlation, which can be achieved by Hansch linear models.

Hansch–Free–Wilson mixed models were also proposed [Kubinyi, 1976a] by combining the two approaches in a single model. A quadratic term accounting for hydrophobic interactions (usually $\log P$ or π Hansch–Fujita constant) can be added to the Free–Wilson (or Fujita–Ban) model as

$$\hat{y}_i = b_0 + \sum_{s=1}^S \sum_{k=1}^{N_s} b_{ks} \cdot I_{i,ks} + b' \cdot (\log P_i)^2$$

where the first part is the Free–Wilson model, S and N_s , respectively, the number of substitution sites and substituent groups in each s th site, and $I_{i,ks}$ is an indicator variable for the i th compound denoting presence (1) or absence (0) of the k th group in the s th site.

Mixed models can also be obtained by mixing the two approaches, each describing a different group of substituents:

$$\hat{y}_i = b_0 + \sum_{s=1}^S \sum_{k=1}^{N_s} b_{ks} \cdot I_{i,ks} + \sum_{j=1}^J b'_j \cdot \Phi_{ij} + b'' \cdot (\log P_i)^2$$

where the Free–Wilson part accounts for a set of substituents and the Hansch part for another set.

Another mixed model, called here **Site-Property analysis** (*SP analysis*), can be obtained [Todeschini and Consonni, 2000]; it represents information regarding the presence of each substituent group in each site by the corresponding physico-chemical properties, that is, the information of the indicator variables $I_{i,ks}$ of the Fujita–Ban analysis is preserved in each site but

is represented by the set of selected properties:

$$\hat{y}_i = b_0 + \sum_{s=1}^S \sum_{j=1}^J b_{sj} \cdot \phi_{is,j}$$

where S is the number of substitution sites, J the number of properties, and $\phi_{is,j}$ the j th substituent group property in the s th site for the i th compound, that is,

$$\phi_{is,j} = \sum_{k=1}^{N_s} \phi_{ks,j} \cdot I_{i,ks}$$

where

$$\sum_{k=1}^{N_s} I_{i,ks} = 1$$

Therefore, *SP analysis* can be performed only if all substituent group constants are available for all the substituents in the data set. The total number of variables is $S \times J$. This approach allows complete $\rightarrow$ *reversible decoding*, that is, the possibility to interpret by the model *how* and *where* the response is increased/decreased.

By assuming that a response would depend on both the holistic properties of molecules and the local specific group contributions a mixed **Global-Site-Property analysis** (*GSP analysis*) can be achieved by a generalized model such as

$$\hat{y}_i = b_0 + \sum_{l=1}^{p'} b'_l \cdot \Phi_{il} + \sum_{s=1}^S \sum_{j=1}^J b_{sj} \cdot \phi_{is,j}$$

where Φ are generic global properties, that is, global descriptors obtained by any method, p' the number of selected descriptors, and $\phi_{is,j}$ the j th substituent group property in the s th site for the i th compound. The total number of variables is $S \times J + p'$.

In Hansch analysis and related approaches, the statistical problems due to the relatively high number of variables with respect to the number of compounds have to be faced using $\rightarrow$ *variable selection* procedures.

Although the predictive power of a model is considered to be a criterion for the relevance of QSAR models, the main purpose of Hansch analysis and related approaches such as Free–Wilson analysis concerns not prediction, but a better understanding of the chemical problem.

 Additional references are collected in the thematic bibliography (see Introduction).

- **Hansch bilinear models** $\rightarrow$ Hansch analysis
- **Hansch descriptors** $\rightarrow$ Hansch analysis
- **Hansch–Free–Wilson mixed models** $\rightarrow$ Hansch analysis
- **Hansch–Fujita hydrophobic substituent constants** $\rightarrow$ lipophilicity descriptors
- **Hansch linear model** $\rightarrow$ Hansch analysis
- **Hansch nonlinear model** $\rightarrow$ Hansch analysis
- **Hansch parabolic model** $\rightarrow$ Hansch analysis
- **Harary–Balaban index** $\rightarrow$ distance matrix

- Harary Cluj detour indices → Harary indices
- Harary Cluj distance indices → Harary indices
- Harary connectivity index → distance matrix
- Harary detour/distance indices → Harary indices
- Harary detour indices → Harary indices
- Harary index → distance matrix
- Harary index → Harary indices

■ Harary indices

The **Harary index** H [Plavšić, Nikolić *et al.*, 1993b], also called **RDSUM index** [Ivanciuc, Balaban *et al.*, 1993b], is a molecular topological index derived from the → reciprocal distance matrix $\mathbf{D}^{-1}$ by the → Wiener operator W_i :

$$H \equiv \text{RDSUM} \equiv W_i(\mathbf{D}^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij}^{-1} = \frac{1}{2} \sum_{i=1}^A \text{RDS}_i \quad j \neq i$$

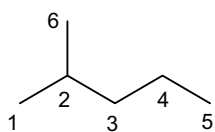
where RDS_i is the → reciprocal distance sum of the i th vertex.

The Harary index increases with both molecular size and → molecular branching; it is therefore a measure of molecular compactness like the → Wiener index. However, the Harary index seems to be a more discriminating index than the Wiener index. A variant H' of the Harary index, called **Harary number**, was derived from the → reciprocal square distance matrix $\mathbf{D}^{-2}$ [Mihalić and Trinajstić, 1992; Plavšić, Nikolić *et al.*, 1993b], still from the Wiener operator W_i :

$$H' \equiv W_i(\mathbf{D}^{-2}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij}^{-2} \quad j \neq i$$

Example H1

Reciprocal distance matrix $\mathbf{D}^{-1}$ and Harary index H for 2-methylpentane.



$\mathbf{D}^{-1} =$

Atom	1	2	3	4	5	6	RDS_i
1	0	1	0.50	0.33	0.25	0.50	2.58
2	1	0	1	0.50	0.33	1	3.83
3	0.50	1	0	1	0.50	0.50	3.50
4	0.33	0.50	1	0	1	0.33	3.16
5	0.25	0.33	0.50	1	0	0.25	2.33
6	0.50	1	0.50	0.33	0.25	0	2.58

$$H = \frac{1}{2} \cdot \sum_{i=1}^6 \sum_{j=1}^6 [\mathbf{D}^{-1}]_{ij} = \frac{18}{2} = 9$$

By generalization, Harary indices and **hyper-Harary indices** (or **hyper-Harary numbers**) are all → molecular descriptors derived from the application of the Wiener operator to reciprocal → graph-theoretical matrices; Harary indices are obtained from edge-type matrices, whose nonvanishing off-diagonal elements are only those corresponding to pairs of adjacent vertices, while the hyper-Harary indices are calculated from path-type matrices, whose nonvanishing

elements correspond to all pairs of vertices [Diudea, 1997c]. Other topological indices based on a modified reciprocal distance matrix are the $\rightarrow$ *constant interval reciprocal indices*.

The most important Harary indices are listed below.

- **Harary Wiener indices**

The Harary index and hyper-Harary index, defined only for acyclic graphs, are obtained from $\rightarrow$ *reciprocal Wiener matrix* [Diudea, 1997c; Diudea and Gutman, 1998]. The two following indices are derived, respectively, from the reciprocal edge-Wiener matrix $\mathbf{W}_e^{-1}$ and reciprocal path-Wiener matrix $\mathbf{W}_p^{-1}$:

$$H_{W_e} \equiv Wi(\mathbf{W}_e^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{W}_e^{-1}]_{ij}$$

$$H_{W_p} \equiv Wi(\mathbf{W}_p^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{W}_p^{-1}]_{ij}$$

It must be noted that while the indices obtained by applying the Wiener operator to the distance matrix $\mathbf{D}$ and to the $\rightarrow$ *edge-Wiener matrix* $\mathbf{W}_e$ are equal (i.e., the $\rightarrow$ *Wiener index*), the corresponding Harary indices are not, that is, $H \neq H_{W_e}$.

- **hyper-Harary distance index**

This is obtained from the $\rightarrow$ *reciprocal distance-path matrix* $\mathbf{D}_p^{-1}$ as

$$H_{D_p} \equiv Wi(\mathbf{D}_p^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}_p^{-1}]_{ij}$$

For acyclic graphs, the equality between the $\rightarrow$ *hyper-distance-path index* D_p and the $\rightarrow$ *hyper-Wiener index* WW is not true for the corresponding hyper-Harary indices, that is, $H_{D_p} \neq H_{W_p}$.

- **Harary detour indices**

These are obtained from the $\rightarrow$ *reciprocal detour matrix* Δ^{-1} as [Diudea, Katona *et al.*, 1998]

$${}^1H_{\Delta} \equiv Wi(\Delta_e^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot [\Delta^{-1}]_{ij} \quad \text{and} \quad H_{\Delta} \equiv Wi(\Delta^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta^{-1}]_{ij}$$

where Δ_e^{-1} is the reciprocal edge-detour matrix that accounts only for pairs of adjacent vertices and a_{ij} are the elements of the adjacency matrix equal to one for pairs of adjacent vertices and zero otherwise.

- **Harary detour-distance indices**

These are obtained from the $\rightarrow$ *reciprocal detour-distance combined matrix* as

$${}^1H_{\Delta D} = \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot [\Delta \wedge \mathbf{D}^{-1}]_{ij} \quad H_{\Delta D} = \sum_{i=1}^A \sum_{j=1}^A [\Delta \wedge \mathbf{D}^{-1}]_{ij}$$

where a_{ij} are the elements of the adjacency matrix equal to one for pairs of adjacent vertices, and zero otherwise. The same index values are, obviously, obtained from the corresponding transpose matrix, that is, the $\rightarrow$ *reciprocal distance-detour combined matrix*.

• Harary Cluj-distance indices

These are molecular indices derived from $\rightarrow$ *reciprocal Cluj matrices*. The Harary-type index is calculated from the reciprocal edge-Cluj-distance matrix $\mathbf{CJD}_e^{-1}$ as

$$H_{CJD_e} \equiv Wi(\mathbf{CJD}_e^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{CJD}_e^{-1}]_{ij}$$

while the hyper-Harary-type index from the reciprocal path-Cluj-distance matrix $\mathbf{CJD}_p^{-1}$ as

$$H_{CJD_p} \equiv Wi(\mathbf{CJD}_p^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{CJD}_p^{-1}]_{ij}$$

In acyclic graphs, the Harary edge-Cluj-distance index H_{CJD_e} coincides with the Harary Wiener index H_{W_e} and the Harary Szeged index H_{SZ_e} ($H_{CJD_e} = H_{W_e} = H_{SZ_e}$), for the corresponding hyper-Harary indices the following relationships hold: $H_{CJD_p} = H_{W_p} \neq H_{SZ_p}$; in cyclic graphs, only $H_{CJD_e} = H_{SZ_e}$, while the other indices give distinct values [Diudea, Pârv *et al.*, 1997b].

• Harary Cluj-detour indices

These are other molecular indices derived from $\rightarrow$ *reciprocal Cluj matrices*. The Harary-type index is calculated from the reciprocal edge-Cluj-detour matrix $\mathbf{CJA}_e^{-1}$ as [Diudea, Katona *et al.*, 1998]

$$H_{CJA_e} \equiv Wi(\mathbf{CJA}_e^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{CJA}_e^{-1}]_{ij}$$

and the hyper-Harary-type index from the reciprocal path-Cluj-detour matrix $\mathbf{CJA}_p^{-1}$:

$$H_{CJA_p} \equiv Wi(\mathbf{CJA}_p^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{CJA}_p^{-1}]_{ij}$$

• Harary Szeged indices

These are molecular indices derived from $\rightarrow$ *reciprocal Szeged matrices*. The Harary-type index is obtained from the reciprocal edge-Szeged matrix $\mathbf{SZ}_e^{-1}$ as [Diudea, 1997c]

$$H_{SZ_e} \equiv Wi(\mathbf{SZ}_e^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{SZ}_e^{-1}]_{ij}$$

and the hyper-Harary-type index from the reciprocal path-Szeged matrix $\mathbf{SZ}_p^{-1}$:

$$H_{SZ_p} \equiv Wi(\mathbf{SZ}_p^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{SZ}_p^{-1}]_{ij}$$

In acyclic graphs, the Harary Szeged index H_{SZ_e} coincides with the Harary Wiener index H_{W_e} ($H_{SZ_e} = H_{W_e}$) while the corresponding hyper-Harary indices are different ($H_{SZ_p} \neq H_{W_p}$); in cyclic graphs all these indices differ [Diudea, Pârv *et al.*, 1997b].

• Harary walk indices

The Harary walk indices are obtained from the $\rightarrow$ reciprocal walk matrix $\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}^{-1}$:

$$H_{W_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}} \equiv W_i(\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}^{-1}]_{ij}$$

where $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ are square $A \times A$ matrices [Diudea, 1997c].

📖 [Estrada and Rodriguez, 1997]

- **Harary matrix** $\equiv$ reciprocal distance matrix $\rightarrow$ distance matrix
- **Harary number** $\rightarrow$ distance matrix
- **Harary number** $\rightarrow$ Harary indices
- **Harary Szeged indices** $\rightarrow$ Harary indices
- **Harary walk indices** $\rightarrow$ Harary indices
- **Harary Wiener indices** $\rightarrow$ Harary indices
- **hardness density** $\equiv$ local hardness $\rightarrow$ quantum-chemical descriptors ($\odot$ hardness indices)
- **hardness indices** $\rightarrow$ quantum-chemical descriptors
- **harmonic mean** $\rightarrow$ statistical indices ($\odot$ indices of central tendency)
- **harmonic oscillator model of aromaticity index** $\equiv$ HOMA index $\rightarrow$ delocalization degree indices
- **harmonic oscillator stabilization energy** $\equiv$ HOSE index $\rightarrow$ delocalization degree indices
- **harmonic topological index** $\rightarrow$ vertex degree
- **Hartley information** $\rightarrow$ information content
- **HASA index** $\equiv$ SSAA index $\rightarrow$ charged partial surface area descriptors
- **HASA₂ index** $\rightarrow$ charged partial surface area descriptors ($\odot$ SSAA index)
- **hash structural codes** $\rightarrow$ substructure descriptors

■ Hasse diagram

Among the $\rightarrow$ ranking methods, the Hasse diagram is a graphical means of illustrating partial order ranking proposed by Hasse [Hasse, 1952]. It was introduced in environmental sciences and QSAR/QSPR studies by Halfon [Halfon and Reggiani, 1986; Halfon, 1989] and refined by Brüggemann [Brüggemann and Bartel, 1999; Brüggemann, Bücherl *et al.*, 1999; Brüggemann, Pudenz *et al.*, 2001] and Carlsen [Carlsen, Sørensen *et al.*, 2002]. The first applications on chemical structure descriptors were proposed by Klein [Klein and Babic, 1997] and Ivanciuc [Ivanciuc, Ivanciuc *et al.*, 2000e]. Hasse diagrams were also used to represent $\rightarrow$ DNA sequences.

Given a set Q of n elements, each described by a vector $\mathbf{x}$ of p variables (attributes), the two elements s and t belonging to Q are comparable if for all the variables x_j either $x_j(t) \geq x_j(s)$ or $x_j(s) \geq x_j(t)$. If $x_j(t) \geq x_j(s)$ for all x_j ($j = 1, \dots, p$) then $t \triangleright s$, that is, t covers s (or s is covered by t). The request “for all” is very important and is called the *generality principle*:

$$t \triangleright s \Leftrightarrow x_j(t) \geq x_j(s) \quad \forall j \in \{1, p\}$$

The ordering relationships between all the pairs of elements are collected into the **Hasse matrix**; for each pair of elements s and t the entry of this matrix is

$$[\mathbf{H}]_{st} = \begin{cases} +1 & \text{if } x_j(s) \geq x_j(t) \quad \forall j \in \{1, p\} \\ -1 & \text{if } x_j(t) \geq x_j(s) \quad \forall j \in \{1, p\} \\ 0 & \text{otherwise} \end{cases}$$

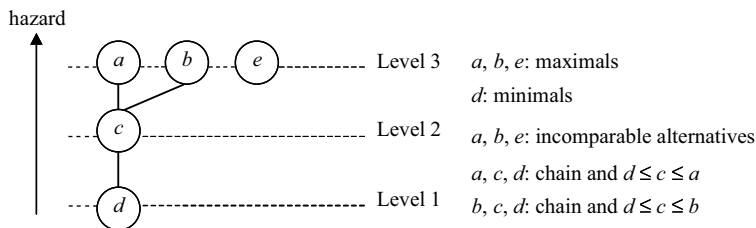
In practice, if the entry $s-t$ contains $+1$, the entry $t-s$ contains -1 ; if the entry $s-t$ contains 0 , also the entry $t-s$ contains 0 . Then, the Hasse matrix is a square $n \times n$ antisymmetric matrix, whose elements take only the values 0 and ± 1 . Moreover, in presence of elements having the same variable values (for all the variables), in both the corresponding entries of the Hasse matrix ($s-t$ and $t-s$), a value equal to 1 is stored. In other words, the object t “dominates” the object s if no contradictions are present in all the variables describing the data; otherwise, if for some variables t “dominates” s and for some others s “dominates” t , the two objects are *not comparable*.

Moreover, if only one criterion is used or all the criteria have a $\rightarrow$ *Spearman rank correlation* equal to one, that is, all the variables provide the same ordering, then a complete or total order is obtained, and all the alternatives are comparable.

Example H2

Hasse matrix $\mathbf{H}$ and Hasse diagram calculated from five objects described by two variables. Dominance is defined by the maximum values for both variables.

Object	Variable x_1	Variable x_2
a	5	7
b	6	5
c	4	4
d	2	1
e	1	8

$$\mathbf{H} = \begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 0 & +1 & +1 & 0 \\ b & 0 & 0 & +1 & +1 & 0 \\ c & -1 & -1 & 0 & +1 & 0 \\ d & -1 & -1 & -1 & 0 & 0 \\ e & 0 & 0 & 0 & 0 & 0 \end{array}$$


- **Hasse matrix** → Hasse diagram
- **HATS indices** → GETAWAY descriptors
- **HATS total index** → GETAWAY descriptors
- **Hausdorff chirality measure** → chirality descriptors
- **HB₁ and HB₂ parameters** → hydrogen-bonding descriptors
- **HBCA index** → charged partial surface area descriptors
- **HB-CPSA descriptors** ≡ *hydrogen-bond charged partial surface area descriptors* → charged partial surface area descriptors
- **H-bonding descriptors** ≡ *hydrogen-bonding descriptors*
- **HB parameter** → hydrogen-bonding descriptors
- **HBSA index** → charged partial surface area descriptors
- **HCD descriptors** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **HDCA index** → charged partial surface area descriptors
- **HDCA₂ index** → charged partial surface area descriptors
- **H-depleted molecular graph** → molecular graph
- **HDSA index** ≡ *SSAH index* → charged partial surface area descriptors
- **HDSA₂ index** → charged partial surface area descriptors (⊙ HDCA₂ index)
- **Henry's law constant** → physico-chemical properties
- **Hermite-like wave functions** → characteristic polynomial-based descriptors
- **Herndon resonance energy** → delocalization degree indices
- **HE-state fields** → electrotopological state indices
- **HE-state index** ≡ *hydrogen electrotopological state index* → electrotopological state indices
- **heteroatom-corrected extended adjacency matrix** → extended adjacency matrices
- **heteroatom/multiplicity-corrected extended adjacency matrix** → extended adjacency matrices
- **H-filled molecular graph** → molecular graph
- **hierarchical fragment description** → substructure descriptors
- **hierarchically ordered extended connectivities algorithms** → canonical numbering
- **hierarchical QSAR approach** → Structure/Response Correlations
- **higher order map matrices** → biodescriptors (⊙ proteomics maps)
- **higher order Wiener numbers** → Wiener matrix
- **higher order χ matrices** → weighted matrices (⊙ weighted adjacency matrices)
- **highest occupied molecular orbital** → quantum-chemical descriptors
- **highest occupied molecular orbital energy** → quantum-chemical descriptors
- **highest scoring common substructure** → maximum common substructure
- **Higuchi–Davis model** → Hansch analysis

■ Hildebrand solubility parameter (δ_H)

A measure of the intermolecular interactions between solute molecules and their environment, defined as

$$\delta_H = \sqrt{\frac{-E_c}{\bar{V}}} = \sqrt{\frac{\Delta H_v - RT}{\bar{V}}}$$

where E_c is the cohesion energy between liquid molecules defined as a function of → *polarizability*, → *ionization potential*, and → *dipole moment*; $\bar{V}$ is the → *molar volume* of the compound;

ΔH_v the vaporisation enthalpy of the liquid at 298°K, T the absolute temperature; and R the gas universal constant [Hildebrand and Scott, 1950]. For apolar or moderately polar compounds, the vaporization enthalpy can be estimated by their boiling points (bp, °K) as

$$\Delta H_v(\text{cal/mole}) = -2950 + 23.7 \cdot \text{bp} + 0.02 \cdot \text{bp}^2$$

Often referred to as the **solvent cohesive energy density**, the Hildebrand solubility parameter is considered a measure of the solvent contribution to the $\rightarrow$ *cavity term*, and is used as a correction factor in the $\rightarrow$ *solvatochromic equation*. It is related to the general definition of **London cohesive energy** between two interacting species:

$$\epsilon_L = \frac{3 \cdot \alpha_i \cdot \alpha_j}{2 \cdot r_{ij}^6} \cdot \frac{IP_i \cdot IP_j}{IP_i + IP_j}$$

where α and IP are the $\rightarrow$ *polarizability* and the $\rightarrow$ *ionization potential* of the two species, respectively, and r is their $\rightarrow$ *geometric distance*.

Moreover, for large molecules or polymeric systems, a solubility parameter can be calculated from group contributions as [Small, 1953]

$$\delta_s = \frac{\sum_k F_k}{\bar{V}}$$

where F_k is the molar attraction constant of the k th substituent group of the compound and the sum runs over all groups; $\bar{V}$ is the molar volume.

📖 [Kamlet, Carr *et al.*, 1981; Pussemier, De Borger *et al.*, 1989; Mutelet, Ekulu *et al.*, 2002; Stefanis, Constantinou *et al.*, 2004]

- **Hill potential function** $\rightarrow$ molecular interaction fields (⊙ steric interaction fields)
- **H indices** $\rightarrow$ GETAWAY descriptors
- **HINT** $\equiv$ *Hydropatic INTeractions* $\rightarrow$ molecular interaction fields (⊙ hydrophobic fields)
- **HLOGP** $\rightarrow$ lipophilicity descriptors
- **HNSO_T index** $\rightarrow$ hydrogen-bonding descriptors
- **H total index** $\rightarrow$ GETAWAY descriptors
- **HOC algorithms** $\equiv$ *hierarchically ordered extended connectivities algorithms* $\rightarrow$ canonical numbering
- **HOC rank descriptors** $\rightarrow$ canonical numbering (⊙ hierarchically ordered extended connectivities algorithms)
- **Hodes statistical-heuristic method** $\rightarrow$ scoring functions
- **Hodgkin similarity index** $\rightarrow$ quantum similarity
- **holograms** $\equiv$ *molecular holograms* $\rightarrow$ substructure descriptors (⊙ fingerprints)
- **Hologram QSAR** $\rightarrow$ substructure descriptors (⊙ fingerprints)
- **holographic vectors** $\rightarrow$ vectorial descriptors
- **Holtz–Stock inductive constant** $\rightarrow$ electronic substituent constants (⊙ inductive electronic constants)
- **HOMA index** $\rightarrow$ delocalization degree indices
- **homeomorphic graphs** $\rightarrow$ graph
- **HOMO–LUMO energy gap** $\rightarrow$ quantum-chemical descriptors

- **HOMO–LUMO energy fraction** → quantum-chemical descriptors
- **HOSE index** → delocalization degree indices
- **Hosoya graph decomposition** → Hosoya Z index
- **Hosoya ID number** → Hosoya Z matrix
- **Hosoya matrix** $\equiv$ *Hosoya Z matrix*
- **Hosoya nonadjacent number** $\equiv$ *nonadjacent number* → Hosoya Z index
- **Hosoya mean information index** → Hosoya Z index
- **Hosoya operator** → characteristic polynomial-based descriptors
- **Hosoya resonance energy** → delocalization degree indices
- **Hosoya total information index** → Hosoya Z index
- **Hosoya-type indices** → characteristic polynomial-based descriptors
- **Hosoya–Wiener index** → double invariants
- **Hosoya–Wiener polynomial** $\equiv$ *Wiener polynomial* → Wiener index
- **Hosoya Z' index** → characteristic polynomial-based descriptors

■ **Hosoya Z index** ($\equiv$ *Z index*)

The Hosoya Z index of a graph G is derived by a combinatorial algorithm and is defined as [Hosoya, 1971]

$$Z = \sum_{k=0}^{\lfloor A/2 \rfloor} a(G, k)$$

where $a(G, k)$, called **nonadjacent number** (or **Hosoya nonadjacent number**), is the number of ways through which k edges may be selected from all of the B edges of graph G , so that no two of them are adjacent, that is, the number of $\rightarrow k$ -*matchings*. A is the number of graph vertices and the Gaussian brackets $\lfloor \]$ represent the greatest integer not exceeding $A/2$. The Z index is calculated by summing the $a(G, k)$ coefficients over all different k values. For any graph, $a(G, 0) = 1$ and $a(G, 1) = B$.

The Hosoya Z index depends on the molecular size as well as on branching and ring closure. It was also found to correlate well with the boiling point.

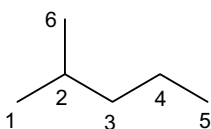
Hosoya graph decomposition is a graph edge decomposition defined as

$$\{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_j, \dots, \mathcal{E}_{N_k}\}_k$$

where N_k is the nonadjacent number $a(G, k)$ and $\mathcal{E}_j$ is the j th $\rightarrow k$ -*matching* of the graph, that is, a subset of k nonadjacent edges.

Example H3

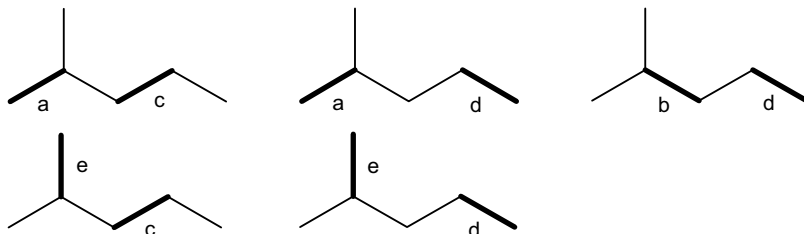
For 2-methylpentane, the H-depleted molecular graph and the nonadjacent numbers are:



$$\begin{aligned} a(G; 0) &= 1 \\ a(G; 1) &= 5 \\ a(G; 2) &= 5 \\ a(G; 3) &= 0 \end{aligned}$$

The number of graph vertices is six, then $[6/2] = 3$, which is the greatest integer not exceeding $6/2 = 3$, and thus $k = 0, 1, 2, 3$.

The computation of $a(G, 2)$ is presented. There are five 2-matchings as shown in the graphs below where the two nonadjacent edges are indicated by bold lines.



The Hosoya graph decomposition for $k = 2$ is

$$E_1 = \{a, c\}, \quad E_2 = \{a, d\}, \quad E_3 = \{b, d\}, \quad E_4 = \{e, c\}, \quad E_5 = \{e, d\}$$

The Hosoya Z index of 2-methylpentane is $Z = 1 + 5 + 5 + 0 = 11$

Hosoya found that the values of the Z index for linear graphs coincide with the $\rightarrow$ *Fibonacci numbers*, that is, $Z = F_A$, where A is the number of vertices in the molecular graph [Hosoya, 1973]; therefore, for a linear graph, the Z index is closely related to the $\rightarrow$ *Merrifield–Simmons index* [Gutman, Hosoya *et al.*, 1992; Randić, Morales *et al.*, 1996].

Using nonadjacent numbers $a(G, k)$ as coefficients, the **Z-counting polynomial** Q of G is a $\rightarrow$ *counting polynomial* defined as

$$Q(G; x) = \sum_{k=0}^{[A/2]} a(G, k) \cdot x^k$$

where the square brackets refer to the largest integer of $A/2$.

The Hosoya Z index can also be defined as the value of the Q polynomial for $x = 1$.

A **modified Hosoya index** (or **Z^* index**) was defined by a generalization of the Z-counting polynomial by treating the powers of x as independent variables [Randić and Zupan, 2001]:

$$Z^*(x_1, x_2, \dots) = \sum_k a(G, k) \cdot x_k$$

where x_k are integer weights representing the number of times each edge has appeared in all disjoint edge patterns.

Closely related to the Z-counting polynomial, the **matching polynomial** (or **acyclic polynomial** or **reference polynomial**) was defined in terms of nonadjacent numbers $a(G, k)$ as [Gutman, Milun *et al.*, 1977; Gutman, 1979; Gutman, Graovac *et al.*, 1982; Hosoya, 1988; Ivanciuc, 1998d; Graovac, Vukicević *et al.*, 2005]

$$M(G; x) = \sum_{k=0}^{[A/2]} (-1)^k a(G, k) \cdot x^{A-2k}$$

For acyclic graphs, the matching polynomial coincides with the $\rightarrow$ graph characteristic polynomial. Moreover, it was demonstrated the following relationship between the Z-counting and matching polynomials [Hosoya, 2003]:

$$M(G; x) = x^A \cdot Q(G; x) \cdot \left(-\frac{1}{x^2}\right)$$

For acyclic graphs the Z-counting polynomial coefficients $a(G, k)$ coincide with the absolute values of the coefficients of the characteristic polynomial of the adjacency matrix (i.e., $\rightarrow$ graph characteristic polynomial) [Nikolić, Plavšić *et al.*, 1992]. Therefore, for any graph, the Hosoya Z index can also be calculated from the matching polynomial coefficients m_{2k} as

$$Z = \sum_{k=0}^{[A/2]} a(G, k) = \sum_{k=0}^{[A/2]} |m_{2k}|$$

and, only for acyclic graphs, from the coefficients c_i of the characteristic polynomial of the adjacency matrix:

$$Z = \sum_{k=0}^{[A/2]} a(G, k) = \sum_{i=0}^A |c_i|$$

By generalization of this last definition of the Hosoya Z index to any graph and any $\rightarrow$ graph-theoretical matrix $\mathbf{M}$, the $\rightarrow$ Hosoya-type indices were proposed as the sum of the absolute values of the coefficients of the characteristic polynomial of the matrix $\mathbf{M}$.

Example H4

The Z-counting polynomial of 2-methylpentane is

$$Q(G; x) = 1 + 5x + 5x^2$$

The matching polynomial is

$$M(G; x) = x^6 - 5x^4 + 5x^2$$

The graph characteristic polynomial is

$$Ch(G; x) = x^6 - 5x^4 + 5x^2$$

From the matching polynomial or the characteristic polynomial, the Hosoya Z index is

$$Z = |1| + |5| + |5| = 11$$

To calculate the Z index for large graphs a composition principle for Z was developed [Hosoya, 1971]: the Z index value of the graph G is obtained as the product of the Z values of graphs G' and G'' , derived from G by cutting an edge, plus the product of the Z values of all graphs derived from G' and G'' by cutting all edges incident to the edge b in the original graph G . The Z value for an empty graph is set at one. It was demonstrated that the Z value for the graph G is uniquely obtained independently of the choice of the edge b to cut in the first step.

The Hosoya Z index for an edge-weighted graph takes into account the edge weights w_b , defining the nonadjacent numbers $a(G, w, k)$ as

$$a(G, w, k) = \sum_{\mathcal{E}_j} \left(\prod_{b=1}^k w_b \right)_{\mathcal{E}_j}$$

where the product is over all edges b of a set $\mathcal{E}$ comprised of k nonadjacent edges and the summation runs over all the subsets of the Hosoya graph decomposition. Obviously, it follows that $a(G, w, 1) = \sum_b w_b$ and, by definition, $a(G, w, 0) = 1$.

The **Hosoya total information index**, denoted as I_Z , is calculated on the Hosoya graph decomposition and is based on the distribution of $a(G, k)$ coefficients. It is an index of $\rightarrow$ *total information content* defined as the following:

$$I_Z = Z \cdot \log_2 Z - \sum_{k=0}^{[A/2]} a(G, k) \cdot \log_2 a(G, k)$$

where Z is the Hosoya index. Analogously, the **Hosoya mean information index** is defined as $\rightarrow$ *mean information content*:

$$\bar{I}_Z = - \sum_{k=0}^{[A/2]} \frac{a(G, k)}{Z} \cdot \log_2 \frac{a(G, k)}{Z}$$

It is noteworthy that I_Z and $\bar{I}_Z$ indices coincide with $\rightarrow$ *information indices on polynomial coefficients* for acyclic graphs.

Example H5

Calculation of Hosoya total and mean information indices is shown for 2-methylpentane. Data from Example H3.

$$I_Z = 11 \times \log_2 11 - 1 \times \log_2 1 - 2 \times (5 \times \log_2 5) = 14.834$$

$$\bar{I}_Z = - \frac{1}{11} \times \log_2 \frac{1}{11} - 2 \times \left(\frac{5}{11} \times \log_2 \frac{5}{11} \right) = 1.349$$

Generalized Hosoya indices Z_m were proposed [Hermann and Zinn, 1995] as counts of nonadjacent molecular paths in the graph G :

$$Z_m = \sum_{k=0}^G a(G, k)_m$$

where the subscript m refers to the order of the index and indicates the path length, $a(G, k)_m$ is the number of all possible combinations of k nonadjacent paths of length m , $a(G, 0) = 1$ by definition, and G is the maximum possible number of k , dependent on the selected path length and the molecule size. The Z_1 index coincides with the original Hosoya Z index and the Z_0 index counts all possible combinations of nonadjacent vertices. These indices must not be confused with the $\rightarrow$ *m Z numbers* based on the sequential erasure of each path from the original graph.

Hosoya Z index was also used to define $\rightarrow$ resonance indices and $\rightarrow$ graphical bond order.

Additional references are collected in the thematic bibliography (see Introduction).

■ Hosoya Z matrix

A square symmetric matrix of dimension $A \times A$, A being the number of vertices in the $\rightarrow$ *H-depleted molecular graph* G . The original Hosoya Z matrix is defined only for acyclic graphs; each off-diagonal element is equal to the $\rightarrow$ *Hosoya Z index* of the subgraph G' obtained from the graph G by erasing all edges along the path connecting two vertices v_i and v_j as

$$[Z]_{ij} = \begin{cases} Z(G') & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The diagonal entries are zero by definition [Randić, 1994b]. If more than one subgraph is obtained by the erasure procedure, the matrix element is calculated by summing up all of the Hosoya Z indices of the subgraphs.

A general definition of the Hosoya Z matrix (**generalized Hosoya Z matrix**) able to represent both acyclic and cyclic graphs is the following [Plavšić, Šoškić *et al.*, 1997]:

$$[Z]_{ij} = \begin{cases} \frac{\sum_{\min p_{ij}} Z(G')}{\min p_{ij}} & \text{if } i \neq j \\ Z(G) & \text{if } i = j \end{cases}$$

where $Z(G')$ is the Z index of the graph G' obtained from the graph G by erasing all edges along the shortest path connecting the vertices v_i and v_j , that is, the $\rightarrow$ *geodesic* $\min p_{ij}$, and the summation goes over all $\min p_{ij}$ geodesics between the considered vertices. The diagonal entries are simply equal to $Z(G)$, that is, the Hosoya Z index of the original graph.

It is interesting to observe that the magnitude of the entries in the matrix Z decreases as the separation between the vertices increases, it can therefore be expected to simulate the interactions between the pairs of vertices well.

The Z'/Z index is among the $\rightarrow$ *graphical bond order descriptors* and can be obtained from the Hosoya Z matrix only by considering the entries relative to pairs of adjacent vertices (i.e., bonds):

$$\frac{Z'}{Z} = \frac{1}{Z} \cdot \text{Wi}(Z_e) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \left(\frac{[Z_e]_{ij}}{Z} \right) = \sum_{b=1}^B \left(\frac{Z(G')}{Z} \right)_b$$

where Z is the $\rightarrow$ *Hosoya Z index* of the whole graph and Wi is the $\rightarrow$ *Wiener operator* applied to the **edge-Hosoya matrix**, denoted as Z_e , where the only the nonvanishing elements correspond to pairs of adjacent vertices:

$$Z_e = Z \otimes A$$

A is the $\rightarrow$ *adjacency matrix*, $\otimes$ the $\rightarrow$ *Hadamard matrix product*, and B is the total number of graph edges. Each term in the sum is a $\rightarrow$ *graphical bond order*.

Other graph invariants derived from the Hosoya Z matrix are the $\rightarrow$ *eigenvalues* and the coefficients of the $\rightarrow$ *characteristic polynomial*. Moreover, sequences of weighted paths and the $\rightarrow$ *weighted path counts* were defined using as the path weights the magnitude of the Z

matrix elements. ${}^mZ$ numbers are calculated as the sum of the magnitude of the entries corresponding to pairs of vertices separated by the shortest path of length m :

$${}^mZ = \sum_{m_{p_{ij}}} [Z]_{ij} = \sum_{m_{p_{ij}}} Z(G - {}^m p_{ij})$$

where the term in the second summation is the Hosoya Z number of the graph G from which the path ${}^m p_{ij}$ is erased. ${}^mZ$ numbers must not be confused with $\rightarrow$ *generalized Hosoya indices* Z_m .

Summing up all ${}^mZ$ numbers, a global molecular descriptor called **Hosoya ID number** ZID is obtained:

$$ZID = A + \sum_m {}^mZ$$

where A is the number of vertices corresponding to ${}^0Z$ and the summation goes over all path lengths.

From row sums of the Z matrix using the $\rightarrow$ *Ivanciuc–Balaban operator* IB , a Balaban-like index called **L_Z index** is calculated as

$$L_Z \equiv IB(Z) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (VS_i(Z) \cdot VS_j(Z))^{-1/2}$$

where B is the number of graph edges, C the $\rightarrow$ *cyclomatic number*, and VS_i and VS_j the row sums of the Hosoya matrix Z corresponding to the vertices v_i and v_j ; a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* which are equal to one for adjacent vertices, and zero otherwise.

Analogously, applying the $\rightarrow$ *Wiener operator* Wi to the matrix Z , a Wiener-type index called **K_Z index** is obtained as

$$K_Z = Wi(Z) \equiv \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [Z]_{ij}$$

📖 [Plavšić, Šoškić *et al.*, 1996a; Janežič, Lučić *et al.*, 2007]

- **Hou fitness function** $\rightarrow$ regression parameters
- **HRNCG index** $\rightarrow$ charged partial surface area descriptors
- **HRNCS index** $\rightarrow$ charged partial surface area descriptors
- **HRPCG index** $\rightarrow$ charged partial surface area descriptors
- **HRPCS index** $\rightarrow$ charged partial surface area descriptors
- **H_1 topological index** $\rightarrow$ connectivity indices
- **Hückel resonance energy** $\rightarrow$ delocalization degree indices
- **Hückel's rule** $\rightarrow$ delocalization degree indices
- **Hurvich–Tsai criterion** $\rightarrow$ regression parameters (☉ Table R1)
- **Hutter likeliness score** $\rightarrow$ scoring functions
- **Hu–Xu ID number** $\rightarrow$ ID numbers
- **Hu–Xu vertex degree** $\rightarrow$ vertex degree
- **hydrated surface area** $\rightarrow$ molecular surface (☉ solvent-accessible molecular surface)

■ Hydration Free Energy Density (HFED)

The empirical hydration free energy density is expressed by a linear combination of some physical properties calculated around the molecule with net atomic charges, polarizabilities, dispersion coefficients of the atoms in the molecule, and solvent accessible surface [Son, Han *et al.*, 1999]. These physical properties are the result of the interaction of the molecule with its environment. To calculate the HFED of a molecule a grid model was proposed; a shell of critical thickness r_C was defined around the solvent-accessible surface with a number of grid points inside (e.g., 8 points/Å³).

The hydration free energy density, denoted by g_k , is calculated at each k th grid point of the shell as

$$g_k = \frac{b_0}{N_G} + \frac{b_1}{N_G} \cdot \sum_{k=1}^{N_G} R_k + b_2 \cdot \left| \sum_{i=1}^A \frac{q_i}{r_{ik}} \right| + b_3 \cdot \sum_{i=1}^A \frac{q_i^2}{r_{ik}^2} + b_4 \cdot \sum_{i=1}^A \frac{\alpha_i}{r_{ik}^3} + b_5 \cdot \sum_{i=1}^A \frac{D_i}{r_{ik}^6}$$

where N_G and A are the numbers of grid points and molecule atoms, respectively; R_k the distance between the center of mass of the molecule and the k th grid point; r_{ik} the distance between the i th atom and the k th grid point; q_i the net atomic charge of the i th atom; and α and D the atomic polarizabilities and the dispersion coefficients. $b_0, \dots, b_5$ are regression coefficients to be determined by multivariate regression analysis.

The atomic polarizability is the charge dependent effective atomic polarizability (CDEAP) calculated by an empirical method as a linear function of the net atomic charge q_i :

$$\alpha_i = \alpha_i^0 - a_i \cdot q_i$$

where α^0 and a are the effective atomic polarizability of a neutral atom and the charge coefficient, respectively [No, Cho *et al.*, 1993]. The **atomic dispersion coefficient** D is calculated by the Slater–Kirkwood formula [Slater and Kirkwood, 1931]:

$$D_i = \frac{3}{4} \cdot \left(\frac{e \cdot h}{m \cdot \sqrt{e}} \right) \cdot \frac{\alpha_i^2}{\sqrt{\alpha_i / N_{el}}}$$

where h , m , and e are the Planck constant, the mass, and the charge of the electrons; N_{el} is the number of effective electrons of the i th atom.

The hydration free energy ΔG_{HYD} is obtained by summing over the HFED within a threshold distance r_t :

$$\Delta G_{HYD} = \sum_{k=1}^{N_G} g_k$$

where N_G is the number of grid points. The quantity ΔG_{HYD} is the scalar representation of the field around the molecule given by the hydration free energy density; to encode information on the spatial distribution of this physical property the **free energy of hydration density tensor** was also proposed. The elements of the tensor in Cartesian coordinates are defined as

$$\vec{g}_{xx} = \frac{1}{2} \cdot \sum_{k=1}^{N_G} g_k \cdot (2x_k^2 - y_k^2 - z_k^2)$$

$$\vec{g}_{xy} = \frac{3}{2} \cdot \sum_{k=1}^{N_G} g_k \cdot x_k \cdot y_k$$

where $\vec{g}_{xx}$ and $\vec{g}_{xy}$ are xx and xy components of the tensor, and x , y , and z the coordinates of the grid point.

- **hydride group** → molecular graph
- **hydrogen-bond acceptors** → hydrogen-bonding descriptors
- **hydrogen-bond acceptor number** → hydrogen-bonding descriptors
- **hydrogen-bond charged partial surface area descriptors** → charged partial surface area descriptors
- **hydrogen bond acidity** → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **hydrogen bond basicity** → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **hydrogen-bond donor number** → hydrogen-bonding descriptors
- **hydrogen-bond donors** → hydrogen-bonding descriptors
- **hydrogen-bond electron-acceptor power** $\equiv$ *hydrogen bond basicity* → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **hydrogen-bond electron-drawing power** $\equiv$ *hydrogen bond acidity* → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **hydrogen-bond index** → hydrogen-bonding descriptors
- **hydrogen-bonding ability constants** → hydrogen-bonding descriptors

■ **hydrogen-bonding descriptors** ($\equiv$ *H-bonding descriptors*)

The hydrogen bond is the bond arising from the interaction between a hydrogen and an electron donor atom, such as oxygen and nitrogen; hydrogen-bonding modifies the electron distribution of the neighbor of the electron-donor atom, thus influencing reactivity. Hydrogen-bonding causes an association of molecules, that is, large aggregates of single molecules. This association influences several → *physico-chemical properties*, such as the compressibility factor, vaporization energy, density, surface tension, parachor, conductivity, dielectric constant, molar refractivity, and boiling and melting points. Moreover, the hydrogen-bonding ability of molecules has long been recognized as being very important in biological reactions, including drug actions.

The theory of hydrogen bonding was fully discussed by Pimentel and McClellan [Pimentel and McClellan, 1960] and Vinogradov and Linnell [Vinogradov and Linnell, 1971].

Intramolecular and intermolecular hydrogen bonds can occur in biological and chemical systems. Moreover, functional groups in the molecule can be distinguished into **Hydrogen-Bond Donor** (HBD) and **Hydrogen-Bond Acceptor** (HBA), the former group having strong electron-withdrawing substituents such as $-\text{OH}$, $-\text{NH}$, $-\text{SH}$, and $-\text{CH}$ and the latter different groups such as $-\text{PO}$, $-\text{SO}$, $-\text{CO}$, $-\text{N}$, $-\text{O}$, $-\text{S}$, and $-\text{F}$; even a π -electron rich system can be considered a H-bond acceptor. Some groups are amphiprotic, that is, with both acceptor and donor ability, such as $-\text{OH}$ and $-\text{NH}$ [Gutmann, 1978].

Hydrogen-bonding ability within a congeneric series of compounds having a common H-bond acceptor or donor can be correlated with the electronic effects of substituents using either

→ *electronic substituent constants* or other physico-chemical properties such as pK_a values. In particular, pK_a is the → *acid dissociation constant*, that is, a measure of the extent of ionization of weakly acid organic compounds. Most approaches for estimating pK_a are based on → *group contribution methods*; other approaches are based on quantum-chemical calculations [Grüber and Buss, 1989; Dixon and Jurs, 1993; Sixt, Altschuh *et al.*, 1996; Schüürmann, Segner *et al.*, 1997; Duboc, 1978; Amić, Davidović-Amić *et al.*, 1999].

Hydrogen-bonding descriptors were introduced in → *Hansch analysis* as well as in → *grid-based QSAR techniques* in the form of → *hydrogen-bonding fields*; moreover, → *hydrogen bond acidity* and → *hydrogen bond basicity* scales, which are among the → *solvatochromic parameters*, were derived both for solutes and solvents by an empirical approach. → *Hydrogen-bond charged partial surface area descriptors* were derived from → *computational chemistry* based on surface areas and partial charges of HBA and HBD atoms or groups.

In analogy with solvatochromic parameters but based on quantum theoretical chemistry, a set of → *quantum-chemical descriptors* intended to describe the hydrogen bonding effects of molecules by → *Theoretical Linear Solvation Energy Relationships* (TLSE) were proposed.

Moreover, H-bond donor ability was estimated by using atomic charge on the most positively charged hydrogen atom in the molecule (Q_H) in conjunction with the → *lowest unoccupied molecular orbital energy* ϵ_{LUMO} ; in an analogous way, H-bond acceptor ability was estimated by using the charge of the most negatively charged atom which is also capable of hydrogen bonding (Q_{MN}) in conjunction with the → *highest occupied molecular orbital energy* ϵ_{HOMO} [Dearden and Ghafourian, 1995; Urrestarazu Ramos, Vaes *et al.*, 1998]. H-bond donor ability was also estimated by using electron → *donor superdelocalizability* S^+ and → *self-atom polarizability* P [Dearden, Cronin *et al.*, 1997].

The other most popular hydrogen-bonding descriptors are listed below. Reviews about hydrogen-bonding parameters are [Hadzi, Kidrić *et al.*, 1990; Dearden and Ghafourian, 1999; Winiwarter, Ax *et al.*, 2003].

• HB parameter

The simplest hydrogen-bonding molecular descriptor, defined as binary variable and accounting for the general ability of the molecule to give hydrogen bonds [Fujita, Nishioka *et al.*, 1977]. A modified HB parameter was proposed [Charton and Charton, 1982] on the basis of the number of hydrogen bonds that a molecule or substituent is capable of forming; for example, HB($-NH_2$) = 2 as a proton donor and HB($-NH_2$) = 1 as a proton acceptor, and HB($-OH$) = 1 as a proton donor and HB($-OH$) = 2 as a proton acceptor.

• I_{HA} and I_{HD} parameters

The simplest hydrogen-bonding substituent descriptors defined as binary variables accounting for the ability of the substituent to give hydrogen bonds [Hansch and Leo, 1979]. I_{HA} is equal to one if the substituent includes at least one H-bond acceptor, otherwise, zero. In the same way, I_{HD} is equal to one if the substituent includes at least one H-bond donor, otherwise, zero.

• Hydrogen-Bond Acceptor number (HBA)

A measure of the hydrogen-bonding ability of a molecule expressed in terms of the number of possible hydrogen-bond acceptors. In particular, it is calculated as the count of lone pairs on oxygen and nitrogen atoms in the molecule.

- **Hydrogen-Bond Donor number (HBD)**

A measure of the hydrogen-bonding ability of a molecule expressed in terms of the number of possible hydrogen-bond donors. In particular, it is calculated as the count of hydrogen atoms bonded to oxygen and nitrogen atoms in the molecule.

📖 [Winiwarter, Bonham *et al.*, 1998]

- **HB₁ and HB₂ parameters**

Substituent descriptors of the hydrogen-bonding ability of functional groups, defined by a set of rules [Yang, Lien *et al.*, 1986].

HB₁ parameter is a count descriptor based on atoms in a group, which possess the ability to form hydrogen bonds. This includes both H-bond acceptors and H-bond donors. The rules for HB₁ are

- oxygen atom is counted as 1, but as zero in $-\text{OCF}_3$;
- nitrogen atom is counted as 1, but as zero if it is bonded to an oxygen atom; moreover, the fragment $\text{N}-\text{N}$ is counted as 1 and the fragment $-\text{N}_3$ as zero;
- hydrogen atom when bonded to oxygen or nitrogen atoms is counted as 1; moreover, it is also counted as 1 in $-\text{C}\equiv\text{C}-\text{H}$ fragment.

For example, HB₁ values for $-\text{NO}$, $-\text{NO}_2$, $-\text{SO}_2\text{NHCH}_3$, and $-\text{CONH}_2$ are 1, 2, 4, and 4, respectively.

The HB₂ parameter is defined as the number of atoms in a group able to form hydrogen bonds multiplied by the value of the strength of hydrogen bond, then divided by 10. The total number of the H-bond acceptors and H-bond donors is calculated by following the rules defined for HB₁. The multiplicative parameters accounting for H-bond strength are 6.05 for oxygen atoms, 5.5 for nitrogen atoms, and 2.5 for hydrogen atoms.

For example, HB₂ values for $-\text{NO}$, $-\text{NO}_2$, $-\text{SO}_2\text{NHCH}_3$, and $-\text{CONH}_2$ are 0.61, 1.21, 2.01, and 1.66, respectively.

📖 [Basak, Niemi *et al.*, 1990b; Basak, 1990]

- **hydrogen-bonding ability constants (I_{H})**

Substituent descriptors defined by measuring the additive contributions of molecular fragments to the hydrogen bonding ability of a molecule [Seiler, 1974]. Such descriptors are calculated from the difference in $\log P$ value in two solvent/water systems:

$$\Delta \log P = \log P_{\text{octanol}} - \log P_{\text{solvent}} = b_0 + \sum_k (I_{\text{H}})_k \cdot N_k$$

where $(I_{\text{H}})_k$ and N_k are the hydrogen-bonding ability constant and the number of occurrences of the k th fragment in the molecule, respectively. b_0 and $(I_{\text{H}})_k$ are regression coefficients estimated by multivariate regression analysis. These substituent descriptors reflect both H-bond donor and H-bond acceptor ability, and are also a function of molecule polarity [Dearden and Ghafourian, 1999].

A set of 23 hydrogen-bonding constants was determined using octanol/water and cyclohexane/water systems; the calculated model was derived from 195 compounds, with intercept $b_0 = -0.16$, $r^2 = 0.935$ and $s = 0.333$. Some I_{H} substituent values are reported in Table H1.

Table H1 Hydrogen-bonding ability constants for some substituent groups.

Substituent	I _H
–N=N–NH– (triazole)	4.24
Aliphatic –COOH	2.88
Aromatic –COOH	2.87
Aromatic –OH	2.60
–CONH–	2.56
–SO ₂ NH–	1.93
Aliphatic –OH	1.82
Aliphatic –NH ₂	1.33
Aromatic –NH ₂	1.18
=N–	1.01
–CO–CH ₂ –CO	0.59
–NR ₁ R ₂ (R ₁ , R ₂ ≠ H)	0.55
–NO ₂	0.45
>C=O	0.31
–C≡N	0.23
–O–	0.11
Ortho-substitution to –OH, –COOH, –NR ₁ R ₂	–0.62

Moreover, $\Delta \log P$ was proposed as a molecular descriptor for modeling hydrogen bonding capacity; it corresponds to the difference between the partition coefficient experimentally determined in octanol/water ($\log P_{ow}$) and partition coefficients determined in other systems, such as octanol/water–heptane/water ($\log P_{alk}$) systems and octanol/water–chloroform/water ($\log P_{CH_3Cl}$) systems [El Tayar, Tsai *et al.*, 1991b; Winiwarter, Ax *et al.*, 2003].

• Raevsky H-bond indices

A set of descriptors characterizing relative H-bond donor and H-bond acceptor abilities of compounds calculated to reproduce the free energy ΔG and enthalpy ΔH of the hydrogen bond complex formation as defined in the thermodynamic equation:

$$\Delta G = \Delta H - T \cdot \Delta S$$

where ΔS is the entropy of complexation and T is the temperature in Kelvin degrees. ΔG was thought of as a multiplicative function of H-bond donor and H-bond acceptor ability as

$$\Delta G = b_0 + b_1 \cdot C_{HD} \cdot C_{HA}$$

where b_0 and b_1 are regression coefficients and C_{HD} and C_{HA} are the free energy H-bond donor and H-bond acceptor factors, respectively [Raevsky, Grigor'ev *et al.*, 1992a; Raevsky, 1997a]. Based on known experimental ΔG values, C_{HD} and C_{HA} values were estimated for 414 and 1298 compounds, respectively, by using the HYBOT program. A value of $C_{HA} = 4.00$ was selected for standard H-bond acceptor (hexamethylphosphoramide) and a value of $C_{HD} = -2.50$ was selected for standard H-bond donor (phenol).

The enthalpy contributions were also estimated and H-bond donor E_{HD} and H-bond acceptor E_{HA} enthalpy factors were also calculated.

Based on these four H-bond factors, Raevsky H-bond indices were therefore proposed as

$$\begin{array}{cccc}
 C_{\text{HD}}^{\text{max}} & C_{\text{HA}}^{\text{max}} & E_{\text{HD}}^{\text{max}} & E_{\text{HA}}^{\text{max}} \\
 \sum C_{\text{HD}} & \sum C_{\text{HA}} & \sum E_{\text{HD}} & \sum E_{\text{HA}} \\
 \frac{\sum C_{\text{HD}}}{\text{MW}} & \frac{\sum C_{\text{HA}}}{\text{MW}} & \frac{\sum E_{\text{HD}}}{\text{MW}} & \frac{\sum E_{\text{HA}}}{\text{MW}}
 \end{array}$$

where the first four descriptors are free energy and enthalpy factors for the strongest H-bond donor atom and H-bond acceptor atom in the molecule; the second set is based on the sums of the free energy and enthalpy factors for all H-bond donor atoms and H-bond acceptor atoms in the molecule; and the third set is constituted by the second set normalized on the molecular weight MW.

📖 [Raevsky, Grigor'ev *et al.*, 1992b, 1993; Schneider, Rüdiger *et al.*, 1993; Raevsky, Dolmatova *et al.*, 1995; Raevsky, 1999; Raevsky, Fetisov *et al.*, 2000; Schaper, Zhang *et al.*, 2001; Raevsky and Skvortsov, 2002]

• Hydrogen-Bond Index (HBI)

An empirical index proposed for chloro–fluoro hydrocarbons (CFC) and defined as [Toropov and Toropova, 2004]

$$HBI = 5000 + N_{\text{H}} - N_{\text{Cl}} - N_{\text{F}}$$

where N_{H} , N_{Cl} , and N_{F} are the number hydrogens, chlorine, and fluorine atoms, respectively; the offset 5000 was added to numerically distinguish this descriptor from other descriptors.

A simple generalization of this index can be proposed taking into account all the halogen atoms in a molecule rather than only fluorine atoms:

$$HBI = 5000 + N_{\text{H}} - N_{\text{Halogens}}$$

• HNSO_T index

A hydrogen-bonding descriptor calculated by adding the total number of lone pairs on oxygen, nitrogen, and sulfur atoms to the number of hydrogen atoms that can be donated by O, N, and S atoms of the molecule in a hydrogen-bonding interaction [MOE – Chemical Computing Group, Inc., 1999; Deretey, Feher *et al.*, 2002].

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **hydrogen bonding fields** → molecular interaction fields
- **hydrogen-bond parameters** → Linear Solvation Energy Relationships
- **hydrogen-depleted molecular graph** $\equiv$ *H-depleted molecular graph* → molecular graph
- **hydrogen electrotopological state index** → electrotopological state indices
- **hydrogen-filled molecular graph** $\equiv$ *H-filled molecular graph* → molecular graph
- **hydrogen-included molecular graph** $\equiv$ *H-filled molecular graph* → molecular graph
- **hydrophathic atom constants** → molecular interaction fields (⊙ hydrophobic fields)

- **hydropathic interactions** $\equiv$ *Kellogg and Abraham interaction field* $\rightarrow$ molecular interaction fields ($\odot$ hydrophobic fields)
- **hydropathy** $\rightarrow$ molecular interaction fields ($\odot$ hydrophobic fields)
- **hydrophilic effect** $\equiv$ *Moriguchi polar parameter* $\rightarrow$ lipophilicity descriptors

■ hydrophilicity index (H_y)

A simple empirical index related to hydrophilicity of compounds based on $\rightarrow$ *count descriptors* [Todeschini, Vighi *et al.*, 1997]. It is defined as

$$H_y = \frac{(1 + N_{Hy}) \cdot \log_2(1 + N_{Hy}) + N_C \cdot \left(\frac{1}{A} \cdot \log_2 \frac{1}{A}\right) + \sqrt{\frac{N_{Hy}}{A^2}}}{\log_2(1 + A)}$$

where N_{Hy} is the number of hydrophilic groups ($-\text{OH}$, $-\text{SH}$, $-\text{NH}$), N_C the number of carbon atoms, and A the number of atoms (hydrogen excluded). The lowest value of the H_y index is -1 for alkanes with an infinite number of carbon atoms (Table H2).

Table H2 Hydrophilicity values H_y for some compounds.

Compound	N_{Hy}	N_C	A	H_y
2,3,4,5,6-hydroxyphenol	6	6	12	4.881
H_2O_2	2	0	2	3.446
H_2O	1	0	1	3.000
4-OH	4	4	8	3.268
Triols	3	3	6	2.492
Carbonic acid	2	3	6	1.317
Diols	2	2	4	1.769
Methanol	1	1	2	1.262
Ethanol	1	2	3	0.638
Decanediol	2	10	12	0.509
Propanol	1	3	4	0.323
Butanol	1	4	5	0.132
Pentanol	1	5	6	0.004
Methane	0	1	1	0.000
$N_{Hy} = 0$ and $N_C = 0$	0	0	2	0.000
Decanol	1	10	11	-0.294
Ethane	0	2	2	-0.631
Pentane	0	5	5	-0.898
Decane	0	10	10	-0.960
Alkane with $N_C = 1000$	0	1000	1000	-1.000

📖 [Gramatica, Corradi *et al.*, 2000; Yao, Zhang *et al.*, 2002; Put, Perrin *et al.*, 2003; Jelcic, 2004; Stanton, Mattioni *et al.*, 2004; Hancock, Put *et al.*, 2005]

- **hydrophilic-lipophilic balance** $\rightarrow$ grid-based QSAR techniques ($\odot$ VolSurf descriptors)
- **hydrophobic fields** $\rightarrow$ molecular interaction fields

- **hydrophobic fragmental constants** $\equiv$ *Nys-Rekker hydrophobic fragmental constants* $\rightarrow$ lipophilicity descriptors
- **hydrophobicity** $\rightarrow$ lipophilicity descriptors
- **hydrophobic substituent constants** $\equiv$ *Hansch-Fujita hydrophobic substituent constants* $\rightarrow$ lipophilicity descriptors
- **hydropoles** $\rightarrow$ Comparative Molecular Moment Analysis
- **hyper-Cluj-detour index** $\rightarrow$ Cluj matrices
- **hyper-Cluj-distance index** $\rightarrow$ Cluj matrices
- **hyperconjugation effect** $\rightarrow$ electronic substituent constants
- **hyper-detour index** $\rightarrow$ detour matrix
- **hyper-distance-path index** $\rightarrow$ distance-path matrix
- **hyper-Harary distance index** $\rightarrow$ Harary indices
- **hyper-Harary indices** $\rightarrow$ Harary indices
- **hyper-Harary numbers** $\equiv$ *hyper-Harary indices* $\rightarrow$ Harary indices
- **hypermolecule** $\rightarrow$ hyperstructure-based QSAR techniques
- **hyperstructure** $\rightarrow$ hyperstructure-based QSAR techniques

■ hyperstructure-based QSAR techniques

These are QSAR techniques based on the construction of a **hyperstructure** defined as a virtual structure built by overlapping the training set structures such that some atoms and bonds of different structures coincide.

A hyperstructure built by overlapping molecular graphs is called **molecular supergraph** (MSG) and it can be considered as a certain graph such that each training set structure can be represented as its subgraph. A 3D hyperstructure based on the $\rightarrow$ *molecular geometry* of the training set compounds is called **hypermolecule**.

The most important QSAR techniques based on a hyperstructure are $\rightarrow$ *minimal topological difference*, $\rightarrow$ *DARC/PELCO analysis*, and $\rightarrow$ *molecular field topology analysis*.

- **hyper-Szeged index** $\rightarrow$ Szeged matrices
- **hyper-Wiener index** $\rightarrow$ Wiener matrix

■ hyper-Wiener-type indices

These are molecular descriptors calculated from the $\rightarrow$ *H-depleted molecular graph* by analogy with the $\rightarrow$ *hyper-Wiener index*. The general formula, called by Ivanciuc **hyper-Wiener operator** [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc, 2001c], to calculate hyper-Wiener-type indices from vertex- and/or edge-weighted molecular graphs is

$$HyWi(\mathbf{M}; w) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=i}^A ([\mathbf{M}(w)]_{ij}^2 + [\mathbf{M}(w)]_{ij})$$

where A is the number of graph vertices and $\mathbf{M}(w)$ is any square symmetric $\rightarrow$ *graph-theoretical matrix*, calculated with the $\rightarrow$ *weighting scheme* w . Note that, unlike the definition of the original hyper-Wiener index proposed by Klein [Klein, Lukovits *et al.*, 1995], also the diagonal elements of the considered matrix $\mathbf{M}$, which are usually different from zero in the case of weighted graphs, are taken into account in the hyper-Wiener operator.

If $\mathbf{M}$ is the $\rightarrow$ *distance matrix*, the classical hyper-Wiener index is obtained, $HyWi(\mathbf{D}) = WW$.

An example of the hyper-Wiener-type index is the **Lu index** proposed to describe multiple bond and heteroatom-containing molecules [Lu, Guo *et al.*, 2006a, 2006b, 2006c]. It is calculated from the $\rightarrow$ *bond length-weighted distance matrix* $\mathbf{D}(r^*)$ and defined as

$$Lu \equiv HyWi[\mathbf{D}(r^*)] = A^{1/2} \cdot \log \left[\frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [(d_{ij}(r^*))^2 + d_{ij}(r^*)] \right]$$

where $d_{ij}(r^*)$ is the bond length-weighted interatomic distance calculated by adding the relative bond lengths of the edges along the shortest path. The relative bond length r^* is calculated as the ratio of each bond length r_{ij} over the bond length of C–C bond (1.54 Å).

Note that, unlike the definition of the original hyper-Wiener index, a log transformation is applied and the summations are over all the matrix elements and not only on the effective different distances in the graph.

The Lu index was demonstrated to well correlate with boiling point, molar refraction, and gas heat capacity of a number of organic compounds including aliphatic aldehydes and ketones.

Another hyper-Wiener-type index is the $\rightarrow$ *resistance distance hyper-Wiener index*.

➤ **hyper-Wiener operator** $\rightarrow$ hyper-Wiener-type indices

- **I_{HA} and I_{HD} parameters** → hydrogen-bonding descriptors
- **I_R aromaticity indices** $\equiv$ *Bird aromaticity indices* → delocalization degree indices
- **Iball index** → biological activity indices
- **ICD descriptors** → molecular descriptors ($\odot$ invariance properties of molecular descriptors)
- **IDentification numbers** $\equiv$ *ID numbers*
- **identity matrix** → algebraic operators

■ ID numbers ($\equiv$ IDentification numbers)

Molecular ID numbers (MID) are molecular descriptors defined as → *weighted path counts* or weighted → *walk counts*, mainly proposed to univocally identify a molecule by a single real number, the aim being of obtaining highly discriminatory power suitable for chemical documentation, storage, and retrieval [Randić, 1984b; Szymanski, Müller *et al.*, 1986a; Carter, Trinajstić *et al.*, 1987].

These indices are usually calculated by assigning a weighting factor w_{ij} to each → *path* or → *walk* of the → *molecular graph* with v_i and v_j as endpoints. By adding all the weighted paths (or walks) starting from the i th vertex, **atomic ID numbers (AID)** are generated as → *local vertex invariants* characterizing the atomic environment.

In most the cases, molecular identification numbers are calculated as the half-sum of atomic identification numbers over all graph vertices. Otherwise, by adding atomic ID numbers only to the atoms belonging to molecular fragments such as functional groups, **fragment ID numbers** (or **subgraph ID numbers**) are obtained.

A list of molecular identification numbers is given below. Other important ones are → *Hosoya ID number*, → *hyper-Wiener index*, → *restricted walk ID number*, and → *total topological state*, defined elsewhere.

Note. To avoid confusion among different acronyms, some ID number names have been changed with respect to the original author definition. In particular, the Randić Connectivity ID number was changed from ID into CID, the ID numbers proposed by Balaban from SID to BID, from MINID to MINCID, and from MINSID to MINBID.

• **Randić Connectivity ID number (CID)** ($\equiv$ *Connectivity ID number*)

It is the molecular identification number proposed first [Randić, 1984b; Szymanski, Müller *et al.*, 1985; Randić and Trinajstić, 1993a] and is defined as a weighted molecular path count:

$$\text{CID} = A + \sum_{m,p_{ij}} w_{ij}$$

where A is the number of graph vertices, ${}^m p_{ij}$ denotes a path of length m (i.e., a sequence of m edges) from the vertex v_i to vertex v_j , and w_{ij} is the path weight. The sum runs over all paths of the graph; each path of length zero is given a unit weight.

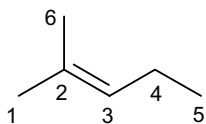
The weight w_{ij} is calculated by multiplying the $\rightarrow$ edge connectivity (Table I1) of all m edges of the path ${}^m p_{ij}$ as

$$w_{ij} = \prod_{b=1}^m (\delta_{b(1)} \cdot \delta_{b(2)})^{-1/2}_b$$

where $\delta_{b(1)}$ and $\delta_{b(2)}$ are the $\rightarrow$ vertex degree of the two vertices incident to the b th edge and b runs over all the m edges of the path. The use of edge weights smaller than one results in a gradual attenuation of the role of paths of longer lengths, therefore the Randić ID number, unlike the molecular path count, is a graph invariant in which local features are more pronounced.

Example I1

Weighted path counts and Randić connectivity ID numbers for 2-methyl-2-pentene.



Atom	${}^0 p_i^w$	${}^1 p_i^w$	${}^2 p_i^w$	${}^3 p_i^w$	${}^4 p_i^w$	p_i^w
1	1	0.5774	1.3905	0.1178	0.0680	3.1537
2	1	1.5630	0.2041	0.1178	—	2.8849
3	1	0.9082	0.7601	—	—	2.6683
4	1	1.0774	0.2041	0.2357	—	2.5172
5	1	0.5774	0.2887	0.1178	0.1361	2.1200
6	1	0.5774	1.3905	0.1178	0.0680	3.1537
	${}^0 p^w$	${}^1 p^w$	${}^2 p^w$	${}^3 p^w$	${}^4 p^w$	CID
Total	6	2.6404	2.1190	0.3535	0.1361	11.2489

• Prime ID number (PID)

This is a modification of the $\rightarrow$ Randić Connectivity ID number, aimed at improving the discriminating power [Randić, 1986b]. The Prime ID number is analogously defined as

$$\text{PID} = A + \sum_{m,p_{ij}} w_{ij}$$

where A is the number of graph vertices, ${}^m p_{ij}$ denotes a path of length m (i.e., a sequence of m edges) from the vertex v_i to vertex v_j , and w_{ij} is the path weight. The summation goes over all paths of the graph. Unlike Randić ID number, weight w_{ij} is calculated by multiplying the edge weights of all m edges of the path ${}^m p_{ij}$ defined in a different way:

$$w_{ij} = \prod_{b=1}^m p n_b^{-1/2}$$

where $p n_b$ is a prime number associated with the edge b , according to the scheme of Table I1, that is, the prime number is chosen according to the $\rightarrow$ vertex degrees of the

that is, the sum of all weighted paths starting from vertex v_i of any length, including zero. The weight w_{ij} is calculated by multiplying the conventional bond order π^* of all m edges of the path ${}^m p_{ij}$:

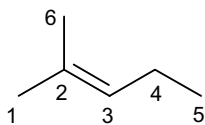
$$w_{ij} = \prod_{b=1}^m \pi_b^*$$

This ID number accounts for multiple bonds in the molecule; for saturated molecules each bond weight is equal to one, therefore the ID number coincides with the $\rightarrow$ *total path count*.

An alternative ID number is calculated in the same way using $\rightarrow$ *fractional bond order* instead of conventional bond order to accomplish a gradual attenuation of the role of paths of longer lengths.

Example I3

Weighted path counts and conventional bond order ID numbers for 2-methyl-2-pentene.



Atom	${}^0 p_i^w$	${}^1 p_i^w$	${}^2 p_i^w$	${}^3 p_i^w$	${}^4 p_i^w$	p_i^w
1	1	1	3	2	2	9
2	1	4	2	2	—	9
3	1	3	5	—	—	9
4	1	2	2	4	—	9
5	1	1	1	2	4	9
6	1	1	3	2	2	9
	${}^0 p^w$	${}^1 p^w$	${}^2 p^w$	${}^3 p^w$	${}^4 p^w$	π ID
Total	6	6	8	6	4	30

• Ring ID number (RID)

This is the $\rightarrow$ *Randić Connectivity ID number* (CID) restricted to path contributions from vertices of a ring in the graph, that is, it is calculated by summing up the atomic ID numbers of the vertices belonging to the selected ring [Randić, 1988c]. The ring ID number is a particular case of $\rightarrow$ *fragment ID numbers*.

• Balaban ID number (BID)

This is a molecular identification number defined as [Balaban, 1987]

$$\text{BID} = A + \sum {}^m p_{ij} w_{ij}$$

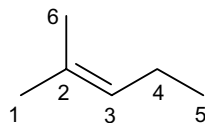
where A is the number of graph vertices, ${}^m p_{ij}$ denotes a path of length m (i.e., a sequence of m edges) from the vertex v_i to vertex v_j , and w_{ij} is the path weight. The summation goes over all paths of the graph. The weight w_{ij} is calculated by multiplying the edge weights of all m edges of the path ${}^m p_{ij}$ by the following:

$$w_{ij} = \prod_{b=1}^m (\sigma_{b(1)} \cdot \sigma_{b(2)})_b^{-1/2}$$

where $\sigma_{b(1)}$ and $\sigma_{b(2)}$ are the $\rightarrow$ *vertex distance degree* of the two vertices incident to the edge b and b runs over all the m edges of the path.

Example I4

Weighted path counts and Balaban ID numbers for 2-methyl-2-pentene.



Atom	${}^0P_i^w$	${}^1P_i^w$	${}^2P_i^w$	${}^3P_i^w$	${}^4P_i^w$	P_i^w
1	1	0.1021	0.0232	0.0014	0.0001	1.1268
2	1	0.3292	0.0140	0.0012	—	1.3444
3	1	0.2368	0.0350	—	—	1.2718
4	1	0.1963	0.0140	0.0029	—	1.2132
5	1	0.0845	0.0094	0.0012	0.0002	1.0953
6	1	0.1021	0.0232	0.0014	0.0001	1.1268
	${}^0P^w$	${}^1P^w$	${}^2P^w$	${}^3P^w$	${}^4P^w$	BID
Total	6	0.5255	0.0594	0.0041	0.0002	6.5892

• MINCID

This is a molecular ID number of a graph with A vertices derived from the $\rightarrow$ *Randić Connectivity ID number* with the aim of obtaining much faster calculations, also for large polycyclic graphs [Ivanciuc and Balaban, 1996b].

It is analogously defined as a weighted molecular path count:

$$\text{MINCID} = A + \sum_{\min p_{ij}} w_{ij}$$

where A is the number of graph vertices, $\min p_{ij}$ denotes the shortest path (i.e., $\rightarrow$ *geodesic*) of length d_{ij} from the vertex v_i to vertex v_j , and w_{ij} is the path weight. The sum runs over all the shortest paths of the graph.

The weight w_{ij} is calculated by multiplying the $\rightarrow$ *edge connectivity* of all edges of the shortest path $\min p_{ij}$:

$$w_{ij} = \prod_{b=1}^{d_{ij}} (\delta_{b(1)} \cdot \delta_{b(2)})_b^{-1/2}$$

where $\delta_{b(1)}$ and $\delta_{b(2)}$ are the $\rightarrow$ *vertex degrees* of the two vertices incident to the b th edge and b runs over all the d_{ij} edges of the path.

For cycle-containing molecular graphs, the set of geodesics represents a subset of the set of paths, while, for acyclic graphs, the two sets are identical; therefore, for a graph G ,

$$\text{CID}(G) > \text{MINCID}(G) \quad \text{for a cyclic graph}$$

$$\text{CID}(G) = \text{MINCID}(G) \quad \text{for an acyclic graph}$$

• MINBID

This is a molecular ID number of a graph with A vertices derived from the $\rightarrow$ *Balaban ID number* with the aim of obtaining much faster calculations, also for large polycyclic graphs [Ivanciuc and Balaban, 1996b]. It is analogously defined as a weighted molecular path count:

$$\text{MINCID} = A + \sum_{\min p_{ij}} w_{ij}$$

where A is the number of graph vertices, ${}^{\min}p_{ij}$ denotes the shortest path (i.e., $\rightarrow$ *geodesic*) of length d_{ij} from the vertex v_i to vertex v_j , and w_{ij} is the path weight. The sum runs over all the shortest paths of the graph. The weight w_{ij} is calculated by multiplying the edge weights of all edges of the path ${}^{\min}p_{ij}$

$$w_{ij} = \prod_{b=1}^{d_{ij}} (\sigma_{b(1)} \cdot \sigma_{b(2)})_b^{-1/2}$$

where $\sigma_{b(1)}$ and $\sigma_{b(2)}$ are the $\rightarrow$ *vertex distance degrees* of the two vertices incident to the edge b and b runs over all the d_{ij} edges of the path.

For cycle-containing molecular graphs, the set of geodesics represents a subset of the set of paths, while, for acyclic graphs, the two sets are identical; therefore, for a graph G ,

$$\text{BID}(G) > \text{MINBID}(G) \quad \text{for a cyclic graph}$$

$$\text{BID}(G) = \text{MINBID}(G) \quad \text{for an acyclic graph}$$

• Weighted ID number (WID)

This is a molecular ID number of a graph with A vertices defined as a function of the sum of weighted walks [Szymanski, Müller *et al.*, 1986b]. The weight w_{ij} of an edge connecting vertices v_i and v_j is the reciprocal of the square root of the product of the $\rightarrow$ *vertex distance degrees* σ of the vertices incident to the edge. These edge weights are the off-diagonal entries corresponding to pairs of adjacent vertices of the $\rightarrow$ *distance-sum-connectivity matrix*.

An auxiliary identification number ID^* is defined as

$$\text{ID}^* = \sum_{i=1}^A \sum_{j=1}^A w_{ij}^* \quad A \leq \text{ID}^* \leq A^2$$

where A is the number of vertices and w_{ij}^* are the entries of the matrix $\mathbf{W}^*$ that is obtained by adding the different k th powers of the distance-sum-connectivity matrix ${}^{\sigma}\chi$:

$$\mathbf{W}^* = \sum_{k=0}^{A-1} {}^{\sigma}\chi^k = \mathbf{I} + \sum_{k=1}^{A-1} {}^{\sigma}\chi^k$$

where ${}^{\sigma}\chi^0 = \mathbf{I}$, $\mathbf{I}$ being the $\rightarrow$ *identity matrix*.

Each ${}^{\sigma}\chi^k$ matrix contains, at position $i-j$, the sum of the weights of all walks $w_{ij(k)}$ of length k from vertex v_i to vertex v_j ; therefore

$$[{}^{\sigma}\chi^k]_{ij} = \sum_{w_{ij(k)}} \prod_{b=1}^k (\sigma_{b(1)} \cdot \sigma_{b(2)})_b^{-1/2}$$

where $\sigma_{b(1)}$ and $\sigma_{b(2)}$ are the vertex distance degrees of the two vertices incident to the edge b and b runs over all the k edges of the walk and the summation goes over all the walks of length k between the vertices v_i and v_j . The diagonal entry $i-i$ is the sum of all weighted $\rightarrow$ *self-returning walks* for vertex v_i of length k . Thus, the matrix $\mathbf{W}^*$ contains, at position $i-j$, the sum of the weights of all walks from vertex v_i to vertex v_j of length less than A ; the i th diagonal entry corresponds to the sum of the weights of all self-returning walks for vertex v_i of length less than A .

The weighted walk ID number is defined as

$$\text{WID} = A - \frac{1}{A} + \frac{\text{ID}^*}{A^2} \quad A \leq \text{WID} < A + 1$$

where $\text{ID}^* = A^2$ for complete graphs and $\text{ID}^* = A$ for a null graph.

The **Self-returning ID number** (SID) is a molecular ID number of a graph defined as the sum of weighted self-returning walks of any length [Müller, Szymanski *et al.*, 1993]:

$$\text{SID} = \sum_{i=1}^A w_{ii}^*$$

where w_{ii}^* are the diagonal entries of the matrix $\mathbf{W}^*$.

• **Extended Adjacency ID number** (EAID)

This is a complex ID number calculated by the powers of an extended adjacency matrix, derived by weighting the vertices of the graph [Hu and Xu, 1996].

Let $\mathbf{W}$ be the edge-weighted matrix defined as

$$w_{ij} = \begin{cases} \sqrt{\frac{S_i}{S_j}} + \sqrt{\frac{S_j}{S_i}} & \text{if } (i, j) \in \mathcal{E}(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{E}(\mathcal{G})$ is the set of graph edges; the local vertex invariant S_i accounting for heteroatoms is defined as

$$S_i = lcv_{i0} + \sum_{k=1}^D lcv_{ik} \cdot lcb_{ik} \cdot 10^{-k}$$

where lcv_{ik} and lcb_{ik} are the entries of the $\rightarrow$ *connectivity valence layer matrix* $\mathbf{LCV}$ and the $\rightarrow$ *connectivity bond layer matrix* $\mathbf{LCB}$, respectively, and D is the number of layers around the focused i th vertex.

An $\rightarrow$ *extended adjacency matrix* $\mathbf{EA}$ is then defined by the following:

$$[\mathbf{EA}]_{ij} = \begin{cases} (\sqrt{\pi_{ij}^*} \cdot w_{ij})/6 & \text{if } (i, j) \in \mathcal{E}(\mathcal{G}) \\ \sqrt{R_i}/6 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where R_i is the $\rightarrow$ *covalent radius* of the i th atom and π_{ij}^* is the $\rightarrow$ *conventional bond order* of the i - j edge; w_{ij} are the elements of the edge-weighted matrix defined above.

The extended adjacency ID number EAID is calculated as

$$\text{EAID} = \sum_{i=1}^A [\mathbf{EA}^*]_{ii}$$

where $[\mathbf{EA}^*]_{ii}$ are the diagonal entries of the matrix $\mathbf{EA}^*$ that is obtained by adding A power matrices $\mathbf{EA}^k$:

$$\mathbf{EA}^* = \sum_{k=0}^{A-1} \mathbf{EA}^k = \mathbf{I} + \sum_{k=1}^{A-1} \mathbf{EA}^k$$

where $\mathbf{EA}^0 = \mathbf{I}$, the $\rightarrow$ identity matrix.

• **Hu–Xu ID number (HXID)**

This is a molecular ID number defined in terms of paths ${}^m p_{ij}$ of length m weighted by the following [Hu and Xu, 1997]:

$$w_{ij} = \prod_{a=2}^{m+1} \left(\frac{\pi_{a-1,a}^*}{a} \cdot \frac{1}{\delta'_{a-1} \cdot \delta'_a} \right)^{1/2}$$

where π^* is the $\rightarrow$ conventional bond order, a is the sequence number of the vertices along the path between vertices v_i and v_j , and δ' is the $\rightarrow$ Hu–Xu vertex degree, defined as

$$\delta'_a = \delta_a \cdot \sqrt{Z_a}$$

where Z_a is the atomic number of the considered atom and δ_a the $\rightarrow$ vertex degree. To avoid singularities in the path weight w_{ij} , δ_a was assigned the value $1/2$ for isolated vertices [Alikhanidi and Takahashi, 2006].

Then, the molecular ID number is defined as

$$\text{HXID} = \sum_{i=1}^A \text{AID}_i^2$$

AID being the atomic ID number obtained by adding the weights of all paths starting from the i th vertex,

$$\text{AID}_i = \sum_{j=1}^A w_{ij}$$

where w_{ij} are the path weights and j runs over all vertices different from vertex v_i .

📖 [Randić, 1979, 1984a; Hendrickson and Toczko, 1983; Razinger, 1986; Carter, Trinajstić *et al.*, 1987; Szymanski, Müller *et al.*, 1987; Elk, 1990, 1995; Elk and Gutman, 1994; Ivanciuc and Balaban, 1999c]

- **Idoux steric constant** $\rightarrow$ steric descriptors ($\odot$ number of atoms in substituent specific positions)
- **immanant** $\rightarrow$ algebraic operators ($\odot$ determinant)
- **IMPI** $\equiv$ inner molecular polarizability index $\rightarrow$ electric polarization descriptors ($\odot$ polarizability effect index)
- **Inamoto–Masuda inductive constant** $\rightarrow$ electronic substituent constants ($\odot$ inductive electronic constants)

■ incidence matrices

Together with $\rightarrow$ *vertex matrices* and $\rightarrow$ *edge matrices*, incidence matrices are other important $\rightarrow$ *graph-theoretical matrices* used to determine a molecular graph. These are matrices whose rows can represent either vertices or edges and columns some subgraphs, such as edges, paths, or cycles.

A general definition of incidence matrix $\mathbf{I}$ of a graph $\mathcal{G}$ is [Janežič, Miličević *et al.*, 2007]

$$[\mathbf{I}]_{ij} = \begin{cases} 1 & \text{if } [i \in \mathbf{I}(\mathcal{G})] \cap [j \in \mathbf{J}(\mathcal{G})] \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{I}$ and $\mathbf{J}$ are two sets of graph elements belonging to two different equivalence classes, whose elements are denoted by the indices i and j , respectively. Examples of sets $\mathbf{I}$ and $\mathbf{J}$ are the set of vertices, the set of edges, the set of vertices belonging to cycles, and the set of paths.

A list of the most common incidence matrices is given below.

• vertex-edge incidence matrix

Derived from the $\rightarrow$ *H-depleted molecular graph*, the vertex-edge incidence matrix, denoted by $^{\text{VE}}\mathbf{I}$, is a rectangular, usually unsymmetrical, matrix $A \times B$ whose rows are the vertices (A) and columns the edges (B) of a graph [Bonchev and Trinajstić, 1977]. Its elements equal one if the vertex v_i is incident to the edge e_j , and zero otherwise

$$[^{\text{VE}}\mathbf{I}]_{ij} = \begin{cases} 1 & \text{if } v_i \text{ is incident to } e_j \\ 0 & \text{otherwise} \end{cases}$$

The **total information content on the incidence matrix** I_{INC} and the **mean information content on the incidence matrix** $\bar{I}_{\text{INC}}$ are, respectively, defined as [Bonchev, 1983]

$$I_{\text{INC}} = A \cdot B \cdot \log_2 A - B \cdot (A-2) \cdot \log_2 (A-2) - 2 \cdot B$$

$$\bar{I}_{\text{INC}} = \log_2 A - \frac{A-2}{A} \cdot \log_2 (A-2) - \frac{2}{A}$$

It must be noted that the mean information content does not depend on the number and type of edges, therefore it will be the same for the $\rightarrow$ *multigraph* $\mathcal{MG}$ and the corresponding graph $\mathcal{G}$.

However, the total information content of the multigraph is always greater than that in the corresponding graph by the quantity

$$I_{\text{INC}}(\mathcal{MG}) - I_{\text{INC}}(\mathcal{G}) = A \cdot \log_2 A - (A-2) \cdot \log_2 (A-2) - 2$$

• edge-vertex incidence matrix

This is the transpose of the vertex-edge incidence matrix. Denoted by $^{\text{EV}}\mathbf{I}$, it is a rectangular unsymmetrical matrix $B \times A$ whose rows represent the edges (B) and columns the vertices (A) of a molecular graph. Its elements are equal to one if the edge e_i is incident to the vertex v_j , and zero otherwise

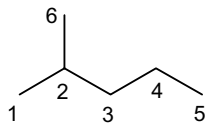
$$[^{\text{EV}}\mathbf{I}]_{ij} = \begin{cases} 1 & \text{if } v_j \text{ is incident to } e_i \\ 0 & \text{otherwise} \end{cases}$$

An interesting relationship between the vertex-edge incidence matrix, the edge-vertex incidence matrix and the $\rightarrow$ Laplacian matrix $\mathbf{L}$ was found as

$${}^{\text{VE}}\mathbf{I} \cdot {}^{\text{EV}}\mathbf{I} = \mathbf{L}$$

Example I5

Vertex-edge incidence matrix for 2-methylpentane.

	vertex/edge	(1, 2)	(2, 6)	(2, 3)	(3, 4)	(4, 5)
	1	1	0	0	0	0
	2	1	1	1	0	0
	3	0	0	1	1	0
	4	0	0	0	1	1
	5	0	0	0	0	1
	6	0	1	0	0	0

• cycle matrices

Cycle matrices are particular incidence matrices, where each column represents a graph $\rightarrow$ circuit. Two main cycle matrices are defined: the *vertex-cycle incidence matrix*, denoted as ${}^{\text{VC}}\mathbf{I}$, whose rows are the A vertices and the *edge-cycle incidence matrix*, denoted as ${}^{\text{EC}}\mathbf{I}$, whose rows are the B edges of the graph [Bonchev, 1983].

Based on total and mean information content, several $\rightarrow$ topological information indices can be calculated both from the vertex-cycle matrix ($\rightarrow$ information indices on the vertex-cycle incidence matrix) and the edge-cycle matrix ($\rightarrow$ information indices on the edge-cycle incidence matrix).

The **vertex-cycle incidence matrix** (${}^{\text{VC}}\mathbf{I}$) is a rectangular unsymmetrical matrix whose rows represent the A vertices and columns the circuits of the graph; this has dimension $A \times C^+$, where C^+ is the $\rightarrow$ cyclicity, that is, the number of circuits, and its elements are formally defined as

$$[{}^{\text{VC}}\mathbf{I}]_{ij} = \begin{cases} 1 & \text{if } v_i \text{ is incident to } j\text{th cycle} \\ 0 & \text{otherwise} \end{cases}$$

The sum over all the entries in the i th row is called **vertex cyclic degree** γ :

$$\gamma_i \equiv VS_i({}^{\text{VC}}\mathbf{I}) = \sum_{j=1}^{C^+} [{}^{\text{VC}}\mathbf{I}]_{ij}$$

where VS is the $\rightarrow$ row sum operator and C^+ is the number of circuits. The sum over all the vertex cyclic degrees is called **total vertex cyclicity** C_{VC} of the graph:

$$C_{\text{VC}} = \sum_{i=1}^A \gamma_i = \sum_{i=1}^A \sum_{j=1}^{C^+} [{}^{\text{VC}}\mathbf{I}]_{ij}$$

where A the number of graph vertices. This index increases rapidly with the number of cycles and can be used as a general descriptor for $\rightarrow$ molecular cyclicity.

The **cycle-vertex incidence matrix**, denoted as ${}^{CV}\mathbf{I}$, is the transpose of the vertex-cycle incidence matrix.

The **edge-cycle incidence matrix** (${}^{EC}\mathbf{I}$) is a rectangular unsymmetrical matrix whose rows represent the edges and columns the circuits of the graph. This matrix of dimension $B \times C^+$, where C^+ is the graph cyclicity, is formally defined as

$$[{}^{EC}\mathbf{I}]_{ij} = \begin{cases} 1 & \text{if } e_i \text{ is incident to } j\text{th cycle} \\ 0 & \text{otherwise} \end{cases}$$

The sum over all the entries in the i th row is called **edge cyclic degree** γ_i^e :

$$\gamma_i^e \equiv VS_i({}^{EC}\mathbf{I}) = \sum_{j=1}^{C^+} [{}^{EC}\mathbf{I}]_{ij}$$

where VS is the row sum operator and C^+ the cyclicity. The sum over all the edge cyclic degrees is called **total edge cyclicity** C_{EC} of the graph:

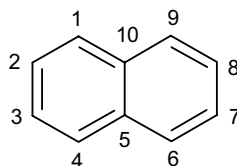
$$C_{EC} = \sum_{i=1}^B \gamma_i^e = \sum_{i=1}^B \sum_{j=1}^{C^+} [{}^{EC}\mathbf{I}]_{ij}$$

where B is the number of graph edges. Like the total vertex cyclicity, this descriptor was proposed as a measure of $\rightarrow$ *molecular cyclicity*.

The **cycle-edge incidence matrix**, denoted as ${}^{CE}\mathbf{I}$, is the transpose of the edge-cycle incidence matrix.

Example 16

Vertex-cycle and edge-cycle incidence matrices for anthracene.



	vertex/circuit	1	2	3	γ_i		edge/circuit	1	2	3	γ_i^e
$VC_I =$	1	1	0	1	2	$EC_I =$	(1,2)	1	0	1	2
	2	1	0	1	2		(2,3)	1	0	1	2
	3	1	0	1	2		(3,4)	1	0	1	2
	4	1	0	1	2		(4,5)	1	0	1	2
	5	1	1	1	3		(5,6)	1	1	1	3
	6	0	1	1	2		(6,7)	0	1	1	2
	7	0	1	1	2		(7,8)	0	1	1	2
	8	0	1	1	2		(8,9)	0	1	1	2
	9	0	1	1	2		(9,10)	0	1	1	2
	10	1	1	1	3		(1,10)	1	0	1	2
						(5,10)	1	1	0	2	

• **vertex-path incidence matrix**

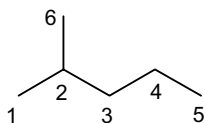
The vertex-path incidence matrix, denoted as ^{VP}I , is an extension of the $\rightarrow$ vertex-edge incidence matrix to paths of any length in the graph. The vertex-path incidence matrix is defined as [Janežič, Miličević *et al.*, 2007]

$$[^{VP}I]_{ij} = \begin{cases} n_{ij} & \text{if } v_i \cap \{^j p\} \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

where n_{ij} is the number of incidences of vertex v_i to any path $^j p$ of length j in the graph.

Example 17

Vertex-path incidence matrix for 2-methylpentane.



$$^{VP}I = \begin{array}{c|ccccc} & \text{vertex/path} & ^0p & ^1p & ^2p & ^3p & ^4p \\ \hline 1 & 1 & 1 & 1 & 2 & 1 & 1 \\ 2 & 2 & 1 & 3 & 4 & 3 & 2 \\ 3 & 3 & 1 & 2 & 4 & 3 & 1 \\ 4 & 4 & 1 & 2 & 1 & 2 & 2 \\ 5 & 5 & 1 & 1 & 1 & 1 & 2 \\ 6 & 6 & 1 & 1 & 2 & 1 & 1 \end{array}$$

- **incidences** $\rightarrow$ graph
- **indegree** $\rightarrow$ graph
- **independent edges** $\rightarrow$ graph
- **independent variables** $\rightarrow$ data set
- **indexed graph** $\rightarrow$ graph
- **index of charge and orbital controlled interaction** $\equiv$ *Interaction Index* $\rightarrow$ quantum-chemical descriptors
- **index of hydrogen deficiency** $\rightarrow$ multiple bond descriptors
- **index of refraction** $\equiv$ *refractive index* $\rightarrow$ physico-chemical descriptors

• **indicator variables (I)**

Also known as **dummy variables**, indicator variables are one of the most simple descriptors and are used when the problem of interest cannot be represented by real numerical values. They usually take positive, negative or zero integer values, indicating the states of some quantity. For example, the presence and discrimination of *cis/trans* isomers can be represented by an indicator variable such as

$$I = \begin{cases} +1 & \text{trans-isomer} \\ 0 & \text{no cis/trans-isomer} \\ -1 & \text{cis-isomer} \end{cases}$$

This indicator variable takes the value 1 when a cis-isomer is present, -1 when a trans-isomer is present and 0 when the characteristic is not applicable to the molecule, that is, when no cis/trans isomerism is present. Indicator variables are in general multivalued variables, thus are able to discriminate among multistate quantities.

The most common subcases of indicator variables are the **binary descriptors** which are bivalued variables taking the value of 1 when the considered characteristic is present in the molecule and the value of 0 when the characteristic is absent; these descriptors are usually indicated by the symbol I_{char} where *char* is the considered characteristic.

Binary descriptors are commonly used to represent

- presence/absence of a molecular fragment
- presence/absence of aromatic substructures
- presence/absence of a specified functional group
- cis/trans-isomer discrimination
- chirality discrimination.

Binary descriptors should be used when the considered characteristic is really a dual characteristic of the molecule or when the considered quantity cannot be represented in a more informative numerical form. In any case, the $\rightarrow$ *mean information content* H_{char} of a binary descriptor is low (the maximum value is 1 when the proportions of 0 and 1 are equal), thus the $\rightarrow$ *standardized Shannon's entropy* $H_{\text{char}}^* = H_{\text{char}}/\log_2 n$, where n is the number of elements, gives a measure of the efficiency of the collected information.

Bit-strings are sets of binary descriptors ($\rightarrow$ *vectorial descriptors*) and are often used as $\rightarrow$ *fingerprints* or $\rightarrow$ *structural keys* to define $\rightarrow$ *substructure descriptors* and perform quick data mining.

For specific purposes, single binary variables can also be combined using the logical operators “and” ($\wedge$) and “or” ($\vee$) [Streich and Franke, 1985].

Binary descriptors are also used in $\rightarrow$ *Free–Wilson analysis* and $\rightarrow$ *DARC/PELCO analysis*.

📖 [Silipo and Hansch, 1975; Bodor, Gabanyi *et al.*, 1989; Kim, 1993e; Franke, Rose *et al.*, 1995; King and Srinivasan, 1997; Murray, Auton *et al.*, 1998; Dearden and Ghafourian, 1999; Amić, Davidović-Amić *et al.*, 2003; Schaper, Kunz *et al.*, 2003; Xing, Glen *et al.*, 2003; Leonard and Roy, 2004; Hoover, Acree Jr. *et al.*, 2005; Roy and Leonard, 2005; Lepoittevin and Roy, 2006]

➤ **indices of central tendency** $\rightarrow$ statistical indices

■ indices of differences of path lengths

These molecular descriptors, also called **DP indices**, were proposed to measure the electron mobility within the molecule [Gálvez, García-Domenech *et al.*, 2000]. They are derived from the model of interference between electronic waves, according to which two electrons, moving through a cyclic graph within a diffraction experiments, interfere in a given vertex of the graph. It was demonstrated that the overall sum of the inverse of the squares of the differences of topological distances between all pairs of vertices of the graph is a measure of the mean global kinetic energy of the electrons that are able to give a constructive interference.

For any aromatic system, the first DP descriptor of k th order, has been defined as

$$DP_k = \sum_{i=1}^A \sum_{j=1}^A \left(p_{ij}^{(1)} - p_{ij}^{(2)} \right)^{-2} \cdot \delta(d_{ij}; k)$$

where $p_{ij}^{(1)}$ and $p_{ij}^{(2)}$ are the lengths of the two paths connecting vertices v_i and v_j , and δ is the Kronecker function equal to 1 if the topological distance between the considered vertices is equal to k .

Moreover, to take into account the differences between the atomic electronegativities, **valence DP indices** were defined as

$$DP_k^v = \sum_{i=1}^A \sum_{j=1}^A \left(p_{ij}^{(1)} - p_{ij}^{(2)} \right)^{-2} \cdot [1 + (\chi_i - \chi_j)] \cdot \delta(d_{ij}; k)$$

where χ are the Pauling atomic electronegativities.

These descriptors were used to predict electronic properties related to the mobility of the π electrons, such as resonance energy, polarizability, binding affinities, and so on.

➤ **indices of dispersion** → statistical indices

■ **indices of neighborhood symmetry** ($\equiv$ *multigraph information content indices*)

These are → *topological information indices* of a graph based on neighbor degrees and edge multiplicity. The index based on first neighbor degrees and edge multiplicity was originally proposed by Sarkar [Sarkar, Roy *et al.*, 1978] and called “*information content for a multigraph*.” They are calculated by partitioning graph vertices into → *equivalence classes* defined as the following: two vertices v_i and v_j of a → *multigraph* MG , belonging to the same chemical element and having the same → *vertex degree*, are said to be topologically equivalent with respect to m th order neighborhood if, and only if, to each m th order path ${}^m p_i$ starting from the vertex v_i , there corresponds a distinct m th order path ${}^m p_j$ starting from the vertex v_j characterized by the same → *conventional bond order* of the edges in the path and the same chemical element and vertex degree of the involved vertices [Magnuson, Harriss *et al.*, 1983; Roy, Basak *et al.*, 1984].

In other words, a basic requirement for the topological equivalence of two vertices is that the corresponding neighborhoods of the m th order are the same, where the m th order **neighborhood of a vertex** v_i in the graph MG is the subset $V_{ir}(MG)$ of vertices defined as

$$V_{im}(MG) = \{a | a \in V(MG); d_{ia} < m\}$$

where d_{ia} is the → *topological distance* of the vertex a from the focused vertex v_i and m is any nonnegative real number. The vertex neighborhood can be thought of as an open sphere $S(v_i, m)$ constituted by all the vertices a in the graph, such that their distance from the vertex v_i is less than m . Obviously, $S(v_i, 0) = \emptyset$, $S(v_i, 1) = v_i$ for $0 < m < 1$, and if $1 < m < 2$, then $S(v_i, m)$ is the set consisting of v_i together with all its adjacent vertices.

In practice, an unordered sequence called a *coordinate* is assigned to each graph vertex v_i :

$$\{(a_1, \delta_1, \pi_1^*; \dots; a_m, \delta_m, \pi_m^*)_1; (a_1, \delta_1, \pi_1^*; \dots; a_m, \delta_m, \pi_m^*)_2; \dots; (a_1, \delta_1, \pi_1^*; \dots; a_m, \delta_m, \pi_m^*)_{P_i}^m\}$$

where the coordinate is composed of ordered sequences each representing a distinct path of length m starting from the vertex v_i , ${}^m P_i$ is the total number of m th order paths starting from v_i (i.e., the m th order → *atomic path count*), a and δ are the chemical element and the vertex degree of the vertices involved in the considered path, and π^* is the conventional bond order of the edge connecting the considered neighbor of v_i and the previous one. The ordered subscripts 1, 2, ..., m of the chemical element, vertex degree and bond order in each path refer to 1st, 2nd, ..., m th neighbors of the vertex v_i along with the considered path. Therefore, two vertices are topologically equivalent if their coordinates are the same.

From the obtained equivalence classes in the hydrogen-filled multigraph, for each m th order (usually $m = 0-6$), the m th order **neighborhood Information Content** IC_m is calculated as defined by → *Shannon's entropy*:

$$IC_m = - \sum_{g=1}^G \frac{A_g}{A} \cdot \log_2 \frac{A_g}{A} = - \sum_{g=1}^G p_g \cdot \log_2 p_g$$

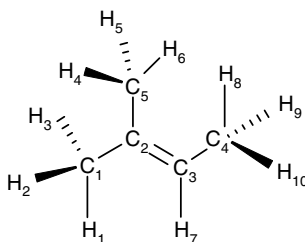
where the summation goes over the G equivalence classes, A_g is the cardinality of the g th equivalence class, A is the total number of vertices, and p_g is the probability of randomly selecting a vertex of the g th class. It represents a measure of structural complexity per vertex.

This descriptor calculated for the $\rightarrow H$ -depleted molecular graph coincides with the $\rightarrow$ vertex orbital information content in the case of atoms of the same chemical element and maximal order of neighborhood:

$$\bar{I}_{\text{ORB}} = IC_m \quad \text{for } m = \max(m)$$

Example 18

Indices of neighborhood symmetry up to the third order for 2-methyl-but-2-ene.



Order	Equivalent vertices	Probability	Descriptors
0	(C ₁ , C ₂ , C ₃ , C ₄ , C ₅)	5/15	$IC_0 = 0.9183$
	(H ₁ , H ₂ , H ₃ , H ₄ , ..., H ₁₀)	10/15	$CIC_0 = 2.9886$ $TIC_0 = 13.7744$ $SIC_0 = 0.2350$ $BIC_0 = 0.2350$
1	(C ₁ , C ₄ , C ₅)	3/15	$IC_1 = 1.3753$
	(C ₂)	1/15	$CIC_1 = 2.5316$
	(C ₃)	1/15	$TIC_1 = 20.6292$
	(H ₁ , H ₂ , H ₃ , H ₄ , ..., H ₁₀)	10/15	$SIC_1 = 0.3520$ $BIC_1 = 0.3520$
2	(C ₁ , C ₅)	2/15	$IC_2 = 1.8716$
	(C ₂)	1/15	$CIC_2 = 2.0353$
	(C ₃)	1/15	$TIC_2 = 28.0740$
	(C ₄)	1/15	$SIC_2 = 0.4790$
	(H ₁ , H ₂ , H ₃ , H ₄ , H ₅ , H ₆ , H ₈ , H ₉ , H ₁₀)	9/15	$BIC_2 = 0.4790$
	(H ₇)	1/15	
3	(C ₁ , C ₅)	2/15	$IC_3 = 2.4226$
	(C ₂)	1/15	$CIC_3 = 1.4843$
	(C ₃)	1/15	$TIC_3 = 36.3387$
	(C ₄)	1/15	$SIC_3 = 0.6201$
	(H ₁ , H ₂ , H ₃ , H ₄ , H ₅ , H ₆)	6/15	$BIC_3 = 0.6201$
	(H ₇)	1/15	
	(H ₈ , H ₉ , H ₁₀)	3/15	

From the m th order neighborhood information content, the following information indices were also derived:

- **neighborhood Total Information Content (TIC_m)**

The m th order TIC_m is defined as A times IC_m :

$$TIC_m = A \cdot IC_m$$

where A is the number of graph vertices. This descriptor represents a measure of the graph complexity.

- **Structural Information Content (SIC_m)**

The m th order SIC_m is defined in a normalized form of the information content to delete the influence of graph size:

$$SIC_m = \frac{IC_m}{\log_2 A}$$

where A is the number of graph vertices.

- **Bonding Information Content (BIC_m)**

The m th order BIC_m is defined in a normalized form as the SIC_m index, but taking into account the number of edges and their multiplicity,

$$BIC_m = \frac{IC_m}{\log_2 \left(\sum_{b=1}^B \pi_b^* \right)}$$

where B is the number of edges and π_b^* is the conventional bond order of the edge b . In the original definition, the denominator was simply considered to be the edge number B .

- **Complementary Information Content (CIC_m)**

The m th order CIC_m measures the deviation of IC_m from its maximum value, that corresponds to the vertex partition into equivalence classes containing one element each:

$$CIC_m = \log_2 A - IC_m$$

where A is the number of graph vertices.

- **redundant information content (R_m)**

This is a measure of relative redundancy of a graph obtained by normalizing the complementary information content, defined as [Roy, Basak *et al.*, 1984]

$$R_m = \frac{CIC_m}{\log_2 A} = 1 - SIC_m$$

- **order of neighborhood (O)**

The order of neighborhood O is a molecular descriptor defined as the order m of the IC_m index when it reaches the maximum value:

$$O = m \quad \text{where } m : \max_m(IC_m)$$

To account for steric effects in molecule–receptor interactions, the **weighted information indices by volume** have been proposed [Ray, Gupta *et al.*, 1985]. These molecular descriptors are calculated in the same way as the indices of neighborhood symmetry above defined using the atomic van der Waals volumes to get the probabilities of the equivalence classes. In other words, the van der Waals volumes of the atoms belonging to each equivalent class are summed up to give a molecule subvolume then divided by the total molecule volume. For example, the weighted information content by volume is defined as

$$IC_m^V = - \sum_{g=1}^G \frac{V_g}{V^{vdw}} \cdot \log_2 \frac{V_g}{V^{vdw}}$$

where V^{vdw} is the $\rightarrow$ *van der Waals volume* of the molecule and V_g is the sum of the effective van der Waals volumes of the atoms in the g th class. The effective van der Waals volume of an atom is defined as the van der Waals volume of the atom minus half the sphere overlapping of the atom due to covalent bonding of the adjacent atoms in the molecule.

Two other information indices derived from the indices of neighborhood symmetry were proposed modifying the classical definition of Shannon's entropy [King, 1989].

In particular, the **Modified Information Content index** (or **MIC index**) was proposed using the atomic masses as the $\rightarrow$ *weighting scheme*:

$$MIC = - \sum_{g=1}^G m_g \cdot (p_g \log_2 p_g)$$

where m_g is the atomic mass of all the equivalent atoms in the g th class and p_g is the probability of selecting a vertex of class g .

The **Z-Modified Information Content index** (or **ZMIC index**) was analogously defined as

$$ZMIC = - \sum_{g=1}^G n_g \cdot Z_g \cdot (p_g \log_2 p_g)$$

where Z_g is the atomic number and n_g the number of atoms in the g th class.

📖 [Bonchev, Kamenski *et al.*, 1976; Ray, Basak *et al.*, 1981, 1982, 1983; Basak and Magnuson, 1983; Roy, Raychaudhury *et al.*, 1983; Basak, Gieschen *et al.*, 1984; Basak, Harriss *et al.*, 1984; Basak, Monsrud *et al.*, 1986; Basak, Magnuson *et al.*, 1987, 1988; Basak, 1987, 1990, 1999; Basak, Niemi *et al.*, 1990b, 1991; Niemi, Basak *et al.*, 1992; Basak, Bertelsen *et al.*, 1994, 1995; Boecklen and Niemi, 1994; Basak, Gute *et al.*, 1996b, 1999a; POLLY – Basak, Harriss *et al.*, 1988; Basak and Gute, 1997; Basak, Balaban *et al.*, 2000; Luan, Zhang *et al.*, 2005a]

- **induced dipole moment** $\rightarrow$ electric polarization descriptors
- **induced polarization** $\rightarrow$ electric polarization descriptors
- **induction parameter** $\rightarrow$ multiple bond descriptors
- **inductive effect** $\rightarrow$ electronic substituent constants
- **inductive electronic constants** $\rightarrow$ electronic substituent constants
- **inertia matrix** $\rightarrow$ principal moments of inertia

- **inertia principal moments** $\equiv$ *principal moments of inertia*
- **inertial shape factor** $\rightarrow$ shape descriptors
- **influence/distance matrix** $\rightarrow$ GETAWAY descriptors
- **informational energy content** $\rightarrow$ information content

■ **information bond index** (I_B)

This is proposed by Dosmorov [Dosmorov, 1982] by analogy with the $\rightarrow$ *total information index on atomic composition* of a molecule as

$$I_B = B \cdot \log_2 B - \sum_{g=1}^G B_g \cdot \log_2 B_g$$

where B is the number of bonds in a molecule and B_g the number of bonds of type g ; the summation goes over all G different types of bonds in the molecule. A simple partition of molecule bonds is according to the $\rightarrow$ *conventional bond order*, that is, single, double, triple, and aromatic bonds [Bonchev, 1983].

- **information centrality** $\rightarrow$ center of a graph

■ **information connectivity indices**

These are $\rightarrow$ *topological information indices* based on the partition of the edges in a graph according to the equivalence and the magnitude of their $\rightarrow$ *edge connectivity* values [Bonchev, Mekenyan *et al.*, 1981c].

The **mean information content on the edge equality** ${}^E\bar{I}_\chi^E$ is based on the partition of edges according to edge connectivity equivalence and is defined as

$${}^E\bar{I}_\chi^E = - \sum_{g=1}^G \frac{B_g}{B} \cdot \log_2 \frac{B_g}{B}$$

where B_g is the number of edges having the same edge connectivity, G is the number of different connectivity values and B the number of edges.

The **mean information content on the edge magnitude** ${}^E\bar{I}_\chi^M$ is based on the magnitude of edge connectivities and is defined as

$${}^E\bar{I}_\chi^M = - \sum_{b=1}^B \frac{(\delta_i \cdot \delta_j)_b^{-1/2}}{{}^1\chi} \cdot \log_2 \frac{(\delta_i \cdot \delta_j)_b^{-1/2}}{{}^1\chi}$$

where ${}^1\chi$ is the $\rightarrow$ *Randić connectivity index*, $(\delta_i \cdot \delta_j)_b^{-1/2}$ is the edge connectivity, δ_i and δ_j being the vertex degree of the two vertices connected by the edge b , and B the total number of edges.

■ **information content**

The information content of a system having n elements is a measure of the degree of diversity of the elements in the set $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ [Klir and Folger, 1988]; it is defined as

$$I_C = \sum_{g=1}^G n_g \log_2 n_g$$

where G is the number of different $\rightarrow$ equivalence classes and n_g is the number of elements in the g th class and

$$n = \sum_{g=1}^G n_g$$

Each g th equivalence class is built by the definition of some relationships among the elements of the system. The logarithm is taken at base 2 for measuring the information content in bits.

The information content is zero, that is, no equivalence relationships are known, if all the elements are different from each other, that is, there are $G = n$ different equivalence classes. On the contrary, the information content is maximal if all the elements of the set are recognized as belonging to the same class ($G = 1$). This quantity is called **maximal information content** $^{\max}I_C$ and represents the information content needed to characterize all the n alternatives, that is, the elements of the considered set:

$$^{\max}I_C = n \log_2 n$$

The **total information content** (or **negentropy**) of a system having n elements is defined by the following:

$$I = ^{\max}I_C - I_C = n \log_2 n - \sum_{g=1}^G n_g \log_2 n_g = n \cdot H$$

When the total information content is calculated on molecules, n being the total number of atoms and n_g the number of equivalent atoms of type g , it is often referred to as **molecular negentropy**. The term H is *Shannon's entropy* that is defined below.

The total information content represents the residual information contained in the system after G relationships are defined among the n elements.

The **mean information content** $\bar{I}$, also called **Shannon's entropy** H [Shannon and Weaver, 1949], is the most common measure of uncertainty and is defined as

$$\bar{I} \equiv H = \frac{I}{n} = - \sum_{g=1}^G \frac{n_g}{n} \log_2 \frac{n_g}{n} = - \sum_{g=1}^G p_g \log_2 p_g$$

where p_g is the probability of randomly selecting an element of the g th class, and I is the total information content (Table I2). Moreover, the following condition must hold:

$$\sum_{g=1}^G p_g = 1$$

The maximum value of the entropy is $\log_2 n$, obtained when $n_g = 1$ for all G equivalence classes.

This quantity, called **Hartley information**, is defined as

$$I_n = \log_2 n$$

where n can be interpreted as the number of alternatives regardless of whether they are realized by one selection from a set or by a sequence of selections [Hartley, 1928]. The Hartley information is based on the uncertainty associated with the choice among a certain number

n of alternatives and is a simple measure of nonspecificity: it represents the information content needed to characterize one of the n alternatives.

The **standardized Shannon's entropy** (or **standardized information content**) is the ratio of the actual mean information content over the maximum available information content (i.e., the Hartley information):

$$H^* = \frac{H}{I_n} = \frac{H}{\log_2 n} = \frac{I}{n \cdot \log_2 n} = \frac{I}{\max I_C} \quad 0 \leq H^* \leq 1$$

The standardized Shannon's entropy is a measure of the relative efficiency of the collected information, that is, the mean information per unit.

Moreover, if the n elements are partitioned into k bins (with $k < n$), the maximum information content is obtained when the elements are uniformly distributed into the k bins, that is, the standardized Shannon's entropy is calculated as

$$H^* = \frac{H}{\log_2 k} \quad 0 \leq H^* \leq 1$$

The **joint entropy** $JH(X_1, X_2, \dots, X_K)$ of the sets $X_1, X_2, \dots, X_K$ is defined as

$$JH(X_1, X_2, \dots, X_K) = - \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_K \in X_K} p(x_1, x_2, \dots, x_K) \cdot \log_2 p(x_1, x_2, \dots, x_K)$$

where $p(x_1, x_2, \dots, x_K)$ is the joint probability for the set of values $x_1, x_2, \dots, x_K$.

From the mean information content, Brillouin [Brillouin, 1962] defined a complementary quantity, called **Brillouin redundancy index** (or **redundancy index**), denoted as R , to measure the information redundancy of the system:

$$R = 1 - \frac{H}{\log_2 n} = 1 - H^*$$

where H^* is the standardized Shannon's entropy.

Both total and mean information content are widely used as molecular descriptors and are called $\rightarrow$ *information indices*.

Table I2 Elemental contributions to the information content functions (n , number of elements; p , probability).

n	$\log_2 n$	$n \log_2 n$	p	$-\log_2 p$	$-p \log_2 p$
0	$-\infty$	0.000	0.0	∞	0.000
1	0.000	0.000	0.1	3.322	0.332
2	1.000	2.000	0.2	2.322	0.464
3	1.585	4.755	0.3	1.737	0.521
4	2.000	8.000	0.4	1.322	0.529
5	2.322	11.610	0.5	1.000	0.500
6	2.585	15.510	0.6	0.737	0.442
7	2.807	19.651	0.7	0.515	0.360
8	3.000	24.000	0.8	0.322	0.258
9	3.170	28.529	0.9	0.152	0.137
10	3.322	33.219	1.0	0.000	0.000

Another measure of entropy is given by the **Gini index** G defined as

$$G = \sum_{g \neq k} p_g \cdot p_k \quad 0 \leq G \leq \frac{n-1}{2n}$$

where g and k are two different equivalence classes. The Gini index increases as the diversity of the system elements increases. A complementary quantity to the Gini index is the **informational energy content** defined as [Onicescu, 1966]

$$I_E = \sum_g p_g^2 \quad \frac{1}{n} \leq I_E \leq 1$$

It corresponds to a redundancy measure whose maximum and minimum values are 1 and $1/n$, respectively.

Both Shannon's entropy and Gini index are used as $\rightarrow$ *classification parameters* to evaluate the quality of classification results.

Example 19

Some indices of information content. Let a set of seven elements ($n = 7$) be $\{3, 3, 4, 4, 4, 5, 8\}$. Four equivalence classes have been defined as shown below, together with the corresponding probability and the related quantities.

Equivalence class	n_g	Probability	$n_g \log_2 n_g$	$-\log_2 p_g$	$-p_g \log_2 p_g$
{3}	2	$p_1(3) = 2/7 = 0.286$	2.000	1.806	0.516
{4}	3	$p_2(4) = 3/7 = 0.429$	4.755	1.221	0.524
{5}	1	$p_3(5) = 1/7 = 0.143$	0.000	2.806	0.401
{8}	1	$p_4(8) = 1/7 = 0.143$	0.000	2.806	0.401

$$\max I_C = 7 \times \log_2 7 = 19.651 \quad I_C = \sum_{g=1}^4 n_g \log_2 n_g = 6.755$$

$$I = \max I_C - I_C = 7 \cdot H = 11.896$$

$$H = \frac{I}{7} = -\sum_{g=1}^4 p_g \log_2 p_g = 1.842 \quad H^* = \frac{H}{\log_2 7} = 0.656 \quad R = 1 - H^* = 0.344$$

$$G = \sum_{g \neq k} p_g \cdot p_k = 0.348 \quad I_E = \sum_g p_g^2 = 0.307$$

A nonmetric measure of relative entropy is the **Kullback–Leibler divergence** (also called **information divergence**, **information gain**, or **relative entropy**) that is a measure of the difference between two probability distributions P and Q , defined as

$$D_{KL}(P||Q) = \sum_{g=1}^G p_g \cdot \log \frac{p_g}{q_g}$$

Often P is an experimental probability distribution and Q is the theoretical one.

In general terms, the expected information gain (I_G) is the change in entropy from a prior state Q to a state P that takes some information: $I_G(P||Q) = H(Q) - H(P||Q)$.

Several descriptors are based on the concepts of information content and entropy; among these are the $\rightarrow$ *topological information indices*, $\rightarrow$ *indices of neighborhood symmetry*, $\rightarrow$ *Shannon Entropy Descriptors* (SHED), $\rightarrow$ *graph entropy*, and some descriptors among the $\rightarrow$ *molecular complexity indices* and the $\rightarrow$ *GETAWAY descriptors*.

The Shannon's entropy is also largely applied to the analysis of the information content of molecular descriptors within data sets of molecules and to assess the diversity of chemical libraries [Lin, 1996b; Bajorath, 2001].

Entropy measures have been used to evaluate the variability of descriptors for $\rightarrow$ *variable selection* purposes (high degenerate descriptors have low entropy values) using histograms comprised of a number of bins (e.g., 100) to represent each variable range [Godden, Stahura *et al.*, 2000; Godden and Bajorath, 2000].

To identify molecular descriptors that are sensitive to intrinsic differences in two collections of compounds A and B , a **differential Shannon's entropy** DSE has also been proposed as [Godden and Bajorath, 2001, 2002, 2003; Stahura, Godden *et al.*, 2002]

$$DSE_j \equiv dH_j = H_{AB,j} - \frac{H_{A,j} + H_{B,j}}{2}$$

where $H_{A,j}$ and $H_{B,j}$ are the Shannon's entropies of the j th molecular descriptor in the compound collection A and B , respectively; $H_{AB,j}$ is the entropy of the descriptor when its values in both collections are considered together. To calculate the entropy value, the descriptor value range is divided into a predefined number of bins (e.g., 100). Low values of DSE indicate that little or no difference between the two compound databases is detected with respect to the considered descriptor. The standardized DSE, denoted as sDSE, is calculated as

$$sDSE_j = \frac{DSE_j}{\log_2(N_{bins})}$$

where N_{bins} is the total number of bins.

📖 [Shannon, 1948a, 1948b; Dancoff and Quastler, 1953; Rashevsky, 1960; Bonchev and Kamenska, 1978; Kier, 1980b; Agrafiotis, 1997; Rouvray, 1997; Graham and Schacht, 2000; Stahura, Godden *et al.*, 2000; Oprea, 2002b; Graham, 2002; Xue, Godden *et al.*, 2003b; Tarasov, Mustafaev *et al.*, 2005; David and Mihailciuc, 2006; Landon and Schaus, 2006; Baldi, Benz *et al.*, 2007; Batista and Bajorath, 2007]

- **information content based on center** $\rightarrow$ centric indices
- **information content ratio** $\rightarrow$ model complexity
- **information distance index** $\rightarrow$ topological information indices ($\odot$ vertex distance complexity)
- **information divergence** $\equiv$ *Kullback–Leibler divergence* $\rightarrow$ information content
- **information energy content** $\rightarrow$ information content
- **information functional** $\rightarrow$ graph entropy

- **information gain** $\equiv$ *Kullback–Leibler divergence* $\rightarrow$ information content
- **information gain in classification** $\rightarrow$ classification parameters
- **information index on amino acid composition** $\rightarrow$ biodescriptors ($\odot$ peptide sequences)
- **information index on isotopic composition** $\rightarrow$ atomic information indices

■ **information index on molecular conformations** (I_{CONF})

This is a molecular index defined as $\rightarrow$ *total information content* based on the number of conformations of a molecule (usually below a cutoff energy value) [Bonchev, 1983]:

$$I_{\text{CONF}} = N_{\text{CONF}} \cdot \log_2 N_{\text{CONF}}$$

where N_{CONF} is the number of molecular conformations. The corresponding **mean information index on molecular conformations** is defined as

$$\bar{I}_{\text{CONF}} = \log_2 N_{\text{CONF}}$$

For rigid molecules that have a unique conformation $I_{\text{CONF}} = 0$ and $\bar{I}_{\text{CONF}} = 0$; for molecules with two conformations (e.g., chair/boat) $I_{\text{CONF}} = 2$ and $\bar{I}_{\text{CONF}} = 1$ [Dosmorov, 1982].

- **information index on molecular symmetry** $\rightarrow$ symmetry descriptors
- **information indices on polynomial coefficients** $\rightarrow$ characteristic polynomial-based descriptors
- **information index on proton–neutron composition** $\rightarrow$ atomic information indices
- **information index on size** $\rightarrow$ atom number

■ **information indices**

These are molecular descriptors calculated by applying formulas as $\rightarrow$ *information content* to molecules. Different criteria are used for defining $\rightarrow$ *equivalence classes*, that is, equivalency of atoms in a molecule such as chemical identity, ways of bonding through space, molecular topology and symmetry, $\rightarrow$ *local vertex invariants* [Bonchev, Kamenska *et al.*, 1976; Bonchev, 1983, 2005].

Among these molecular descriptors, the most popular are $\rightarrow$ *topological information indices*.

Other information indices are $\rightarrow$ *atomic composition indices*, $\rightarrow$ *information bond index*, $\rightarrow$ *Morowitz information index*, $\rightarrow$ *information index on size*, $\rightarrow$ *information index on molecular symmetry*, $\rightarrow$ *information index on amino acid composition*, $\rightarrow$ *information index on molecular conformations*, $\rightarrow$ *Bertz complexity index*, $\rightarrow$ *Dosmorov complexity index*, $\rightarrow$ *Bonchev complexity index*, $\rightarrow$ *atomic information indices*, $\rightarrow$ *electropy index*, $\rightarrow$ *information theoretic topological index*, $\rightarrow$ *information layer index*. Information indices are also the $\rightarrow$ *Shannon Entropy Descriptors* (SHED), while some information indices are among the $\rightarrow$ *GETAWAY descriptors*.

- **information indices on the adjacency matrix** $\rightarrow$ topological information indices
- **information indices on the distance matrix** $\rightarrow$ topological information indices
- **information indices on the edge adjacency matrix** $\rightarrow$ topological information indices
- **information indices on the edge-cycle incidence matrix** $\rightarrow$ topological information indices
- **information indices on the edge distance matrix** $\rightarrow$ topological information indices
- **information indices on the vertex-cycle incidence matrix** $\rightarrow$ topological information indices

- **information layer index** → topological information indices (⊙ vertex distance complexity)
- **information on the possible valence bonds** → Morovitz information index

■ **information theoretic topological index** (I_k)

The information theoretic topological index I_k is derived from a → *signed graph* as [Sahu and Lee, 2004]

$$I_k = m \cdot \log_2 m - n \cdot \log_2 n - (n-m) \cdot \log_2 |n-m|$$

where m and n represent the total number of positive signs and the total number of negative signs of the edges, respectively, and k is the molecular orbital level.

From this definition, other indices are derived: the **bonding orbital information index** I_b

$$I_b = \sum_{k=1}^{K^+} I_k$$

where the summation runs over the I_k terms having positive signs, and the **antibonding orbital information index** I_a :

$$I_a = \sum_{k=1}^{K^-} I_k$$

where the summation runs over the I_k terms having negative signs.

- **information Wiener index** ≡ *mean information content on the distance magnitude* → topological information indices
- **inner molecular polarizability index** → electric polarization descriptors (⊙ Polarizability Effect Index)
- **integrated molecular transform** → molecular transforms
- **integrated spatial difference in field potential** → molecular shape analysis
- **integrity moments** → GRID-based QSAR techniques (⊙ VolSurf descriptors)
- **interatomic distances** → molecular geometry
- **interatomic interaction spectrum** → molecular transform
- **interaction energy values** → grid-based QSAR techniques
- **interaction fields** ≡ *molecular interaction fields*
- **interaction geodesic matrices** → weighted matrices (⊙ weighted distance matrices)
- **interaction graph matrices** → weighted matrices (⊙ weighted distance matrices)
- **interaction index** → quantum-chemical descriptors
- **Interaction Pharmacophore Elements** → 4D-Molecular Similarity Analysis
- **interactive polar parameter** → lipophilicity descriptors
- **interactive variable selection for PLS** → variable selection
- **intermediate least squares regression** → variable selection
- **intermolecular interatomic distances** → molecular geometry
- **intermolecular solute-membrane interaction descriptors** → Membrane Interaction QSAR analysis
- **internal coordinates** → molecular geometry
- **internal fragment topological indices** → fragment topological indices

- **interquartile range** → statistical indices (⊙ indices of dispersion)
- **interstitial volume** → molecular surface (⊙ solvent-accessible molecular surface)
- **intramolecular interatomic distances** → molecular geometry
- **intramolecular solute descriptors** → Membrane Interaction QSAR analysis
- **intricacy numbers** → Schultz molecular topological index
- **intrinsic molecular volume** $\equiv$ *van der Waals volume* → volume descriptors
- **intrinsic state** → electrotopological state indices
- **intrinsic state** → vertex degree
- **intrinsic state pseudoconnectivity indices** → electrotopological state indices
- **intrinsic state sum** → electrotopological state indices
- **invariance properties of molecular descriptors** → molecular descriptors

■ invariant moments

Invariant moments are descriptors derived from a statistical analysis of the distribution and value of the pixels that make up the image to be compared [Hu, 1962; Robinson, Barlow *et al.*, 1997b]. This distribution takes the form of the following equations:

$$m_{p,q} = \iint x^p \cdot y^q \cdot \mathcal{P}(x, y) \cdot dx dy \quad \text{for continuous data}$$

or

$$m_{p,q} = \sum \sum x^p \cdot y^q \cdot \mathcal{P}(x, y) \quad \text{for discrete data}$$

where $\mathcal{P}(x, y)$ is any property projected onto a bidimensional map, x and y being the map coordinates; p and q are integer parameters. The invariance to translation is simply obtained by centering the coordinates, then the following central moments derive as

$$\mu_{p,q} = \iint (x-\bar{x})^p (y-\bar{y})^q \cdot \mathcal{P}(x, y) \cdot dx dy \quad \text{for continuous data}$$

or

$$\mu_{p,q} = \sum \sum (x-\bar{x})^p \cdot (y-\bar{y})^q \cdot \mathcal{P}(x, y) \quad \text{for discrete data}$$

The invariance to rotation is obtained by a complex procedure that combines the different central moments, leading to seven invariant moments used as image descriptors:

$$\begin{aligned} \eta_0 &= \mu_{20} + \mu_{02} & \eta_1 &= \sqrt{|(\mu_{20} - \mu_{02})^2 + 4 \cdot \mu_{11}^2|} \\ \eta_2 &= \sqrt{|(\mu_{30} - 3 \cdot \mu_{12})^2 + (3 \cdot \mu_{21} - \mu_{03})^2|} & \eta_3 &= \sqrt{|(\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2|} \\ \eta_4 &= \sqrt[4]{\left| \frac{(\mu_{30} - 3 \cdot \mu_{12}) \cdot (\mu_{30} + \mu_{12}) \cdot [(\mu_{30} + \mu_{12})^2 - 3 \cdot (\mu_{21} + \mu_{03})^2] + (3 \cdot \mu_{21} - \mu_{03}) \cdot (\mu_{21} + \mu_{03}) \cdot [3 \cdot (\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]}{(\mu_{30} - 3 \cdot \mu_{12}) \cdot (\mu_{30} + \mu_{12}) \cdot [(\mu_{30} + \mu_{12})^2 - 3 \cdot (\mu_{21} + \mu_{03})^2] + (3 \cdot \mu_{21} - \mu_{03}) \cdot (\mu_{21} + \mu_{03}) \cdot [3 \cdot (\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]} \right|} \\ \eta_5 &= \sqrt[3]{(\mu_{20} - \mu_{02}) \cdot [(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4 \cdot \mu_{11} \cdot (\mu_{30} + \mu_{12}) \cdot (\mu_{21} + \mu_{03})} \end{aligned}$$

$$\eta_6 = \sqrt[4]{\left| \frac{(3 \cdot \mu_{21} - \mu_{03}) \cdot (\mu_{30} + \mu_{12}) \cdot [(\mu_{30} + \mu_{12})^2 - 3 \cdot (\mu_{21} + \mu_{03})^2] + (3 \cdot \mu_{12} - \mu_{30}) \cdot (\mu_{21} + \mu_{03}) \cdot [3 \cdot (\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]}{(3 \cdot \mu_{12} - \mu_{30}) \cdot (\mu_{21} + \mu_{03}) \cdot [3 \cdot (\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]} \right|}$$

The final vector comprising of the seven invariant moments represents the distribution of the property $\mathcal{P}$ on the bidimensional map. This representation was used to compare bidimensional molecule projections by $\rightarrow$ *Carbó similarity index* and $\rightarrow$ *Hodgkin similarity index*.

- **inverse QSAR** $\equiv$ *reversible decoding* $\rightarrow$ structure/response correlations
- **ionization energy** $\equiv$ *ionization potential* $\rightarrow$ quantum-chemical descriptors
- **ionization potential** $\rightarrow$ quantum-chemical descriptors
- **IPE** $\equiv$ *Interaction Pharmacophore Elements* $\rightarrow$ 4D-Molecular Similarity Analysis
- **irregular graph** $\rightarrow$ graph
- **ISIDA descriptors** $\rightarrow$ substructure descriptors ($\odot$ fingerprints)
- **ISIS keys** $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **isocodal graphs** $\rightarrow$ graph
- **isolated vertices** $\rightarrow$ graph
- **isomorphic graphs** $\rightarrow$ graph
- **isospectral graphs** $\rightarrow$ graph
- **isotropic surface area** $\rightarrow$ molecular surface ($\odot$ solvent-accessible molecular surface)

■ Iterated Line Graph Sequence (ILGS)

This is an ordered sequence of line graphs $L(\mathcal{G})$ obtained by an iterative procedure starting from the $\rightarrow$ *molecular graph* $\mathcal{G}$ [Diudea, Horvath *et al.*, 1992]:

$$\{L_0, L_1, L_2, \dots, L_m\}$$

where L_0 is the $\rightarrow$ *line graph* of zero-order coinciding with the original molecular graph $\mathcal{G}$; L_1 is the line graph of $\mathcal{G}$; L_2 is the line graph of the first line graph L_1 ; L_m is the m th line graph of $\mathcal{G}$, that is, $L_m(\mathcal{G}) = L(L_{m-1}(\mathcal{G}))$.

The numbers of vertices A_m and edges B_m in the graph L_m are given by the following relations:

$$A_m = B_{m-1}$$

$$B_m = \sum_{i=1}^{A_{m-1}} \binom{\delta_i}{2} = \frac{1}{2} \cdot \sum_{i=1}^{A_{m-1}} \delta_i^2 - B_{m-1}$$

where A_{m-1} and B_{m-1} represent the number of vertices and edges in the line graph of $(m-1)$ th order, respectively; δ_i is the $\rightarrow$ *vertex degree*; it can be noted that the number of edges in the m th line graph L_m coincides with the $\rightarrow$ *connection number* of the $(m-1)$ th line graph L_{m-1} (Figure I1).

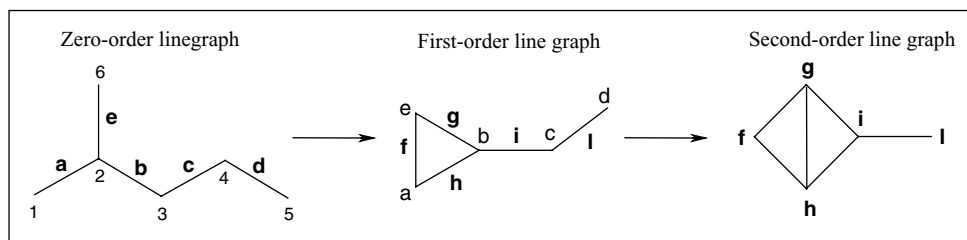


Figure 11 First- and second-order line graphs of the 2-methylpentane.

Each vertex i_m of the current line graph L_m denotes a pair of vertices of the lower order graph L_{m-1} :

$$i_m = (j_{m-1}, k_{m-1})$$

where the two vertices j and k in L_{m-1} are necessarily connected by an edge of the graph and are themselves pairs of vertices of graph L_{m-2} . This relation between the i th edge and the corresponding vertices j and k can be represented as

$$j_{m-1} \in i_m \quad \text{and} \quad k_{m-1} \in i_m$$

The relatedness of vertices in the line derivative process can be represented by the Kronecker delta as

$$\delta(i_m, i_{m+1}) = \begin{cases} 1 & \text{if } (i_m \in i_{m+1}) \\ 0 & \text{otherwise} \end{cases}$$

This relation can be extended to any arbitrary rank of derivative m and n ($m \geq n$), stating that $\delta(i_m, i_n) = 1$ only if the vertex i_m appears in at least one of the subsets defining the vertex i_n .

Based on iterated line graph sequence, local and global centricities of molecular graphs were derived on the basis of the $\rightarrow$ *branching layer matrix LB* by the MOLCEN algorithm [Diudea, Horvath *et al.*, 1992]. This algorithm was proposed to obtain $\rightarrow$ *canonical numbering* of subgraphs of various length in molecular graphs.

MOLORD algorithm, proposed by Diudea [Diudea, Horvath *et al.*, 1995a], calculates local and global graph invariants from a series of line graphs $L_0, L_1, \dots, L_m$ derived from a molecular graph G . Let ${}^m\mathcal{L}_i$ be any local invariant of the vertex i and ${}^m\mathcal{D}$ the corresponding global invariant on each L_m within the set of derivative graphs $L_0, L_1, \dots, L_m$, that is,

$${}^m\mathcal{D} = \sum_{i=1}^{A_m} {}^m\mathcal{L}_i$$

A **partial local invariant** ${}^m\mathcal{PL}(i_n)$ of a vertex i_n with respect to the m th order line graph L_m is defined as

$${}^m\mathcal{PL}(i_n) = \frac{{}^n\mathcal{D}}{{}^m\mathcal{D}} \cdot \sum_{i=1}^{A_m} {}^m\mathcal{L}_i \cdot \delta(i_n; i)$$

where ${}^m\mathcal{L}_i$ is the local vertex invariant of the vertex i_m calculated from the topology of the graph L_m . The partial invariant of the vertex i_n of the graph L_n with respect to L_m is calculated by

summing up all local invariants ${}^m I_i$ of those vertices in L_m which are “related” to i_n according to the m - n successive derivatives, $L_n, \dots, L_m$. The ratio ${}^n \mathcal{D}/{}^m \mathcal{D}$ is used as a scaling factor allowing a comparison of $\mathcal{PL}$ values irrespective of the current L_m for which are evaluated.

For a series of successive derivative graphs $L_n, \dots, L_m$, a **local synthetic invariant** ${}^m S\mathcal{L}(i_n)$ of the vertex i_n in the n th order line graph is calculated as

$${}^m S\mathcal{L}(i_n) = \sum_{k=n}^m {}^k \mathcal{PL}(i_n) \cdot f^{n-k}$$

where the superscript m in ${}^m S\mathcal{L}(i_n)$ means that the last line graph L_m has been taken into account; f is an empirical factor used to give a different weighting to the contributions arising from derivatives of various ranks. It must be observed that in the case of $n = m$, the synthetic local invariant ${}^m S\mathcal{L}(i_n)$ reduces to the classical invariant ${}^n \mathcal{L}_i$.

Finally, a **global synthetic invariant** ${}^m S\mathcal{D}(L_n)$ of a graph L_n is defined as

$${}^m S\mathcal{D}(L_n) = \sum_{i_n=1}^{A_n} {}^m S\mathcal{L}(i_n)$$

Studies of these descriptors were performed using, as local invariants, the $\rightarrow$ *centric operator* c_i and the $\rightarrow$ *centrocomplexity operator* x_i calculated on $\rightarrow$ *layer matrices*.

A vector of molecular topological descriptors can be calculated for the whole iterated line graph sequence. For example, a **line graph Randić connectivity index** was calculated as [Estrada, Guevara *et al.*, 1998b]

$$\chi(L_n) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot [\delta_i(L_n) \cdot \delta_j(L_n)]^{-1/2}$$

where the summation runs over all the pairs of vertices in the n th order line graph and δ_i and δ_j are the $\rightarrow$ *vertex degree* of the two vertices v_i and v_j ; a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to one for pairs of adjacent vertices, and zero otherwise.

Note that the χ index of the first line graph L_1 coincides with the $\rightarrow$ *edge connectivity index* ϵ . Moreover, the **line graph connectivity indices** were proposed by analogy with the $\rightarrow$ *Kier–Hall connectivity indices* as

$${}^m \chi_t(L_n) = \sum_{k=1}^K \left(\prod_i \delta_i(L_n) \right)_k^{-1/2}$$

where k runs over all the m th order subgraphs, m being the number of edges in the subgraph; K is the total number of m th order subgraphs; the product is over all the vertex degrees of the vertices involved in the subgraph. The subscript “ t ” for the connectivity indices refers to the type of $\rightarrow$ *molecular subgraph* and is “ ch ” for chain or ring, “ pc ” for path-cluster, “ c ” for cluster, and “ p ” for path (that can also be omitted) [Estrada, Guevara *et al.*, 1998a]. It was shown that line graph connectivity indices are linear combinations of $\rightarrow$ *extended edge connectivity indices* for some molecular graphs.

Spectral moments of iterated line graph sequence were derived from the adjacency matrix A of the line graph L_n of any order n as

$$\mu^k(L_n) = \text{tr}[A^k(L_n)]$$

where μ^k is the k th order spectral moment and “tr” is the $\rightarrow$ trace of the k th power of the adjacency matrix A^k of the considered line graph L_n [Estrada, 1999c]. They were also expressed as linear combinations of some $\rightarrow$ embedding frequencies of the molecular graph, that is, the number of occurrences of specified subgraphs in the original molecular graph G .

Obviously, spectral moments μ^k of the first-order line graph L_1 are the $\rightarrow$ spectral moments of the edge adjacency matrix, and zero-order spectral moments μ^0 of any line graph L_n coincide with the number of vertices in the considered line graph.

📖 [Gutman and Estrada, 1996; Gutman, Popovic *et al.*, 1997, 1998; Godsil and Gutman, 1999; Kuanar, Kuanar *et al.*, 1999b; Basak, Nikolić *et al.*, 2000; Gutman and Tomović, 2000a, 2000b, 2001; Nikolić, Trinajstić *et al.*, 2001a; Tomović and Gutman, 2001a; Dobrynin and Mel'nikov, 2004, 2005a, 2005b]

- **Iterative Vertex and Edge Centricity algorithm** $\rightarrow$ center of a graph
- **Ivanciuc–Balaban operator** $\rightarrow$ Balaban distance connectivity index
- **Ivanciuc weighting schemes** $\rightarrow$ weighting schemes
- **Ivanciuc weighted adjacency matrices** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **Ivanciuc weighted distance matrices** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)



- **Jaccard/Tanimoto coefficient** → similarity/diversity
- **Jaccard–Tanimoto similarity coefficient** → similarity/diversity (Table S9)
- **JEDA** $\equiv$ *Joint Entropy-based Diversity Analysis* → cell-based methods
- **Jenkins steric parameter** → steric descriptors
- **J_p statistics** → regression parameters
- **J_i index** → Balaban distance connectivity index
- **J'/J index** → bond order indices (⊙ graphical bond order)
- **JJ indices** → Wiener matrix
- **Jochum–Gasteiger canonical numbering** → canonical numbering
- **Joint Entropy-based Diversity Analysis** → cell-based methods
- **joint entropy** → information content

■ Joshi electronic descriptors

These are molecular → *electronic descriptors* assuming that the minimum energy conformation of a molecule represents the optimal picture of the electronic charge distribution in the whole molecule [Joshi, Meister *et al.*, 1993, 1994].

The Joshi electronic descriptors (JS1–JS5) are defined as

$$JS1 = \frac{E_R}{E_H} \quad JS2 = \frac{E_R - E_H}{E_H} \quad JS3 = \frac{E_R - E_{HS}}{E_H}$$

$$JS4 = \frac{E_R - \sum_j E_{R_j}}{E_H} \quad JS5 = \frac{E_R - \sum_j E_{R_j} - E_{HS}}{E_H}$$

where E is the ΔH_f conformational energy value of the global minimum energy conformer calculated by → *computational chemistry* (AM1) methods. The subscripts R, H, and HS refer to a R-substituted compound, the unsubstituted compound, and a compound for which the aromatic moiety is unsubstituted but the side chain is substituted in a similar way. E_{R_j} is the energy contribution due to the formation of the *j*th substituent group calculated by subtracting the energy value of methane from that of the corresponding substituted methane. The summation in JS4 and JS5 depends on the series of studied compounds.

- **Joshi steric descriptor** → steric descriptors
- **Julg–François index** → delocalization degree indices

- **Jurs cost function** $\rightarrow$ regression parameters
- **Jurs shape indices** $\equiv$ *shadow indices*

K

- **Kaliszan shape parameter** → shape descriptors
- **Kamlet descriptors** → Linear Solvation Energy Relationships
- **Kamlet's general equation** → Linear Solvation Energy Relationships
- **Kantola-Villar-Loew hydrophobic models** → lipophilicity descriptors
- **K correlation analysis** → variable reduction
- **K correlation index** $\equiv$ *multivariate K correlation index* → statistical indices ($\odot$ correlation measures)

■ **Kekulé number (K)** ($\equiv$ *Kekulé structure count, SC*)

This is the number of Kekulé structures in an aromatic system [Trinajstić, 1992]. It can be calculated by extensive enumeration of the structures or by using appropriate algorithms.

For benzenoid systems, the Kekulé number K is obtained from the positive eigenvalues λ_i of the → *adjacency matrix* as

$$K = \prod_{i=1}^{A/2} \lambda_i$$

where A is the number of atoms. The logarithm of the Kekulé number is related to the resonance energy of the compound and used as a → *resonance index*.

The Kekulé number of alternant hydrocarbons is equal to the sum of “even” K^+ and “odd” K^- Kekulé structures:

$$K = K^+ + K^-$$

The even or odd parity is determined in the Dewar–Longuet-Higgins scheme [Dewar and Longuet-Higgins, 1952] according to whether the number of transpositions of double bonds required to transform one Kekulé structure into another one is even or odd.

The difference between the numbers of even and odd Kekulé structures is called **algebraic structure count** ASC [Wilcox Jr, 1968, 1969] or **corrected structure count** CSC [Herndon, 1973a, 1974b]:

$$ASC = K^+ - K^- \quad K^+ \geq K^-$$

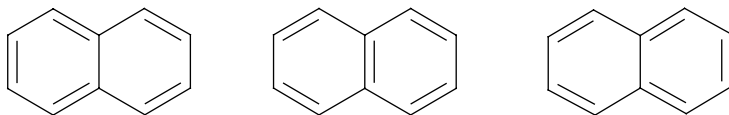


Figure K1 The three Kekulé resonance structures of the naphthalene.

ASC represents a structure count excluding structures that do not contribute to stabilizing resonance interactions. However, the concept of parity and the derived ASC descriptor do not work in nonalternant systems with three odd-membered rings [Randić and Trinajstić, 1993a].

📖 [Kekulé, 1865; Gutman and Trinajstić, 1973b; Herndon, 1973b; Balaban and Tomescu, 1985; Dias, 1992; Ivanciuc and Balaban, 1992b; Balaban, Liu *et al.*, 1993; Guo and Zhang, 1993; Guo, Randić *et al.*, 1996 Mishra and Patra, 1998; Cash, 1998]

- **Kekulé structure count** $\equiv$ *Kekulé number*
- **Kellogg and Abraham interaction field** $\rightarrow$ molecular interaction fields ($\odot$ hydrophobic fields)
- **Kendall rank correlation coefficient** $\rightarrow$ statistical indices ($\odot$ correlation measures)
- **Kier alpha-modified shape descriptors** $\rightarrow$ Kier shape descriptors
- **Kier bond rigidity index** $\rightarrow$ flexibility indices
- **Kier–Hall connectivity indices** $\equiv$ *Molecular Connectivity Indices* $\rightarrow$ connectivity indices
- **Kier–Hall electronegativity** $\rightarrow$ vertex degree
- **Kier–Hall solvent polarity index** $\rightarrow$ electric polarization descriptors
- **Kier molecular flexibility index** $\rightarrow$ flexibility indices
- **Kier steric descriptor** $\rightarrow$ steric descriptors

■ Kier shape descriptors (κ)

These are topological shape descriptors ${}^m\kappa$ defined in terms of the number of graph vertices A and the number of paths ${}^mP$ with length m ($m = 1, 2, 3$) in the $\rightarrow$ *H-depleted molecular graph*, according to the following [Kier, 1985; Kier, 1986b]:

$${}^1\kappa = 2 \cdot \frac{{}^1P_{\max} \cdot {}^1P_{\min}}{({}^1P)^2} = \frac{A \cdot (A-1)^2}{({}^1P)^2} \quad {}^2\kappa = 2 \cdot \frac{{}^2P_{\max} \cdot {}^2P_{\min}}{({}^2P)^2} = \frac{(A-1) \cdot (A-2)^2}{({}^2P)^2}$$

$${}^3\kappa = 4 \cdot \frac{{}^3P_{\max} \cdot {}^3P_{\min}}{({}^3P)^3} = \begin{cases} \frac{(A-3) \cdot (A-2)^2}{({}^3P)^2} & \text{for even } A (A > 3) \\ \frac{(A-1) \cdot (A-3)^2}{({}^3P)^2} & \text{for odd } A (A > 3) \end{cases}$$

where ${}^mP_{\min}$ and ${}^mP_{\max}$ are the minimum and maximum m th order $\rightarrow$ *path count* in the molecular graphs of molecules with the same number A of graph vertices. These extremes are obtained from two reference structures chosen in an isomeric series and, for the i th molecule, is therefore

$${}^mP_{\min} \leq {}^mP_i \leq {}^mP_{\max}$$

The reference structure for ${}^1P_{\min}$ is the $\rightarrow$ *linear graph* while for ${}^1P_{\max}$ it is the $\rightarrow$ *complete graph* in which all vertices are bonded to each other; their numerical values are calculated as follows:

$${}^1P_{\min} = A - 1 \quad {}^1P_{\max} = \frac{A \cdot (A - 1)}{2}$$

The scaling factor of 2 in the numerator of ${}^1\kappa$ index formula makes the value ${}^1\kappa = A$ when there are no cycles in the graph of the molecule. Monocyclic molecules have a lower value and bicyclic structures have an even lower value. The structural information encoded in ${}^1\kappa$ is related to the complexity, or, more precisely, to the number of cycles of a molecule.

The reference structure for ${}^2P_{\min}$ is the linear graph while for ${}^2P_{\max}$ it is the $\rightarrow$ *star graph* in which all vertices but one are adjacent to a central vertex; their numerical values are calculated as follows:

$${}^2P_{\min} = A - 2 \quad {}^2P_{\max} = \frac{(A - 1) \cdot (A - 2)}{2}$$

where A is the total number of vertices in the graph. The scaling factor of 2 in the numerator of ${}^2\kappa$ index formula makes the value ${}^2\kappa = A - 1$ for all linear graphs. The information encoded by ${}^2\kappa$ index is related to the degree of star graph-likeness and linear graph-likeness, that is, ${}^2\kappa$ encodes information about the spatial density of atoms in a molecule.

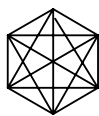
The reference structure for ${}^3P_{\min}$ is the linear graph while for ${}^3P_{\max}$ it is the *twin star graph*; their numerical values are calculated as follows:

$${}^3P_{\min} = A - 3 \quad {}^3P_{\max} = \begin{cases} \frac{(A - 2)^2}{4} & \text{for even } A \\ \frac{(A - 1) \cdot (A - 3)}{4} & \text{for odd } A \end{cases}$$

The scaling factor of 4 is used in the numerator of ${}^3\kappa$ index to bring ${}^3\kappa$ onto approximately the same numerical scale as the other kappa indices. The ${}^3\kappa$ values are larger when $\rightarrow$ *molecular branching* is nonexistent or when it is located at the extremities of a graph; ${}^3\kappa$ encodes information about the centrality of branching.

Example K1

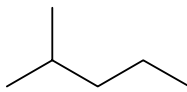
Path counts and Kier shape descriptors for 2-methylpentane (central column).



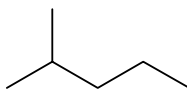
$${}^1P_{\max} = 15$$



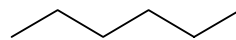
$${}^2P_{\max} = 10$$



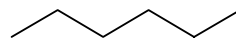
$${}^1P = 5 \quad {}^1\kappa = 6.000$$



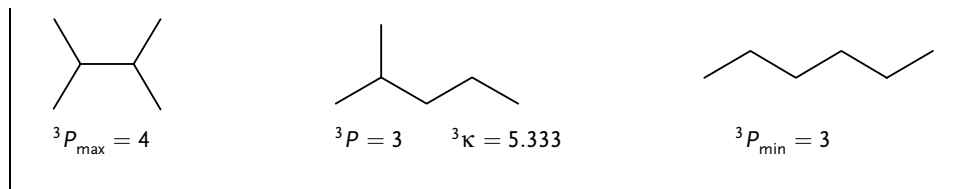
$${}^2P = 5 \quad {}^2\kappa = 3.200$$



$${}^1P_{\min} = 5$$



$${}^2P_{\min} = 4$$



To take into account the different shape contribution of heteroatoms and hybridization states, **Kier alpha-modified shape descriptors**, denoted as ${}^m\kappa_\alpha$ ($m = 1, 2, 3$), were proposed by the following [Kier, 1986a]:

$${}^1\kappa_\alpha = \frac{(A + \alpha) \cdot (A + \alpha - 1)^2}{({}^1P + \alpha)^2} \quad {}^2\kappa_\alpha = \frac{(A + \alpha - 1) \cdot (A + \alpha - 2)^2}{({}^2P + \alpha)^2}$$

$${}^3\kappa_\alpha = \begin{cases} \frac{(A + \alpha - 3) \cdot (A + \alpha - 2)^2}{({}^3P + \alpha)^2} & \text{for even } A (A > 3) \\ \frac{(A + \alpha - 1) \cdot (A + \alpha - 3)^2}{({}^3P + \alpha)^2} & \text{for odd } A (A > 3) \end{cases}$$

where α is a parameter derived from the ratio of the $\rightarrow$ covalent radius R_i of the i th atom relative to the sp^3 carbon atom ($R_{\text{C}_{\text{sp}^3}}$):

$$\alpha = \sum_{i=1}^A \left(\frac{R_i}{R_{\text{C}_{\text{sp}^3}}} - 1 \right)$$

The only nonvanishing contributions to α are given by heteroatoms or carbon atoms with a valence state different from sp^3 (Table K1).

Table K1 Covalent radius R and α parameter values for some atom types.

Atom/hybrid	R (Å)	α	Atom/hybrid	R (Å)	α
C_{sp^3}	0.77	0	P_{sp^3}	1.10	0.43
C_{sp^2}	0.67	-0.13	P_{sp^2}	1.00	0.30
C_{sp}	0.60	-0.22	S_{sp^3}	1.04	0.35
N_{sp^3}	0.74	-0.04	S_{sp^2}	0.94	0.22
N_{sp^2}	0.62	-0.20	F	0.72	-0.07
N_{sp}	0.55	-0.29	Cl	0.99	0.29
O_{sp^3}	0.74	-0.04	Br	1.14	0.48
O_{sp^2}	0.62	-0.20	I	1.33	0.73

Kappa indices can also be calculated for molecular fragments and functional groups X. The calculation of these indices for groups was performed using a “pseudomolecule” X–X: two fragments X of the same kind are linked together, kappa values are calculated for the pseudomolecule, and these are then divided by two.

To quantify the shape of the whole molecule, Kier proposed a linear combination of the above-defined κ indices, each representing a particular shape attribute of the molecule:

$$shape = b_0 \cdot {}^0\kappa + b_1 \cdot {}^1\kappa + b_2 \cdot {}^2\kappa + b_3 \cdot {}^3\kappa$$

where ${}^0\kappa$ is the $\rightarrow$ Kier symmetry index used to encode the shape contribution due to symmetry.

Specific combinations of κ indices were also proposed as indices of molecular flexibility ($\rightarrow$ Kier molecular flexibility index) and steric effects ($\rightarrow$ Kier steric descriptor).

📖 [Kier, 1986c, Kier, 1987a, 1987b, 1987c, 1990, 1997; Gombar and Jain, 1987; Mokrosz, 1989; Hall and Kier, 1991; Skvortsova, Baskin *et al.*, 1993; Hall and Vaughn, 1997]

- **Kier symmetry index** $\rightarrow$ symmetry descriptors
- **K_Z index** $\rightarrow$ Hosoya Z matrix
- **K inflation factor** $\rightarrow$ variable reduction ($\odot$ K correlation analysis)
- **Kirchhoff matrix** $\equiv$ Laplacian matrix
- **Kirchhoff number** $\rightarrow$ resistance matrix
- **Kirchhoff sum index** $\equiv \Omega/D$ index $\rightarrow$ resistance matrix
- **Kirkwood function** $\rightarrow$ physico-chemical properties ($\odot$ dielectric constant)
- **KLOGP** $\rightarrow$ lipophilicity descriptors ($\odot$ Klopman hydrophobic models)
- **Klopman–Henderson cumulative substructure count** $\rightarrow$ scoring functions
- **Klopman hydrophobic atomic constants** $\rightarrow$ lipophilicity descriptors ($\odot$ Klopman hydrophobic models)
- **Klopman hydrophobic models** $\rightarrow$ lipophilicity descriptors
- **Klopman LOG P** $\equiv$ KLOG P $\rightarrow$ lipophilicity descriptors ($\odot$ Klopman hydrophobic models)
- **k-matching** $\rightarrow$ graph
- **Kohonen artificial neural networks** $\equiv$ Self-Organizing Maps
- **Kohonen maps** $\equiv$ Self-Organizing Maps
- **KOKOS descriptors** $\rightarrow$ biodescriptors ($\odot$ amino acid descriptors)
- **Koppel–Paju B scale** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- **Kovats retention index** $\rightarrow$ chromatographic descriptors
- **KOWWIN**[®] $\rightarrow$ lipophilicity descriptors
- **Krygowski bond energy** $\rightarrow$ delocalization degree indices
- **Kubinyi fitness function** $\rightarrow$ regression parameters
- **Kuhn length** $\rightarrow$ size descriptors
- **Kulczynski similarity coefficients** $\rightarrow$ similarity/diversity ($\odot$ Table S9)
- **Kullback–Leibler divergence** $\rightarrow$ information content
- **Kupchik modified connectivity indices** $\rightarrow$ connectivity indices
- **Kupchik vertex degree** $\rightarrow$ vertex degree
- **kurtosis** $\rightarrow$ statistical indices ($\odot$ moment statistical functions)



- **Labeled Hydrogen-Filled molecular Graph** $\equiv$ *H-filled molecular graph* $\rightarrow$ molecular graph
- **Labeled Hydrogen-Suppressed molecular Graph** $\equiv$ *H-depleted molecular graph* $\rightarrow$ molecular graph
- **labyrinthicity** $\rightarrow$ walk counts
- **Laffort solute descriptors** $\rightarrow$ Linear Solvation Energy Relationships
- **Lagrange distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)
- **Lance-Williams distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)
- **Laplacian graph energy** $\rightarrow$ spectral indices

■ **Laplacian matrix (L)** ($\equiv$ *admittance matrix, Kirchhoff matrix, combinatorial Laplacian matrix*)

This is a square $A \times A$ symmetric matrix, A being the number of vertices in the $\rightarrow$ *molecular graph*, obtained as the difference between the $\rightarrow$ *vertex degree matrix V* and the $\rightarrow$ *adjacency matrix A* [Mohar, 1989b, 1989a]:

$$\mathbf{L} = \mathbf{V} - \mathbf{A} = \mathbf{V}^{1/2} \cdot (\mathbf{I} - \mathbf{H}) \cdot \mathbf{V}^{1/2}$$

where $\mathbf{V}$ is the diagonal $\rightarrow$ *vertex degree matrix* of dimension $A \times A$ whose diagonal entries are the $\rightarrow$ *vertex degrees* δ_i . In the last expression, $\mathbf{I}$ indicates the $\rightarrow$ *identity matrix* and $\mathbf{H}$ is a matrix derived from the $\rightarrow$ *random walk Markov matrix MM* by a similarity transformation. The matrix $\mathbf{I} - \mathbf{H}$ is sometimes called the **normalized Laplacian matrix** [Chung, 1997]. Note that the negative half-sum of elements of the normalized Laplacian matrix is the $\rightarrow$ *Randić connectivity index* ${}^1\chi$ [Klein, Palacios *et al.*, 2004]:

$${}^1\chi = -\frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [(\mathbf{I} - \mathbf{H})]_{ij}$$

The entries of the Laplacian matrix formally are

$$[\mathbf{L}]_{ij} = \begin{cases} \delta_i & \text{if } i = j \\ -1 & \text{if } (i, j) \in \mathcal{E}(\mathcal{G}) \\ 0 & \text{if } (i, j) \notin \mathcal{E}(\mathcal{G}) \end{cases}$$

where $\mathcal{E}(\mathcal{G})$ is the set of edges of the molecular graph $\mathcal{G}$.

The Laplacian matrix is also related to the vertex-edge and edge-vertex $\rightarrow$ *incidence matrices*.

The diagonalization of the Laplacian matrix gives A real eigenvalues λ_i that constitute the **Laplacian spectrum** [Mohar, 1991b; Trinajstić, Babic *et al.*, 1994] and are conventionally labeled so that

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_A$$

Among the several properties of the Laplacian eigenvalues, three important ones are

- (a) the Laplacian eigenvalues are nonnegative numbers;
- (b) the last eigenvalue λ_A is always equal to zero;
- (c) the eigenvalue λ_{A-1} is greater than zero if, and only if, the graph G is connected; therefore, for a molecular graph all the Laplacian eigenvalues except the last are positive numbers.

Moreover, the sum of the positive eigenvalues is equal to twice the number B of graph edges, that is,

$$\sum_{i=1}^{A-1} \lambda_i \equiv \text{tr}(\mathbf{L}) = 2 \cdot B$$

The sum of the reciprocal $A - 1$ positive eigenvalues was proposed as a molecular descriptor [Mohar, Babic *et al.*, 1993; Gutman, Yeh *et al.*, 1993] and called the **quasi-Wiener index** W^* [Marković, Gutman *et al.*, 1995]; it is defined as

$$W^* = A \cdot \sum_{i=1}^{A-1} \frac{1}{\lambda_i}$$

For acyclic graphs, the quasi-Wiener index W^* coincides with the $\rightarrow$ *Wiener index* W , that is, $W^* = W$, while for cycle-containing graphs the two descriptors differ. Moreover, it has been demonstrated that the quasi-Wiener index coincides with the $\rightarrow$ *Kirchhoff number* for any graph [Gutman and Mohar, 1996].

The product of the positive $A - 1$ eigenvalues of the Laplacian matrix gives the **spanning tree number** T^* of the molecular graph G as

$$T^* = \frac{1}{A} \cdot \prod_{i=1}^{A-1} \lambda_i = \frac{|a|}{A}$$

where the $\rightarrow$ *spanning tree* is a connected acyclic subgraph containing all the vertices of G [Trinajstić, Babic *et al.*, 1994]. The term a in the second equality is the coefficient of the linear term in the $\rightarrow$ *Laplacian polynomial* [Nikolić, Trinajstić *et al.*, 1996b]. The number of spanning trees of a graph is used as a measure of $\rightarrow$ *molecular complexity* for polycyclic graphs; it increases with the complexity of the molecular structure. It has to be noted that some specific algorithms have been proposed to calculate the number of spanning trees in molecular graphs of cata-condensed systems [John, Mallion *et al.*, 1998].

Moreover, the **spanning-tree density** (STD) and the **reciprocal spanning-tree density** (RSTD) were defined as [Mallion and Trinajstić, 2003]

$$\text{STD} = \frac{T^*}{{}^eN} \quad \text{STD} \leq 1 \quad \text{RSTD} = \frac{{}^eN}{T^*} \quad \text{RSTD} \geq 1$$

where ${}^eN$ is the number of ways of choosing any $A - 1$ edges belonging to the set $\mathcal{E}(G)$ of graph edges. RSTD was proposed as a measure of *intricacy* of a graph, that is, the bigger RSTD is, the more intricate G .

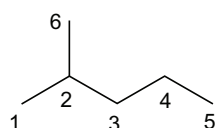
Also derived from the Laplacian matrix are the **Mohar indices** TI_1 and TI_2 , defined as

$$TI_1 = 2 \cdot A \cdot \log\left(\frac{B}{A}\right) \cdot \sum_{i=1}^{A-1} \frac{1}{\lambda_i} = 2 \cdot \log\left(\frac{B}{A}\right) \cdot W^* \quad TI_2 = \frac{4}{A \cdot \lambda_{A-1}}$$

where λ_{A-1} is the first nonzero eigenvalue and W^* the quasi-Wiener index [Trinajstić, Babić *et al.*, 1994]. Being $W^* = W$ for acyclic graphs, the first Mohar index TI_1 is closely related to the Wiener index for acyclic graphs.

Example L1

Laplacian matrix L , its eigenvalues and some related indices for 2-methylpentane.



$$W^* = 6 \times 5.33305 = 31.998 \approx 32 = W$$

$$T^* = \frac{6.000}{6} = 1$$

$$TI_1 = -5.0676 \quad TI_2 = 2.0519$$

Atom	1	2	3	4	5	6
1	1	-1	0	0	0	0
2	-1	3	-1	0	0	-1
3	0	-1	3	-1	0	0
4	0	0	-1	2	-1	0
5	0	0	0	-1	1	0
6	0	-1	0	0	0	1

ID	Eigenvalues
1	4.2143
2	3.0000
3	1.4608
4	1.0000
5	0.3249
6	0.0000

With some analogy to the Laplacian matrix is the **second path matrix**, denoted by S and defined as [John and Diudea, 2004]

$$S = A^2 - V$$

where A^2 is the square adjacency matrix and V is the diagonal matrix of the vertex degrees.

The sum of the entries of the matrix S coincides with the $\rightarrow$ *Platt number* F and twice the $\rightarrow$ *Gordon–Scantlebury index* N_{GS} :

$$\sum_{i=1}^A \sum_{j=1}^A [S]_{ij} = F = 2 \cdot N_{GS}$$

Moreover, the **quasi-Euclidean matrix**, denoted as ρ_{qE} , was defined as [Ivanciuc, Ivanciuc *et al.*, 2001b; Zhu and Klein, 1996]

$$[\rho_{qE}]_{ij} = \begin{cases} [\Gamma^2]_{ii} + [\Gamma^2]_{jj} - 2 \cdot [\Gamma^2]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where Γ is the generalized inverse of the Laplacian matrix L . The off-diagonal elements $[\rho_{qE}]_{ij}$ of this matrix are called **quasi-Euclidean distances**.

📖 [Ivanciuc, 1993; Gutman, Lee *et al.*, 1994; Nikolić, Trinajstić *et al.*, 1996b; Chan, Lam *et al.*, 1997; Xiao, 2004]

- **Laplacian polynomial** → characteristic polynomial-based descriptors
- **Laplacian spectrum** → Laplacian matrix
- **lateral validation** → validation techniques
- **lattice representation** $\equiv$ *stereoelectronic representation* → molecular descriptors

■ **layer matrices (LM)** ($\equiv$ *shell matrices*)

A layer matrix **LM** of a → *molecular graph* G is a rectangular unsymmetrical matrix $A \times (D + 1)$, A being the number of graph vertices and D the → *topological diameter*. The entry i - k (lm_{ik}) is the sum of the weights of the vertices located in the concentric shell (layer) at → *topological distance* k around the vertex v_i [Diudea, Minailiuc *et al.*, 1991; Diudea, 1994; Skorobogatov and Dobrynin, 1988]. The k th layer of the vertex v_i is the set $V_{ik}(G)$ of vertices defined as the following:

$$V_{ik}(G) = \{a | a \in V(G); d_{ia} = k\}$$

where d_{ia} is the topological distance of the a th vertex from v_i .

The entries of the layer matrix are formally defined as

$$lm_{ik} = \sum_{a \in V_{ik}} w_a \quad \text{or} \quad lm_{ik} = \sum_{j=1}^A w_j \cdot \delta(d_{ij}; k)$$

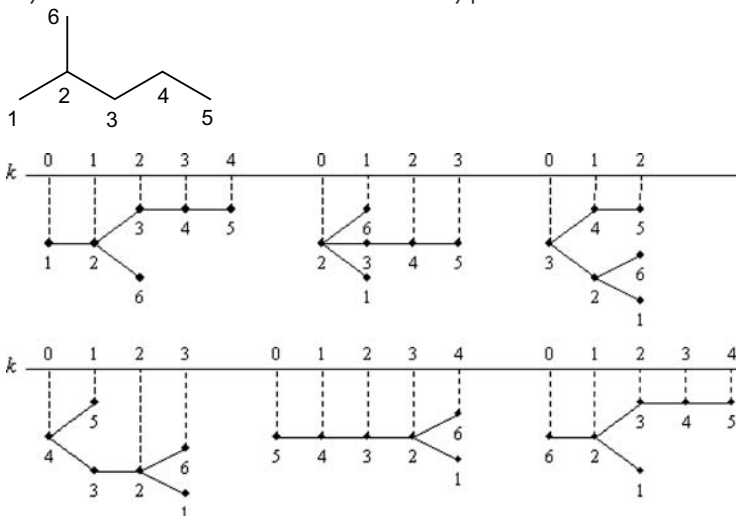
where w is the → *weighting scheme* for graph vertices, and $\delta(d_{ij}; k)$ the Dirac delta function equal to one when the distance d_{ij} between vertices v_i and v_j is equal to k , and zero otherwise.

The columns of the layer matrix are $D + 1$, D being the topological diameter and the case $k = 0$ is also considered, meaning that the property of the focused i th vertex is also considered.

The weights w for graph vertices can be any chemical or topological atomic properties. Examples of chemical → *atomic properties* are → *van der Waals volume*, atomic mass, → *polarizability*, examples of → *local vertex invariants* are → *vertex degree*, → *path degree*, → *walk degree*.

Example L2

Layer structures for the six vertices of 2-methylpentane.



Based on the different atomic properties, several layer matrices can be obtained; the most common are defined below.

• **cardinality layer matrix (LC)**

The simplest layer matrix obtained weighting all vertices by a weight equal to one [Diudea, 1994]. Therefore, the entry $i-k$ of the matrix is the number lc_{ik} of vertices located at distance k from the focused i th vertex. For the cardinality layer matrix, the following relations hold:

$$\sum_{k=0}^D lc_{ik} = \sum_{i=1}^A lc_{i0} = A \quad \text{and} \quad \sum_{i=1}^A lc_{i1} = 2 \cdot B$$

where A is the number of graph vertices, B the number of graph edges and D the molecule diameter that is the largest distance in the graph.

The cardinality layer matrix was originally called **L matrix** [Skorobogatov and Dobrynin, 1988] and **F matrix** [Diudea and Pârnu, 1988]. Moreover, for acyclic graphs, the cardinality layer matrix coincides with the $\rightarrow$ *path-layer matrix*.

By this matrix, the $\rightarrow$ *information layer index* is calculated. Moreover, four different local vertex invariants derived from the cardinality layer matrix have been proposed [Wang, Milne *et al.*, 1994] as the following:

$$\begin{aligned} \gamma_i &= \log \left(\sum_{k=1}^D lc_{ik} \cdot 2^{-k} \right) & \gamma_i &= \log \left(\sum_{k=1}^D lc_{ik} \cdot 4^{-k} \right) \\ \gamma_i &= \log \left(\sum_{k=1}^D lc_{ik} / (k+1) \right) & \gamma_i &= \log \left(\sum_{k=1}^D lc_{ik} / (k+1)^{3/2} \right) \end{aligned}$$

where D is the maximum topological distance in the graph.

• **branching layer matrix (LB)**

A layer matrix obtained weighting all the vertices in the graph by their $\rightarrow$ *vertex degrees* δ [Diudea, Minailiuc *et al.*, 1991]. Therefore, the entry $i-k$ of the matrix is the sum of the vertex degrees over all vertices in the k th layer around the focused i th vertex:

$$lb_{ik} = \sum_{j=1}^A \delta_j \cdot \delta(d_{ij}; k)$$

where δ_j is the vertex degree of the j th atom, d_{ij} is the distance between vertices v_i and v_j , and $\delta(d_{ij}, k)$ the Dirac delta function equal to one when the distance d_{ij} is equal to k and zero otherwise. For acyclic graphs, the elements of the branching layer matrix coincide with the $\rightarrow$ *valence shells* of the graph vertices proposed by Randić to compute weighted path counts [Randić, 2001g].

Each i th row of the matrix **LB** expresses the global state of vertex degrees from the viewpoint of vertex i , that is, the distribution of sums of vertex degrees in shells around the i th vertex. The

sum of each i th row element is a constant equal to $2B$, twice the number of graph edges. Moreover, more branched and $\rightarrow$ *central vertices* show higher values in the first layers within the corresponding rows, while less branched and $\rightarrow$ *terminal vertices* have higher values in the far layers. The vertex degrees of the atoms are in the first column ($k = 0$).

- **connectivity valence layer matrix (LCV)**

Similar to the branching layer matrix, this matrix is obtained weighting the vertices by their $\rightarrow$ *valence vertex degree* δ^v instead of the simple vertex degree δ [Hu and Xu, 1996].

- **edge layer matrix (LE)**

Analogously to the branching layer matrix, the edge layer matrix **LE** is obtained weighting all vertices by the number of distinct edges incident to the vertices of the k th layer around the vertex v_i , without counting any edge already counted in a preceding layer [Diudea, Minailiuc *et al.*, 1991]. The sum of the elements of each i th row is a constant equal to the number of edges B , while the sum of the elements of each k th column is a constant equal to $2B$. The $\rightarrow$ *vertex degrees* of the atoms are in the first column ($k = 0$).

- **connectivity bond layer matrix (LCB)**

A layer matrix whose entry $i-k$ is defined as the sum of the $\rightarrow$ *conventional bond order* π^* of the edges connecting the vertices situated in the k th layer with the vertices of the $(k - 1)$ th layer with respect to the focused i th vertex [Hu and Xu, 1996].

- **sum layer matrix (LS)**

To increase the discriminating power of atomic and molecular descriptors derived from the layer matrices, the sum layer matrix **LS** was also defined, and its entries are the sums of the corresponding entries of the branching layer matrix **LB** (lb_{ik}) and edge layer matrix **LE** (le_{ik}) [Diudea, Minailiuc *et al.*, 1991]:

$$ls_{ik} = lb_{ik} + le_{ik}$$

The sum of the elements of each i th row is a constant equal to $3B$, B being the number of graph edges.

- **distance sum layer matrix (LDS)**

A layer matrix obtained weighting the vertices by their $\rightarrow$ *vertex distance degree* σ , that is, the row sum of the $\rightarrow$ *distance matrix* **D** [Balaban and Diudea, 1993]. Therefore, the entry $i-k$ of the layer matrix is the sum of the vertex distance degrees of the vertices located at distance k from the focused i th vertex. It is obvious that the entries of the first column ($k = 0$) are only the vertex distance degrees. Moreover, the sums over each row in **LDS** are all equal to twice the $\rightarrow$ *Wiener index*, that is, the following relation holds:

$$\sum_{k=0}^D ls_{ik} = \sum_{i=1}^A ls_{i0} = 2 \cdot W$$

where A is the number of graph vertices and W the Wiener index.

- **geometric sum layer matrix (LGS)**

A layer matrix defined by analogy with the distance sum layer matrix deriving the weighting scheme for vertices from the $\rightarrow$ *geometry matrix* **G** instead of the $\rightarrow$ *distance matrix* **D** [Diudea, Horvath *et al.*, 1995b]. Therefore, the entry $i-k$ of this layer matrix is the sum of the $\rightarrow$ *geometric distance degrees* $^G\sigma$ of the vertices located at distance k from the focused i th vertex, where the geometric distances are obtained by methods of $\rightarrow$ *computational chemistry*. The sums over each row in **LGS** as well as the zero-column sum are all equal to twice the $\rightarrow$ *3D-Wiener index*.

- **path degree layer matrix (LPD)**

A layer matrix obtained weighting all vertices by the $\rightarrow$ *path degree* ξ . Therefore, the entry $i-k$ of the matrix is the sum of the path degrees of all vertices located at distance k from the focused i th vertex. The half-sum of the elements in the first column ($k=0$) is equal to the half-sum of the elements in each row of matrix **LPD** and corresponds to a molecular descriptor repurposed as the $\rightarrow$ *all-path Wiener index* W^{AP} , that is, the following relations hold:

$$\sum_{k=0}^D lpd_{ik} = \sum_{i=1}^A lpd_{i0} = 2 \cdot W^{AP}$$

where A is the number of graph vertices.

Note that for acyclic graphs, the path degree layer matrix **LPD** coincides with the distance sum layer matrix **LDS**.

- **walk degree layer matrix ($LW^{(m)}$)**

A layer matrix obtained weighting the vertices by their $\rightarrow$ *walk degree* (i.e., the $\rightarrow$ *atomic walk count* of length m , $awc_i^{(m)}$). Therefore, the entry $i-k$ of the layer matrix (lw_{ik}) is the sum of the walk degrees of the vertices located at distance k from the focused i th vertex [Diudea, Topan *et al.*, 1994]. Different matrices $LW^{(m)}$ can be calculated according to the chosen order m of the walk degrees $awc_i^{(m)}$. The walk degree layer matrix $LW^{(1)}$ coincides with the branching layer matrix **LB**. The elements of the first column ($k=0$) in the matrix $LW^{(m)}$ represent only the walk degrees of order m .

Moreover, the half-sum of both the entries in the first column ($k=0$) and in each row is the total number of walks of length m in the graph, as can be seen from the following relationships:

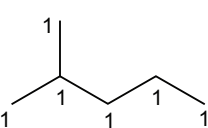
$$\frac{1}{2} \cdot \sum_{k=0}^D lw_{ik}^{(m)} = \frac{1}{2} \cdot \sum_{i=1}^A lw_{i0}^{(m)} = \frac{1}{2} \cdot \sum_{i=1}^A awc_i^{(m)} = mwc^{(m)}$$

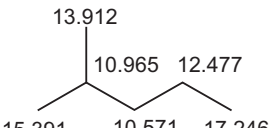
where A is the number of graph vertices, $awc_i^{(m)}$ the atomic walk count, and $mwc^{(m)}$ is the m th order $\rightarrow$ *molecular walk count*.

Analogous layer matrices can be obtained using, as the weighting scheme, $\rightarrow$ *weighted walk degrees* instead of the simple vertex walk degrees.

Example L3

Cardinality layer matrix **LC** and geometric sum layer matrix **LGS** for 2-methylpentane. Matrix rows represent vertices (*i*), while matrix columns represent layers (*k*). Vertices in the graphs are labeled according to the unitary weighting scheme for cardinality layer matrix and geometric distance degrees for geometric sum layer matrix.





LC =

<i>i/k</i>	0	1	2	3	4
1	1	1	2	1	1
2	1	3	1	1	0
3	1	2	3	0	0
4	1	2	1	2	0
5	1	1	1	1	2
6	1	1	2	1	1

LGS =

<i>i/k</i>	0	1	2	3	4
1	15.391	10.955	24.483	12.477	17.246
2	10.955	39.874	12.477	17.246	0
3	10.571	23,432	46.549	0	0
4	12.477	27.817	10.955	29.303	0
5	17.246	12.477	10.571	10.955	29.303
6	13.912	10.955	25.962	12.477	17.246

Derived from layer matrices, two main types of $\rightarrow$ local vertex invariants are defined on the basis of two types of operators, the **centric operator** c_i and the **centrocomplexity operator** x_i ; [Diudea, 1994; Diudea, Horvath *et al.*, 1995b]:

$$c_i(\mathbf{LM}) = \left[\sum_{k=1}^D (lm_{ik})^{k/d} \right]^{-1} \quad \text{and} \quad x_i(\mathbf{LM}) = \left[\frac{1}{w_i} \cdot \sum_{k=0}^D lm_{ik} \cdot 10^{-zk} \pm l_i \right]^{\pm 1} \cdot t_i$$

where

$$l_i = f_i \cdot \left(\frac{lm_{i0}}{10} + \frac{lm_{i1}}{100} \right) \quad \text{and} \quad f_i = \sum_{j=1}^A a_{ij} \cdot (\pi_{ij}^* - 1) \quad \pi_{ij}^* = 0 \text{ if } (i, j) \notin E(\mathcal{G})$$

where d is a specified topological distance larger than the topological diameter D (for example, 10); w_i is the atomic property of the i th vertex, z is the number of figures of the largest entry lm_{ik} value; l_i is a local parameter accounting for multiple bonds, f_i being the $\rightarrow$ atomic multigraph factor obtained by summing up the conventional bond orders π_{ij}^* of the vertices j adjacent to the i th vertex; t_i is a weighting factor accounting for heteroatoms by means of atomic numbers, electronegativities, covalent radii, and so on.

A third type of operator generating local invariants from layer matrices is defined as follows [Balaban and Diudea, 1993]:

$$x_{ji}(\mathbf{LM}) = \sum_{j=1}^A a_{ij} \cdot \left[\frac{\sum_{k=0}^D lm_{ik} \cdot 10^{-zk}}{t_i \cdot (1 + f_i)} \cdot \frac{\sum_{k=0}^D lm_{jk} \cdot 10^{-zk}}{t_j \cdot (1 + f_j)} \right]^{-1/2}$$

where the sum runs over all vertices j adjacent to the i th vertex, that is, located at distance one from v_i .

Centrocomplexity invariant values should measure the location of the considered vertex with respect to a vertex “of importance”, that is, a vertex of highest branching degree, electro-negativity, and so on, while centric invariant values should measure the centrality of a vertex, that is, its location with respect to the $\rightarrow$ graph center.

The local indices obtained by applying centrocomplexity operator to the branching layer matrix **LB** are called **regressive vertex degrees** [Diudea, Minailiuc *et al.*, 1991]. Such indices are an extension of the concept of $\rightarrow$ vertex degree, taking into account the contributions of distant vertices from the focused i th vertex; these contributions decrease with increasing distance, slightly augmenting the value of the classical vertex degree. High values of regressive vertex degrees should correspond to vertices of highest degree, closest to branching sites and to the graph center. The two types of originally proposed centrocomplexity operators were defined as

$$RVD_i \equiv x_i^{R1}(\mathbf{LB}) = \sum_{k=0}^D lb_{ik} \cdot k^{-3} = \delta_i + \sum_{k=1}^D lb_{ik} \cdot k^{-3}$$

$$RVD_i \equiv x_i^{R2}(\mathbf{LB}) = \sum_{k=0}^D lb_{ik} \cdot 10^{-k} = \delta_i + \sum_{k=1}^D lb_{ik} \cdot 10^{-k}$$

where **LB** is the branching layer matrix and δ the vertex degree.

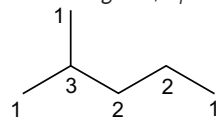
Example L4

Branching layer matrix **LB** and regressive vertex degrees for 2-methylpentane.

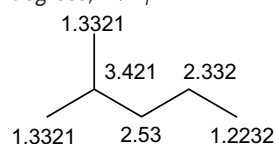
LB =

i/k	0	1	2	3	4
1	1	3	3	2	1
2	3	4	2	1	0
3	2	5	3	0	0
4	2	3	3	2	0
5	1	2	2	3	2
6	1	3	3	2	1

Vertex degrees, δ_i



Regressive vertex degrees, RVD_i



$$RVD_1 = RVD_6 = \sum_{k=0}^4 lb_{1k} \cdot 10^{-k} = 1 + 0.3 + 0.03 + 0.002 + 0.0001 = 1.3321$$

$$RVD_2 = \sum_{k=0}^4 lb_{2k} \cdot 10^{-k} = 3 + 0.4 + 0.02 + 0.001 = 3.421$$

$$RVD_3 = \sum_{k=0}^4 lb_{3k} \cdot 10^{-k} = 2 + 0.5 + 0.03 = 2.53$$

$$RVD_4 = \sum_{k=0}^4 lb_{4k} \cdot 10^{-k} = 2 + 0.3 + 0.03 + 0.002 = 2.332$$

$$RVD_5 = \sum_{k=0}^4 lb_{5k} \cdot 10^{-k} = 1 + 0.2 + 0.02 + 0.003 + 0.0002 = 1.2232$$

By analogy with the regressive vertex degrees, the **regressive distance sums** (or **regressive incremental distance sums**) are local invariants obtained by applying the centrocomplexity operator x_i to the distance sum layer matrix **LDS** [Balaban and Diudea, 1993]. Also in this case, a simplified form of the centrocomplexity operator has been proposed:

$$RDS_i \equiv x_i(\mathbf{LDS}) = \sum_{k=0}^D (lds_{ik} \cdot 10^{-zk}) = \sigma_i + \sum_{k=1}^D (lds_{ik} \cdot 10^{-zk})$$

where σ_i is the distance sum (i.e., vertex distance sum) of the i th vertex and z is the number of figures of the largest entry lds_{ik} value.

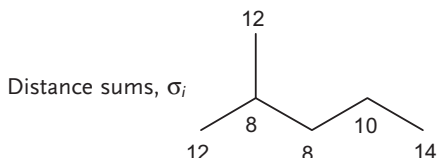
To obtain greater discrimination between the terminal and central vertices, the **regressive decremental distance sums** were proposed [Balaban, 1995b]. They are calculated from the distance sum layer matrix **LDS** by the following:

$$RDDS_i = \sigma_i - \sum_{k=1}^D (lds_{ik} \cdot 10^{-zk})$$

where σ_i is the $\rightarrow$ distance sum of the i th vertex. In this way, the progressively attenuated contributions due to more distant vertices are subtracted from the distance degree of the focused vertex.

Example L5

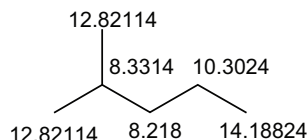
Distance sum layer matrix **LDS**, regressive distance sums, and regressive decremental distance sums for 2-methylpentane.



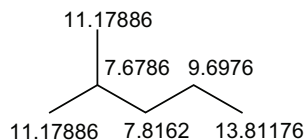
LDS =

i/k	0	1	2	3	4
1	12	8	20	10	14
2	8	32	10	14	0
3	8	18	38	0	0
4	10	22	8	24	0
5	14	10	8	8	24
6	12	8	20	10	14

Regressive distance sums, RDS_i



Regressive decremental distance sums, $RDDS_i$



$$RDS_1 = RDS_6 = \sum_{k=0}^4 lds_{1k} \cdot 10^{-k} = 12 + 0.8 + 0.020 + 0.0010 + .00014 = 12.82114$$

$$RDS_2 = \sum_{k=0}^4 lds_{2k} \cdot 10^{-k} = 8 + 0.32 + 0.010 + 0.0014 = 8.3314$$

$$\begin{aligned}
 RDS_3 &= \sum_{k=0}^4 lds_{3k} \cdot 10^{-k} = 8 + 0.18 + 0.038 = 8.218 \\
 RDS_4 &= \sum_{k=0}^4 lds_{4k} \cdot 10^{-k} = 10 + 0.22 + 0.08 + 0.0024 = 10.3024 \\
 RDS_5 &= \sum_{k=0}^4 lds_{5k} \cdot 10^{-k} = 14 + 0.10 + 0.08 + 0.008 + 0.00024 = 14.18824 \\
 RDDS_1 &= RDDS_6 = \sigma_1 - \sum_{k=0}^4 lds_{1k} \cdot 10^{-k} = 12 - 0.8 - 0.020 - 0.0010 - 0.00014 = 11.17886 \\
 RDDS_2 &= \sigma_2 - \sum_{k=0}^4 lds_{2k} \cdot 10^{-k} = 8 - 0.32 - 0.010 - 0.0014 = 7.6786 \\
 RDDS_3 &= \sigma_3 - \sum_{k=0}^4 lds_{3k} \cdot 10^{-k} = 8 - 0.18 - 0.038 = 7.8162 \\
 RDDS_4 &= \sigma_4 - \sum_{k=0}^4 lds_{4k} \cdot 10^{-k} = 10 - 0.22 - 0.08 - 0.0024 = 9.6976 \\
 RDDS_5 &= \sigma_5 - \sum_{k=0}^4 lds_{5k} \cdot 10^{-k} = 14 - 0.10 - 0.08 - 0.00024 = 13.81176
 \end{aligned}$$

The sums of the local vertex invariants x_i and c_i over all of the vertices give the corresponding molecular descriptors, called **centrocomplexity topological index** X and **centric topological index** C , respectively:

$$X(\mathbf{LM}) = \sum_{i=1}^A x_i(\mathbf{LM}) \quad \text{and} \quad C(\mathbf{LM}) = \sum_{i=1}^A c_i(\mathbf{LM})$$

where $\mathbf{LM}$ represents any layer matrix. These descriptors are related to $\rightarrow$ *molecular complexity*. Normalized centrocomplexity and centric local invariants x'_i and c'_i are obtained dividing each local invariant by the corresponding global topological index.

📖 [Diudea and Bal, 1990; Diudea and Kacso, 1991, 1992; Dobrynin, 1993; Ivanciuc, Balaban *et al.*, 1993b, 1992; Wang, Milne *et al.*, 1994; Dobrynin and Mel'nikov, 2001; Diudea, Jäntschi *et al.*, 2002; Diudea, 2002a; Diudea and Ursu, 2003; Konstantinova and Vidyuk, 2003]

- **LCD descriptors** $\rightarrow$ molecular descriptors ($\odot$ invariance properties of molecular descriptors)
- **LDOS** $\equiv$ *Local Density Of States* $\rightarrow$ quantum-chemical descriptors ($\odot$ EIM descriptors)
- **LEACH index** $\rightarrow$ environmental indices ($\odot$ leaching indices)
- **leaching indices** $\rightarrow$ environmental indices
- **lead compound** $\rightarrow$ drug design
- **leading eigenvalue** $\rightarrow$ spectral indices
- **leading eigenvalue of the distance matrix** $\rightarrow$ spectral indices ($\odot$ eigenvalues of the distance matrix)
- **leading eigenvalue of $(A + D)$** $\rightarrow$ spectral indices

- **lead-like indices** → property filters
- **learning set** $\equiv$ *training set* → data set
- **leave-more-out technique** → validation techniques (⊙ cross-validation)
- **leave-one-out technique** → validation techniques (⊙ cross-validation)
- **length-to-breadth ratio** → shape descriptors (⊙ Kaliszan shape parameter)
- **Lennard–Jones 6–12 potential function** → molecular interaction fields (⊙ steric interaction fields)
- **Leo–Hansch hydrophobic fragmental constants** → lipophilicity descriptors
- **Lethal Concentration** → biological activity indices (⊙ toxicological indices)
- **Lethal Dose** → biological activity indices (⊙ toxicological indices)
- **Lethal Time** → biological activity indices (⊙ toxicological indices)
- **level pattern indices** → spectral indices (⊙ eigenvalues of the adjacency matrix)
- **leverage matrix** → chemometrics (⊙ regression analysis)
- **lhaf(D) index** → algebraic operators (⊙ determinant)
- **ligand** → drug design
- **L index** → combined descriptors
- **L_Z index** → Hosoya Z matrix
- **LIN index** → environmental indices (⊙ leaching indices)
- **linear aromatic substituent reactivity relationships** $\equiv$ *Yukawa–Tsuno equation* → electronic substituent constants (⊙ resonance electronic constants)
- **Linear Free Energy Relationships** → extrathermodynamic approach
- **linear graph** → graph
- **linear indices** → TOMOCOMD descriptors
- **linearity index** → shape descriptors
- **linear notation systems** → molecular descriptors
- **linear similarity index** → quantum similarity

■ Linear Solvation Energy Relationships (LSERs)

Linear solvation energy relationships constitute the basis on which effects of solvent–solute interactions on → *physico-chemical properties* and reactivity parameters are studied. In general, a property $\mathcal{P}$ of a species A in a solvent S can be expressed as

$$\mathcal{P}_{A,S} = \sum_j \varphi_j(A, S)$$

where φ are complex functions of both solvents and solutes [Kamlet, Abboud *et al.*, 1981]. By assuming that these functions can be factorized into two contributions separately dependent on solute and solvent, the property can be represented as

$$\mathcal{P}_{A,S} = \sum_j f_j(A) \cdot g_j(S)$$

where f are functions of the solute and g functions of the solvent.

The underlying philosophy of the linear solvation energy relationships is based on the possibility to study these two functions, after a proper choice of the reference systems and properties. Moreover, it has been recognized that solution properties $\mathcal{P}$ mainly depend on three factors: a cavity term, a polar term, and hydrogen-bond term:

$$P = \text{intercept} + \text{cavity term} + \text{dipolarity/polarizability term} + \text{hydrogen-bond term}$$

Therefore, a typical linear solvation energy relationship is expressed as [Kamlet, Doherty *et al.*, 1987a]

$$P_{A,S} = b_0 + b_1 \cdot (\delta_H^2)_1 \cdot V_2 + b_2 \cdot \Pi_1^* \cdot \Pi_2^* + b_3 \cdot \alpha_1 \cdot \beta_2 + b_4 \cdot \beta_1 \cdot \alpha_2$$

where b are estimated regression coefficients, and the subscripts 1 and 2 in the solvent/solute property parameters refer to the solvent S and the solute A, respectively. This equation is usually known as **Abraham's general equation** or **solvatochromic equation** even if it is extended to cover some nonspectroscopic properties, and the parameters of polarity/dipolarizability and hydrogen-bonding as **solvatochromic parameters**. The term *solvatochromic* is derived from the origin of this approach referring to the effect solvent has on the color of an indicator that is used for quantitative determination of some molecular attributes (*solvatochromic parameters*).

From the general solvatochromic equation, two special cases can be encountered. When dealing with effects of different solvents on properties of a specific solute, the general equation is explicitly on solvent parameters:

$$P_{A,S_i} = b_0 + b_1 \cdot (\delta_H^2)_i + b_2 \cdot \Pi_{1,i}^* + b_3 \cdot \alpha_{1,i} + b_4 \cdot \beta_{1,i}$$

This equation has been used in several correlations of solvent effects on solute properties such as reaction rates and equilibrium constants of solvolyses, energy of electronic transitions, solvent induced shifts in UV/visible, IR, and NMR spectroscopy, fluorescence lifetimes, formation constants of hydrogen-bonded and Lewis acid/base complexes [Kamlet, Doherty *et al.*, 1986c].

Conversely, when dealing with solubilities, lipophilicity, or other properties of a set of different solutes in a specific solvent, the general equation is explicitly on the solute parameters:

$$P_{A_i,S} = b_0 + b_1 \cdot V_i + b_2 \cdot \Pi_{2,i}^* + b_3 \cdot \alpha_{2,i} + b_4 \cdot \beta_{2,i}$$

This equation has been mainly used in correlations of aqueous solubility of compounds, octanol/water partition coefficients and some other $\rightarrow$ *partition coefficients* together with some biological properties [Kamlet, Abraham *et al.*, 1984; Kamlet, Doherty *et al.*, 1986a, 1987a, 1987c, 1988c].

Other two general linear solvation energy relationships for solute physico-chemical properties in a fixed phase [Abraham, Whiting *et al.*, 1990b, 1991a, 1991b; Abraham, 1993d; Abraham, Andonian-Haftvan *et al.*, 1994] are

$$\log(P_{A_i,S}) = b_0 + b_1 \cdot V_{X,i} + b_2 \cdot R_{2,i} + b_3 \cdot \pi_{2,i}^H + b_4 \cdot \alpha_{2,i}^H + b_5 \cdot \beta_{2,i}^H$$

$$\log(P_{A_i,S}) = b_0 + b_1 \cdot L_i^{16} + b_2 \cdot R_{2,i} + b_3 \cdot \pi_{2,i}^H + b_4 \cdot \alpha_{2,i}^H + b_5 \cdot \beta_{2,i}^H$$

where the first can be applied to processes within condensed phases and the second to processes involving gas-condensed phase transfer. The solute descriptors in these relationships are called **Abraham-Klamt descriptors** and were successively denoted as [Zissimos, Abraham *et al.*, 2002c]

$$A \equiv \alpha^H; \quad B \equiv \beta^H; \quad S \equiv \pi^H; \quad E \equiv R; \quad V \equiv V_X$$

Similar equations but based on other solute descriptors were proposed in literature with the aim of better chromatographic data [Abraham, Ibrahim *et al.*, 2004]. In particular, five solute descriptors, here called **Laffort solute descriptors** (Table L1), were defined by Laffort *et al.* using GLC retention data on five stationary phases for 240 compounds [Laffort and Patte, 1976; Patte, Etcheto *et al.*, 1982]. These solute descriptors were used to fit a number of physico-chemical and biochemical properties. Note that in the first paper [Laffort and Patte, 1976], the five solute descriptors were obtained by $\rightarrow$ *Principal Component Analysis* on the data obtained from 25 stationary phases for 75 compounds, thus their numerical values differ from those obtained in the later paper.

Other five solute descriptors, here called **Wilson solute descriptors** (Table L1), were proposed by Wilson *et al.* using ten different HPLC stationary phases, all with acetonitrile 50% as the mobile phase, initially for 67 compounds and then extended to a larger class of compounds [Wilson, Nelson *et al.*, 2002a, 2002b; Wilson, Dolan *et al.*, 2002].

Weckwerth solute descriptors (Table L1) are five solute parameters based on $\rightarrow$ *Kovats retention index* on seven of GC stationary phases for 53 compounds [Weckwerth, Vitha *et al.*, 2001].

Table L1 Laffort, Wilson, and Weckwerth solute descriptors.

Laffort solute descriptors	Wilson solute descriptors	Weckwerth solute descriptors
Volume-sensitive apolar factor (α)	Hydrophobicity (η')	Solute volume (V)
Orientation factor, proportional to dipole moment of simple molecules (ω)	Steric parameter (σ')	Solute polarizability (P)
Electron factor, related to dispersion interactions (ϵ)	Basicity (β')	Solute dipolarity (D)
Hydrogen bond acidity (π)	Acidity (α')	Hydrogen bond acidity (A)
Hydrogen bond basicity (β)	Cation-exchange parameter (κ')	Hydrogen bond basicity (B)

The descriptors of the solvatochromic equation are specified below.

• cavity term

The cavity term is a measure of the endoergic cavity-forming process, that is, the free energy necessary to separate the solvent molecules, overcoming solvent–solvent cohesive interactions, and provide a suitably size cavity for the solute. The magnitude of the cavity term depends on the $\rightarrow$ *Hildebrand solubility parameter* δ_H and $\rightarrow$ *volume descriptors* of the solute. The solute volume can be measured in different ways, such as by $\rightarrow$ *van der Waals volume* V^{vdw} [Leahy, 1986], $\rightarrow$ *molar volume* $\bar{V}$ or $\rightarrow$ *Mc Gowan's characteristic volume* V_X ; in some cases, also $\rightarrow$ *molecular weight* MW has been used. Usually, the volumes are divided by 100 ($\bar{V}/100$) to obtain a more homogeneous scale with respect to the other parameters.

The parameter L^{16} is the $\rightarrow$ *Ostwald solubility coefficient* on *n*-hexadecane at 298 K; it includes both general dispersion interactions and endoergic cavity term and was proposed for modeling properties of solutes in processes involving gas-condensed phase transfer such as gas-liquid chromatographic parameters [Abraham, Grellier *et al.*, 1987].

• **dipolarity/polarizability term** ($\equiv$ *dipole term*)

This term is a measure of the exoergic balance (i.e., release of energy) of solute–solvent and solute–solute dipolarity/polarizability interactions. This term, denoted by Π^* , describes the ability of the compound to stabilize a neighboring charge or dipole by virtue of nonspecific dielectric interactions and is in general given by $\rightarrow$ *electric polarization descriptors* such as $\rightarrow$ *dipole moment* or other empirical $\rightarrow$ *polarity/polarizability descriptors* [Abraham, Grellier *et al.*, 1988]. Other specific polarity parameters empirically derived for linear solvation energy relationships are reported below.

Several **solvent polarity scales** were proposed to quantify the polar effects of solvents on physical properties and reactivity parameters in solution, such as rate of solvolyses, energy of electronic transitions, solvent induced shifts in IR, or NMR spectroscopy. Most of the polarity scales were derived by an empirical approach based on the principles of the $\rightarrow$ *linear free energies relationships* applied to a chosen reference property and system where hydrogen bonding effects are assumed negligible [Reichardt, 1965, 1990; Kamlet, Abboud *et al.*, 1981, 1983].

The most important scales of solvent polarity are:

π^* polarity scale. A solvent polarity parameter (also denoted as π^*_s) proposed by Kamlet, Abboud and Taft [Kamlet, Abboud *et al.*, 1977; Kamlet, Carr *et al.*, 1981], based on the solvatochromic shifts on the frequency maxima of the $\pi \rightarrow \pi^*$ transitions of seven different benzene derivatives. The π^* values are averaged on the seven compounds to prevent the inclusion of specific effects or spectral anomalies and are normalized so that π^* equals zero for cyclohexane and one for dimethylsulfoxide. This scale is one of the most comprehensive for the number of considered solvents and is widely used. Moreover, for solution properties ϕ involving different relative contributions of polarity and polarizability, π^* values can be corrected as $(\pi^* - d\delta_H)$, where δ_H is the $\rightarrow$ *Hildebrand solubility parameter*. The term d is calculated by dividing the difference in ϕ at $\pi^* = 0.7$, as obtained separately for nonpolychlorinated aliphatic and aromatic solvents, by the average of the slopes of the solvatochromic equations for aliphatic and aromatic solvents. The Hildebrand parameter is assumed to be $\delta_H = 0.0$ for nonpolychlorinated aliphatic solutes, $\delta_H = 0.5$ for polychlorinated aliphatics, and $\delta_H = 1.0$ for aromatic solutes; the term d ranges from zero, for maximal polarizability contributions to the studied property, to -0.40 for minimal contributions. The $\rightarrow$ *excess molar refractivity* R_2 has also been used as polarizability correction term instead of the Hildebrand parameter δ_H [Abraham, Lieb *et al.*, 1991].

Y polarity scale. A solvent polarity scale proposed by Grunwald and Winstein [Grunwald and Winstein, 1948] based on solvolytic rate k_0 of *t*-butyl chloride in 80% aqueous ethanol at 298 K. The Y-polarity value for a given solvent is calculated by

$$Y = \log k_s - \log k_0$$

where k_s is the solvolytic rate of *t*-butyl chloride in the considered solvent. Y scale was proposed as measure of an empirical “ionizing power” of solvents.

E_T polarity scale. A solvent polarity scale proposed by Dimroth, Reichardt, and coworkers [Dimroth, Reichardt *et al.*, 1963; Reichardt, 1965] based on the solvatochromic band shifts of the 4-(2,4,6-triphenylpyridinium)-2,6-diphenylphenoxide and its trimethyl derivative. This scale is one of the most comprehensive for the number of considered solvents and is widely used.

E_K polarity scale. A solvent polarity scale proposed by Walther [Walther, 1974] based on the hypsochromic shift of the longest wavelength absorption of a molybdenum complex.

Z polarity scale. A solvent polarity scale proposed by Kosower [Kosower 1958a, 1958b] based on the energy of the electronic transition of the 1-ethyl-4-carbomethoxypyridinium iodide that is strongly solvent-dependent. This is a measure of an internal charge transfer process. The original set of Z values being quite small, it was successively extended by means of other indicators (Table L2).

Table L2 Empirical parameters of the solvent polarity from different sources.

Solvent	γ	Z	E_T	E_K	π^*
H ₂ O	3.493	94.6	63.1		1.09
HCOOH	2.054				0.65
HCONH ₂	0.604	83.3	56.6		
CH ₃ COOH	−1.64	79.2	51.1	55.0	0.62
CH ₃ OH	−1.09	83.6	55.5	56.3	0.60
CH ₃ NO ₂			46.3		0.85
CH ₃ CN		71.3	46.0		0.75
C ₂ H ₅ OH	−2.03	79.6	51.9	55.3	0.54
C ₅ H ₅ N		64.0	40.2	57.0	0.87
CCl ₄			32.5	49.9	0.294
CH ₂ Cl ₂		64.2	41.1	53.9	0.802
C ₆ H ₆		54.0	34.5	53.4	0.588
C ₆ H ₅ Cl			37.5	53.9	0.709
C ₆ H ₅ Br			37.5	53.9	0.794
C ₆ H ₅ CN			42.0		0.933
C ₆ H ₅ NO ₂			42.0		1.006
Cyclo-C ₆ H ₁₂			31.2	49.0	0.000

A wide variety of correlations among solvent polarity scales was studied [Reichardt and Dimroth, 1968]; however, because of the different reference compounds used to define them, direct comparison should be done with caution [Bentley and von Schleyer, 1977].

The **solute polarity parameter** π_2^* was originally taken as identical with the solvent polarity parameter π^* for nonassociated liquids only [Taft, Abraham *et al.*, 1985]. Then, an alternative solute polarity parameter π_2^H (or $\sum \pi_2^H$) was proposed, which was based on experimental procedures that include, at least in principle, all types of solute molecules [Abraham, Whiting *et al.*, 1991a; Abraham and Whiting, 1992]. Values of this solute parameter were determined by back-calculation solving “inverse” solvation equation systems based on 30–70 stationary phases for each solute. π_2^H values refer to a situation in which a solute molecule is surrounded by an excess of solvent molecules and so they are effective values, more correctly denoted as $\sum \pi_2^H$, accounting for combined effects due to polyfunctional groups in the molecule.

• hydrogen-bond parameters

They are measures of the exoergic effects (i.e., release of energy) of the complexation between solutes and solvents.

The $\rightarrow$ *hydrogen bond donor* (HBD) power of a compound is called **hydrogen bond acidity** (or **hydrogen-bond electron-drawing power**) and is denoted by α_1 and α_2 for solvents and solutes, respectively.

The most important scales for **solvent HBD acidity** are here reported.

The α scale proposed to measure solvent hydrogen bond acidity, that is, the ability of a bulk solvent to act as hydrogen bond donor toward a solute, was derived from 16 diverse properties involving 13 solutes as averaged values [Taft and Kamlet, 1976; Kamlet, Abboud *et al.*, 1983].

The **Gutmann's Acceptor Number (AN)** was proposed [Gutmann, 1978] as quantitative empirical parameter of solvent hydrogen bond acidity based on ^{31}P -NMR shifts of triethylphosphine oxide at infinite dilution, calculated as $\text{AN} = -\delta_{\infty}^{\text{corr}} \cdot 2.349$.

For **solute HBD acidity**, different scales were proposed mainly based on complexation constants and enthalpies of complexation. The most important are here reported.

The α_{m} scale was proposed for solute HBD acidity of "monomer" amphihydrogen-bonding compounds acting as non-self-associated solutes [Taft, Abraham *et al.*, 1985; Kamlet, Doherty *et al.*, 1986a]. In particular, α_{m} values were derived from $\log K$ values for complexation with pyridine *N*-oxide in cyclohexane; this set of values was successively extended through various back-calculations using the solvatochromic equation.

The α_2^{H} scale was proposed for solute HBD acidity based on $\log K$ values for 1:1 complexation of series of acids against a given base in dilute solution of CCl_4 [Abraham, Grellier *et al.*, 1989]. Forty-five linear equations were solved for each considered base by a series of acids:

$$\log K_i = b_0^{\text{B}} + b_1^{\text{B}} \cdot \log K_{\text{A}}^{\text{H}_i}$$

where b^{B} are the regression coefficients characterizing each reference base, and $\log K_{\text{A}}^{\text{H}}$ values are characteristics of hydrogen-bonding acids, and hence represent the solute hydrogen bond acidities. All the equations intersect at a "magic" point where $\log K = -1.1$ (K measured on molar scale). The general $\log K_{\text{A}}^{\text{H}}$ values were then transformed into α_2^{H} values suitable for multivariate regression analysis by the following:

$$\alpha_2^{\text{H}} = \frac{\log K_{\text{A}}^{\text{H}} + 1.1}{4.636}$$

A fairly good correlation was found between α_2^{H} scale and α_{m} scale. Moreover, the set of original α_2^{H} values was then enlarged by solving a system of solvatochromic equations on partition coefficients, thus including several new compounds and molecular fragments [Abraham, Chadha *et al.*, 1994b]. The **effective solute hydrogen-bond acidity** $\sum \alpha_2^{\text{H}}$ was back-calculated by a number of multiple linear regression equations for solutes surrounded by a large excess of solvent and hence undergoing multiple hydrogen-bonding. This hydrogen bond descriptor agrees with α_2^{H} values for monofunctional compounds, while for polyfunctional compounds it significantly differs [Abraham, 1993d].

The $\rightarrow$ *hydrogen bond acceptor* (HBA) power of a compound is called **hydrogen bond basicity** (or **hydrogen-bond electron-acceptor power**) and is denoted by β_1 and β_2 for solvents and solutes, respectively.

The most important scales for **solvent HBA basicity** are here reported.

The β scale was proposed to measure solvent hydrogen bond basicity, that is, the ability of a bulk solvent to act as hydrogen bond acceptor. This scale was derived by systematic application of the solvatochromic comparison method; the final β values were calculated by averaging 13 β

parameters for each solvent obtained with different solutes and different physico-chemical properties [Kamlet, Abboud *et al.*, 1981, 1983].

The **Koppel–Paju B scale** was proposed to measure solvent hydrogen bond basicity, based on solvent shifts of the IR stretching frequencies of the free and hydrogen-bonded OH group of phenol in CCl₄ [Koppel and Paju, 1974].

The **Gutmann's Donor Number (DN)** was proposed [Gutmann, 1978] as quantitative empirical parameter for solvent nucleophilicity. For most solvents, it was found to well correlate with β scale.

The most important scale for **solute HBA basicity** are here reported.

The **β_m scale** was proposed for solute HBA basicity of “monomer” amphihydrogen-bonding compounds acting as non-self-associated solutes. In particular, β_m values are taken equal to β values for non-self-associating compounds.

The **β_2^H scale** was proposed for solute HBA basicity based on log K values for 1:1 complexation of series of basis against a number of reference acids in dilute solution of CCl₄ [Abraham, Grellier *et al.*, 1990]. Thirty-four linear equations were solved for each considered reference acid by a series of bases:

$$\log K_i = b_0^A + b_1^A \cdot \log K_B^H$$

where b^A are the regression coefficients characterizing each reference acid, and $\log K_B^H$ values are characteristics of the bases representing the solute hydrogen bond basicities. All the equations intersect at a “magic” point where $\log K = -1.1$ (K measured on molar scale). The general $\log K_B^H$ values were then transformed into β_2^H values suitable for multivariate regression analysis by the following:

$$\beta_2^H = \frac{\log K_B^H + 1.1}{4.636}$$

This transformation was proposed to obtain a basicity scale with a zero-point corresponding to all non-hydrogen-bonding bases, such as alkanes and cycloalkanes. Moreover, on this scale, hexamethylphosphoric triamide basicity is equal to one. The set of original β_2^H values was then enlarged by solving a system of solvatochromic equations on partition coefficients, thus including several new compounds and molecular fragments [Abraham, Chadha *et al.*, 1994b]. The **effective solute hydrogen-bond basicity** $\sum \beta_2^H$ was back-calculated by a number of multiple linear regression equations for solutes surrounded by a large excess of solvent and hence undergoing multiple hydrogen-bonding. This hydrogen bond descriptor agrees with β_2^H values for monofunctional compounds, while for poly-functional compounds it significantly differs [Abraham, Whiting *et al.*, 1991a; Abraham and Whiting, 1992; Abraham, 1993d].

For most solutes, the effective hydrogen-bond basicity is constant over all the solvent systems; however, in the case of some specific solutes, including anilines and pyridines, the effective solute hydrogen-bond basicity varies with the solvent system. Therefore, the descriptor $\sum \beta_2^H$ is preferably used for partition between water and rather nonaqueous solvent systems, while an alternative $\sum \beta_2^0$ can be used for partition between water and aqueous solvent systems [Abraham and Rafols, 1995].

Table L3 LSER parameter values for some solutes. Symbols defined in the text and data taken from [Kamlet, Doherty *et al.*, 1987a; Abraham, Andonian-Haftvan *et al.*, 1994].

Solute	$\bar{V}/100$	V_x	$\log L^{16}$	R_2	$\sum \pi_2^H$	α_m	β_m	$\sum \alpha_2^H$	$\sum \beta_2^H$
Diethyl ether	1.046	0.7309	2.015	0.041	0.25	0.00	0.47	0.00	0.45
Di- <i>n</i> -butyl ether	1.693	1.2945	3.924	0.000	0.25	0.00	0.46	0.00	0.45
1-Propanol	0.757	0.5900	2.031	0.236	0.42	0.33	0.45	0.37	0.48
2-Propanol	0.765	0.5900	1.764	0.212	0.36	0.33	0.51	0.33	0.56
1-Butanol	0.915	0.7309	2.601	0.224	0.42	0.33	0.45	0.37	0.48
2-Butanol	0.917	0.7309	2.338	0.217	0.36	0.33	0.51	0.33	0.56
1-Pentanol	1.086	0.8718	3.106	0.219	0.42	0.33	0.45	0.37	0.48
Cyclopentanol	1.009	0.7630	3.241	0.427	0.54	0.33	0.51	0.32	0.56
Trichloroethene	0.897	0.7146	2.997	0.524	0.40	0.00	0.10	0.08	0.03
1,1,1-Trichloroethane	0.989	0.7576	2.733	0.369	0.41	0.00	0.10	0.00	0.09
Tetrachloromethane	0.968	0.7391	2.823	0.458	0.38	0.00	0.10	0.00	0.00
1,2-Dichloroethane	0.787	0.6352	2.573	0.416	0.64	0.00	0.10	0.10	0.11
Butanone	0.895	0.6879	2.287	0.166	0.70	0.00	0.48	0.00	0.51
Cyclopentanone	0.986	0.7202	3.221	0.373	0.86	0.00	0.52	0.00	0.52
Cyclohexanone	1.136	0.8611	3.792	0.403	0.86	0.00	0.53	0.00	0.56
Acetonitrile	0.521	0.4042	1.739	0.237	0.90	0.15	0.35	0.04	0.33
Benzene	0.989	0.7164	2.786	0.610	0.52	0.00	0.10	0.00	0.14
Phenol	0.989	0.7751	3.766	0.805	0.89	0.61	0.33	0.60	0.31
<i>m</i> -Cresol	1.163	0.9160	4.329	0.822	0.88	0.55	0.33	0.57	0.34
Nitrobenzene	1.127	0.8906	4.557	0.871	1.11	0.00	0.30	0.00	0.28
2-Nitrotoluene	1.217	1.0315	4.878	0.866	1.11	0.00	0.30	0.00	0.28
Benzonitrile	1.120	0.8711	4.039	0.742	1.11	0.00	0.35	0.00	0.33
Tetrahydrofuran	0.911	0.6223	2.636	0.289	0.52	0.55	0.00	0.00	0.48

A **modified LSER** version, called **MLSER**, is based on $\rightarrow$ *quantum-chemical descriptors* and defined as [Wang, Zhai *et al.*, 2005]

$$P_i = b_0 + b_1 \cdot \alpha + b_2 \cdot \mu + b_3 \cdot \varepsilon_{\text{HOMO}} + b_4 \cdot q^- + b_5 \cdot \varepsilon_{\text{LUMO}} + b_6 \cdot q^+$$

where α is the $\rightarrow$ *molecular polarizability*, μ the $\rightarrow$ *dipole moment*, $\varepsilon_{\text{HOMO}}$ and $\varepsilon_{\text{LUMO}}$ the energies of the highest occupied and lowest unoccupied orbitals, respectively, and q^- and q^+ the negative and positive charges, respectively.

 Additional references are collected in the thematic bibliography (see Introduction).

- **linear subfragment descriptors** $\rightarrow$ substructure descriptors
- **line graph** $\rightarrow$ graph
- **line graph connectivity indices** $\rightarrow$ iterated line graph sequence
- **line graph Randić connectivity index** $\rightarrow$ iterated line graph sequence
- **link** $\equiv$ *connection* $\rightarrow$ edge adjacency matrix
- **linking number** $\rightarrow$ polymer descriptors
- **Lipinski drug-like index** $\rightarrow$ property filters ($\odot$ drug-like indices)
- **lipole** $\rightarrow$ lipophilicity descriptors
- **lipophilicity** $\rightarrow$ lipophilicity descriptors

■ lipophilicity descriptors

Lipophilicity, denoted as P , is the measure of the partitioning of a compound between a lipidic and an aqueous phase that depends on solute bulk, polar, and hydrogen-bonding effects [Taylor, 1990].

Compounds for which $P > 1$ or $\log P > 0$ are *lipophilic*, and compounds for which $P < 1$ or $\log P < 0$ are *hydrophilic*.

The most widely used molecular descriptor of lipophilicity is the $\rightarrow$ *octanol–water partition coefficient* K_{ow} (**log K_{ow}** , or also **log P** when no further specifications are given) that is the partition coefficient between 1-octanol and water:

$$\log P \equiv \log K_{ow} = \log \frac{[C]_{1\text{-octanol}}}{[C]_{\text{water}}} = \log [C]_{1\text{-octanol}} - \log [C]_{\text{water}}$$

Other $\rightarrow$ *partition coefficients* related to lipophilicity are defined for n -alkane systems, such as the partition coefficient between n -heptane–water.

Lipophilicity can be factorized into two main terms as [Carrupt, Testa *et al.*, 1997]

$$\text{lipophilicity} = \text{hydrophobicity} - \text{polarity}$$

where **hydrophobicity** refers to nonpolar interactions (such as dispersion forces, hydrophobic interactions, etc.) of the solute with organic and aqueous phases and *polarity* to polar interactions (such as ion–dipole interactions, hydrogen-bonds, induction and orientation forces, etc.). As it can be observed, in this case, the term hydrophobicity is not synonymous with lipophilicity, but is a component of it.

Usually, hydrophobicity is encoded by $\rightarrow$ *steric descriptors* such as molar or $\rightarrow$ *molecular volume*, which account satisfactorily for nonpolar interactions; polarity can be described by polar terms that are negatively related to lipophilicity. An important factorization of lipophilicity is provided by the $\rightarrow$ *solvatochromic parameters*. Moreover, a measure of the global polarity of a given solute was proposed by Testa and coworkers [Testa and Seiler, 1981; El Tayar and Testa, 1993; Vallat, Gaillard *et al.*, 1995] and called **interactive polar parameter Λ** (or **Testa lipophobic constant**). It is defined as the difference between the experimental lipophilicity measure and that estimated for an hypothetical n -alkane of the same molecular volume V as

$$\Lambda = (\log P)_{\text{exp}} - (\log P)_{\text{calc}}$$

where the calculated $\log P$ is obtained from the following equation:

$$(\log P)_{\text{calc}} = 0.0309 \cdot V + 0.346$$

The interactive parameter Λ should by definition encode the same information as the solvatochromic polar parameters; it takes negative values for lipophobic fragments [van de Waterbeemd and Testa, 1987; El Tayar, Testa *et al.*, 1992, 1993].

For apolar compounds, an analogous linear relationship should be expected between $\log P$ and $\rightarrow$ *van der Waals volume* V^{vdw} that accounts for steric contributions to $\log P$ [Moriguchi, 1975; Moriguchi, Kanada *et al.*, 1976, 1977]. An improved relation is obtained by incorporating into V^{vdw} a correction accounting for $\rightarrow$ *molecular branching*. Moreover, to extend the relation to polar compounds, a correction factor called **Moriguchi polar parameter** (or **hydrophilic effect**) V_H was proposed as in the following equation:

$$\log P = 2.71 \times (V^{vdw} - V_H) + 0.12$$

Thus, the hydrophilic effect V_H is calculated as

$$V_H = V^{vdw} - \frac{\log P - 0.12}{2.71}$$

and it accounts for intramolecular hydrophobic bonding; moreover, it was found to well correlate with the interactive polar parameter Λ .

There are several methods developed for the calculation of $\log P$ from molecular structure, based on $\rightarrow$ *substituent constants*, $\rightarrow$ *fragmental constants*, $\rightarrow$ *electronic descriptors*, $\rightarrow$ *steric descriptors*, $\rightarrow$ *connectivity indices*, $\rightarrow$ *surface areas*, $\rightarrow$ *volume descriptors*, $\rightarrow$ *chromatographic descriptors*.

Important reviews and books about lipophilicity are: [Leo, 1990; Hansch, Leo *et al.*, 1995; Carrupt, Testa *et al.*, 1997; Reinhard and Drefahl, 1999; Testa, Crivori *et al.*, 2000; Mannhold, 2003; Caron and Ermondi, 2008; Mannhold and Ostermann, 2008; Pliška, Testa *et al.*, 2008; van de Waterbeemd and Mannhold, 2008].

The most popular approaches to $\log P$ calculation are listed below.

• **Hansch–Fujita hydrophobic substituent constants** ($\equiv$ *hydrophobic substituent constants*, π)

The lipophilicity is calculated by analogy with the $\rightarrow$ *Hammett equation* as

$$\log \frac{P_X}{P_H} = \rho \cdot \pi_X$$

where P_X and P_H are the partition coefficients of a X-substituted and unsubstituted compounds, respectively; π_X is the hydrophobic constant of the substituent X; the ρ constant reflects the characteristics of the solvent system and it is assumed equal to one for octanol/water solvent system [Fujita, Iwasa *et al.*, 1964].

These hydrophobic substituent constants are commonly used in $\rightarrow$ *Hansch analysis* to encode the lipophilic behavior of the substituents; the lipophilicity of the whole molecule is obtained by adding to the lipophilicity of the unsubstituted parent compound ($\log P_H$) the lipophilic contributions of the substituents:

$$\log P(\text{molecule}) = \log P_H + \sum_{s=1}^S \pi_{X_s}$$

where S is the number of substitution sites and π_{X_s} are the hydrophobic constants of the substituents in the molecule. Distinct values of the π constants were defined for aromatic and aliphatic compounds. The Hansch–Fujita hydrophobic constants are still widely used in QSAR studies, but not for calculating $\log P$ values.

The major drawback of this approach is that π values depend on their electronic environment. When electronic interactions of the substituent X with other substituents in the compound are possible, more realistic π values have to be used. In particular, π^- lipophilic constant, also known as **Norrington lipophilic constant** [Norrington, Hyde *et al.*, 1975] measures the lipophilic contribution of strong electron-releasing groups such as $-\text{OH}$, $-\text{NH}_2$, $-\text{NHR}$, or $-\text{NR}_1\text{R}_2$ when they are attached to a conjugated system (usually phenol or aniline); π^+ lipophilic constant measures the lipophilic contribution of strong electron-attracting groups such as cyano or nitro groups, conjugated with the functional group. The use of this last constant is very rare.

Based on the decomposition of $\log P$ into enthalpic P_h and entropic P_s contributions,

$$\log P = \frac{-\Delta G_p^0}{2.303 \cdot RT} = \frac{-\Delta H_p^0}{2.303 \cdot RT} + \frac{\Delta S_p^0}{2.303 \cdot R} = P_h + P_s$$

the hydrophobic substituent constant π was decomposed [Da, Ito *et al.*, 1992; Da, Yanagi *et al.*, 1993] into the **enthalpic hydrophobic substituent constant** π_h and **entropic hydrophobic substituent constant** π_s , respectively:

$$\pi = \pi_h + \pi_s$$

$$\pi_h = (P_h)_X - (P_h)_H \quad P_h = \frac{-\Delta H_p^0}{2.303 \cdot RT}$$

$$\pi_s = (P_s)_X - (P_s)_H \quad P_s = \frac{+\Delta S_p^0}{2.303 \cdot R}$$

where ΔG_p^0 , ΔH_p^0 , and ΔS_p^0 are the Gibbs free energy, enthalpy, and entropy of transfer for partition, respectively; R is the gas constant and T the absolute temperature; the subscripts X and H denote the substituted and unsubstituted compound, respectively. The enthalpic contribution P_h can be interpreted as a new hydrophobic parameter that reflects the heat evolved when a solute is transferred from water to nonaqueous phase. Similarly, the entropic contribution P_s can be interpreted to reflect the change of randomness induced in the solution when a solute is transferred from water to nonaqueous phase.

📖 [Hansch, Maloney *et al.*, 1962; Hansch, Muir *et al.*, 1963; Hansch and Anderson, 1967; Martin and Lynn, 1971; Hansch, Leo *et al.*, 1971, 1972, 1973; Hansch and Dunn III, 1972; Fujita and Nishioka, 1976; Fujita, 1983; Leo, 1993; Gago, Pastor *et al.*, 1994; Amić, Davidović-Amić *et al.*, 1998]

• Nys-Rekker hydrophobic fragmental constants (f)

Also simply called **hydrophobic fragmental constants**, they are measures of the absolute lipophilicity contribution of specific molecular fragments to the lipophilicity of the molecule [Nys and Rekker, 1973, 1974; Rekker, 1977a, 1977b; Rekker and De Kort, 1979].

The $\log P$ of a molecule is calculated by summing up the fragmental contributions and applying the appropriate correction factors as

$$\log P = b_0 + \sum_i f_i \cdot N_i + \sum_j c_j \cdot N_j$$

where f_i and N_i are the hydrophobic constant and the number of occurrences of i th fragment in the considered compound, N_j is the number of occurrences of the j th correction factor. c_j is the value of the considered correction factor describing some special structural features (proximity effects, hydrogen atoms attached to polar groups, aryl-aryl conjugation, etc.); in practice, it can be calculated as

$$c_j = k_j \cdot 0.219$$

where 0.219 is the so-called “magic constant” and k_j is an integer value characterizing the j th correction factor.

Different sets of fragmental constants were derived by multiple regression analysis for fragments depending on their attachment to an aliphatic or aromatic carbon atom. In this approach, the effects of intramolecular interactions on lipophilicity are taken into account by the correction factors and implicitly in the definition of the molecular fragments. However, group interactions are evaluated more by the topological distance between the groups rather than by the chemical nature of the groups and their geometry.

A drawback of this approach is that, for the same compound, different selections of fragments give different $\log P$ values. Moreover, for complex compounds, the decomposition of the molecular structure into appropriate fragments is not unique and is a difficult task.

Calculation of $\log P$ based on revised Nys–Rekker fragmental constants is provided by some software, such as PROLOGP and SANALOGP.

📖 [Mayer, van de Waterbeemd *et al.*, 1982; Takeuchi, Kuroda *et al.*, 1990; Rekker, 1992; Rekker and Mannhold, 1992; Rekker and de Vries, 1993; Mannhold, Rekker *et al.*, 1998; Rekker, Mannhold *et al.*, 1998]

• Leo-Hansch hydrophobic fragmental constants (f')

These are hydrophobic constants calculated for molecular fragments by a “constructionist approach” that consists of determining very accurately the $\log P$ values of simple compounds usually having a single functional group and then calculating fundamental hydrophobic fragmental constants from these values [Hansch and Leo, 1979; Leo, 1987, 1993]. $\log P$ of compound is calculated by using the Rekker’s additive scheme, based on different fragmental constants f'' and correction factors F' . The decomposition of the molecular structure into fragments is performed by using a unique and simple set of rules, thus obtaining a unique solution; the fragments are either atoms or polyatomic groups. Correction factors were derived from compounds with more than one substituent to better approximate experimental $\log P$ values. They take into account proximity effects due to multiple halogenation and groups giving hydrogen-bonds, intramolecular hydrogen-bonds involving oxygen and nitrogen atoms, electronic effects in aromatic systems, unsaturation, branching, chains, rings. Over 200 fragmental constants and 14 correction factors have been determined.

The software version of the Leo-Hansch fragmental method is known as **CLOGP** or **Calculated LOGP** [Chou and Jurs, 1979].

📖 [Leo and Hansch, 1971; Leo, Jow *et al.*, 1975; Lyman, Reehl *et al.*, 1982; Mayer, van de Waterbeemd *et al.*, 1982; Abraham and Leo, 1987; Leo, 1991, 1993]

• Klopman hydrophobic models

The first model for the prediction of $\log P$ proposed by Klopman and Iroff [Klopman and Iroff, 1981] was based on the assumption that partition coefficients of molecules depend on the charge densities on each atom of the molecule. The following equation was proposed including both atom and group counting descriptors and charge density descriptors:

$$\log P = b_0 + \sum_i b_i \cdot N_i + \sum_i b'_i \cdot q_i + \sum_i b''_i \cdot q_i^2 + \sum_j c_j \cdot N_j$$

where the first three summations run over the different types of atoms, b are the estimated regression coefficients, N_i the number of occurrences of the i th atom-type, q_i the charge density

on the i th atom-type; N_j are the occurrences or the presence/absence of some specific functional groups (acid/ester, nitrile, amide groups) whose influence on molecule lipophilicity is described by selected correction factors c_j .

Another model was proposed based on atomic composition of the molecule only, ignoring the influence of the charge density descriptors on the log P calculation [Klopman, Namboodiri *et al.*, 1985]. The best-fitted proposed model is

$$\log P = -0.206 + 0.332 \cdot N_C + 0.071 \cdot N_H - 0.860 \cdot N_O - 1.124 \cdot N_N + 0.688 \cdot N_{Cl} \\ + 0.981 \cdot (N_{ac} + N_{est}) - 0.138 \cdot N_{Ph} + 2.969 \cdot N_{NO_2} + 1.053 \cdot I_{aliph}$$

$$n = 195; \quad r^2 = 0.949; \quad s = 0.293; \quad F = 33$$

where N_C , N_H , N_O , N_N , and N_{Cl} are the numbers of carbon, hydrogen, oxygen, nitrogen, and chlorine atoms, respectively; N_{ac} and N_{est} are the numbers of acid and ester groups, respectively; N_{Ph} , and N_{NO_2} are the numbers of phenyl rings and nitrile groups; I_{aliph} is an indicator variable for aliphatic hydrocarbons.

The regression coefficients of the model relative to the atomic counting descriptors can be viewed as individual log P contributions of each atom; these are the **Klopman hydrophobic atomic constants**. A better evaluation of these atomic contributions was proposed by classifying the atoms also accounting for their environment represented by the first neighbors [Klopman and Wang, 1991]. Moreover, this method uses for the evaluation of log P only those atom-centered groups that are identified by stepwise regression as the most significant groups determining log P .

A further developed model (**KLOGP** or **Klopman LOGP**), was proposed as

$$\log P = b_0 + \sum_i f_i \cdot N_i + \sum_j c_j \cdot N_j$$

where N_i is the number of occurrences of the i th atom-centered fragment in the molecule and N_j are the number of occurrences of particular fragments accounting for the interactions between groups whose influence on molecule lipophilicity is described by calculated correction factors c_j [Klopman, Li *et al.*, 1994]. Basic atom-centered groups are of two types: (a) atomic groups defined by their chemical element, hybridization state, and the number of attached hydrogen atoms; (b) functional groups containing at least one heteroatom. Correction factors are supplementary hydrophobic constants relative to specific substructures with more than two nonhydrogen atoms.

The regression coefficients b_i are the Klopman hydrophobic atomic constants measuring the hydrophobic contributions of atom-types in the same way as the hydrophobic fragmental constants f_i defined in Rekker and Leo–Hansch approaches. The best evaluation of 64 atomic constants plus 30 correction factors was obtained by a training set of 1663 compounds, $r^2 = 0.93$, $s = 0.38$, $F = 218$

The automated recognition of fragments and correction factors is performed by **CASE approach** (*Computer Automated Structure Evaluation*). Basically, CASE is an artificial intelligence system capable of identifying structural fragments that may be associated with the properties of the training molecules, such as biological activity and physico-chemical properties [Klopman, 1984]. **MULTICASE** (or **MCASE**) is the most recent upgraded software version [Klopman, 1992, 1998; Saiakhov, Stefan *et al.*, 2000; Klopman, Zhu *et al.*, 2003].

• Suzuki–Kudo hydrophobic fragmental constants

The contribution method of Suzuki and Kudo [Suzuki and Kudo, 1990; Suzuki, 1991] is based on hydrophobic fragmental constants f_i and it is defined as

$$\log P = b_0 + \sum_i f_i \cdot N_i$$

where N_i is the number of occurrences of the i th fragment in the molecule. A first set of 415 basic hydrophobic constants was proposed, representing the lipophilic contributions of groups, each described by its structural environment. Several basic groups were first defined as CH_3 , CH_2 , CO , SO_2 , and so on, which were further distinguished according to their neighboring atoms with their connectivities. Groups of atoms such as cyano and nitro are considered as univalent heteroatoms. In addition, extended fragments based on the basic fragments plus some other functional groups were selected together with some user-defined fragments. A training set of 1465 compounds plus a test set of 221 compounds were used to evaluate the hydrophobic constants by multivariate regression analysis. The software version of Suzuki–Kudo method is **CHEMICALC** (*Combined Handling of Estimation Methods Intended for Completely Automated Log P Calculation*).

• Broto–Moreau–Vandycke hydrophobic atomic constants

The Broto–Moreau–Vandycke contribution method is based on hydrophobic atomic constants a_k measuring the lipophilic contributions of atoms, each described by its nature, neighboring atoms and associated connectivities, thus implicitly considering some proximity effects and interactions in conjugated systems [Broto, Moreau *et al.*, 1984b]. Hydrogen atoms and correction factors are not explicitly considered. The model is defined as

$$\log P = b_0 + \sum_i a_i \cdot N_i$$

where N_i is the number of occurrences of the i th atom-type in the molecule. The atomic constants a_i were estimated by multivariate regression analysis.

Atom-types are classified according to their structural environment; carbon atom-types are differentiated by their bonds to nonhydrogen atoms; heteroatoms are differentiated by their bonds to nonhydrogen first neighbors and the nature of the neighbors, moreover if the neighbor atom is a carbon atom its bond environment is also accounted for. A conjugation contribution is considered as correction factor for sp^2 carbon atoms in conjugated systems.

Using a training set of 1868 compounds, a set of 222 atomic constants was proposed and the best-fitted model gave a standard error about 0.4 log unit.

The software program **SMILOGP** for the calculation of $\log P$ is based on the Broto–Moreau–Vandycke hydrophobic constants and the SMILES notation for the recognition of molecule atom-centered fragments [Convard, Dubost *et al.*, 1994].

• Ghose–Crippen hydrophobic atomic constants

These are hydrophobic atomic constants a_k measuring the lipophilic contribution of atoms in the molecule, each described by its neighboring atoms [Ghose and Crippen, 1986; Ghose, Pritchett *et al.*, 1988; Viswanadhan, Ghose *et al.*, 1989]. The model for $\log P$ calculation is defined as

$$\log P = \sum_k a_k \cdot N_k$$

where N_k is the number of occurrences of the k th atom-type.

A set of 120 $\rightarrow$ atom-centered fragment descriptors that are the **Ghose–Crippen descriptors**, was proposed (Table L4). Atom-centered fragments were defined for hydrogen atoms, carbon atoms, and heteroatoms. Hydrogen and halogen atoms are classified by the hybridization and oxidation state of the carbon atom to which they are bonded; for hydrogens, heteroatoms attached to a carbon atom in α -position are further considered. Carbon atoms are classified by their hybridization state and depending on whether their neighbors are carbon or heteroatoms.

The corresponding hydrophobic constants were evaluated by multivariate regression analysis using a training set of 8364 compounds, $r^2 = 0.95$, $Q^2 = 0.90$ and $RMSE = 0.555$ [Ghose, Viswanadhan *et al.*, 1998]. The $\log P$ estimated by Ghose–Crippen method is actually called **ALOGP** [Viswanadhan, Reddy *et al.*, 1993]. As in the Broto–Moreau–Vanduycke approach, correction factors are avoided, while hydrogen atoms are instead considered.

Calculation of ALOGP is provided by a number of software programs, such as DRAGON, MOLCAD, PROLOGP, and TSAR.

A smaller set of atom-centered fragments was later proposed to avoid ambiguity sometimes occurring in atom-type assignment. It is comprehensive for the common elements in organic molecules, and also includes metals and noble gases [Wildman and Crippen, 1999].

Ghose–Crippen descriptors were successfully used also to model $\rightarrow$ molar refractivity [Ghose and Crippen, 1987] and solvation free energies [Viswanadhan, Ghose *et al.*, 1999] by $\rightarrow$ group contribution methods.

Table L4 Ghose–Crippen atomic contributions to $\log P$ and molar refractivity (MR).

ID	Description	$\log P$	MR	ID	Description	$\log P$	MR
C-001	CH ₃ R/CH ₄	-1.5603	2.968	O-061	O.. ^a	1.052	1.945
C-002	CH ₂ R ₂	-1.012	2.9116	O-062	O ⁻ (negatively charged)	-0.7941	—
C-003	CHR ₃	-0.6681	2.8028	O-063	R—O—O—R	0.4165	—
C-004	CR ₄	-0.3698	2.6205	Se-064	Any—Se—Any	0.6601	—
C-005	CH ₃ X	-1.788	3.015	Se-065	=Se	—	—
C-006	CH ₂ RX	-1.2486	2.9244	N-066	Al—NH ₂	-0.5427	2.6221
C-007	CH ₂ X ₂	-1.0305	2.6329	N-067	Al ₂ —NH	-0.3168	2.5
C-008	CHR ₂ X	-0.6805	2.504	N-068	Al ₃ —N	0.0132	2.898
C-009	CHRX ₂	-0.3858	2.377	N-069	Ar—NH ₂ /X—NH ₂	-0.3883	3.6841
C-010	CHX ₃	0.7555	2.559	N-070	Ar—NH—Al	-0.0389	4.2808
C-011	CR ₃ X	-0.2849	2.303	N-071	Ar—NAl ₂	0.1087	3.6189
C-012	CR ₂ X ₂	0.02	2.3006	N-072	RCO—N</>N—X=X	-0.5113	2.5
C-013	CRX ₃	0.7894	2.9627	N-073	Ar ₂ NH/Ar ₃ N/ Ar ₂ N—Al/R.. ^b	0.1259	2.7956
C-014	CX ₄	1.6422	2.3038	N-074	R#N/R=N—	0.1349	2.7
C-015	=CH ₂	-0.7866	3.2001	N-075	R—N—R ^c /R—N—X	-0.1624	4.2063
C-016	=CHR	-0.3962	4.2654	N-076	Ar—NO ₂ /R—N(—R)— O ^d /RO—NO	-2.0585	4.0184
C-017	=CR ₂	0.0383	3.9392	N-077	Al—NO ₂	-1.915	3.0009
C-018	=CHX	-0.8051	3.6005	N-078	Ar—N=X/X—N=X	0.4208	4.7142
C-019	=CRX	-0.2129	4.487	N-079	N ⁺ (positively charged)	-1.4439	—

(Continued)

Table L4 (Continued)

ID	Description	log P	MR	ID	Description	log P	MR
C-020	=CX ₂	0.2432	3.2001	U-080	Undefined	—	—
C-021	#CH	0.4697	3.4825	F-081	F ^e attached to C ¹ (sp ³)	0.4797	0.8725
C-022	#CR/R=C=R	0.2952	4.2817	F-082	F ^e attached to C ² (sp ³)	0.2358	1.1837
C-023	#CX	—	3.9556	F-083	F ^e attached to C ³ (sp ³)	0.1029	1.1573
C-024	R--CH--R	-0.3251	3.4491	F-084	F ^e attached to C ¹ (sp ²)	0.3566	0.8001
C-025	R--CR--R	0.1492	3.8821	F-085	F ^e attached to C ² (sp ²)-C ⁴ (sp ²)/ C ¹ (sp)/C ⁴ (sp ³)/X	0.1988	1.5013
C-026	R--CX--R	0.1539	3.7593	Cl-086	Cl ^e attached to C ¹ (sp ³)	0.7443	5.6156
C-027	R--CH--X	0.0005	2.5009	Cl-087	Cl ^e attached to C ² (sp ³)	0.5337	6.1022
C-028	R--CR--X	0.2361	2.5	Cl-088	Cl ^e attached to C ³ (sp ³)	0.2996	5.9921
C-029	R--CX--X	0.3514	3.0627	Cl-089	Cl ^e attached to C ¹ (sp ²)	0.8155	5.3885
C-030	X--CH--X	0.1814	2.5009	Cl-090	Cl ^e attached to C ² (sp ²)-C ⁴ (sp ²)/ C ¹ (sp)/C ⁴ (sp ³)/X	0.4856	6.1363
C-031	X--CR--X	0.0901	—	Br-091	Br ^e attached to C ¹ (sp ³)	0.8888	8.5991
C-032	X--CX--X	0.5142	2.6632	Br-092	Br ^e attached to C ² (sp ³)	0.7452	8.9188
C-033	R--CH..X	-0.3723	3.4671	Br-093	Br ^e attached to C ³ (sp ³)	0.5034	8.8006
C-034	R--CR..X	0.2813	3.6842	Br-094	Br ^e attached to C ¹ (sp ²)	0.8995	8.2065
C-035	R--CX..X	0.1191	2.9372	Br-095	Br ^e attached to C ² (sp ²)-C ⁴ (sp ²)/ C ¹ (sp)/C ⁴ (sp ³)/X	0.5946	8.7352
C-036	Al--CH=X	-0.132	4.019	I-096	I ^e attached to C ¹ (sp ³)	1.4201	13.9462
C-037	Ar--CH=X	-0.0244	4.777	I-097	I ^e attached to C ² (sp ³)	1.1472	14.0792
C-038	Al--C(=X)--Al	-0.2405	3.9031	I-098	I ^e attached to C ³ (sp ³)	—	14.073
C-039	Ar--C(=X)--R	-0.0909	3.9964	I-099	I ^e attached to C ¹ (sp ²)	0.7293	12.9918
C-040	R--C(=X)--X/ R--C#X/X=C=X	-0.1002	3.4986	I-100	I ^e attached to C ² (sp ²)-C ⁴ (sp ²)/ C ¹ (sp)/C ⁴ (sp ³)/X	0.7173	13.3408
C-041	X--C(=X)--X	0.4182	3.4997	F-101	Fluoride ion	—	—
C-042	X--CH..X	-0.2147	2.7784	Cl-102	Chloride ion	-2.6737	—
C-043	X--CR..X	-0.0009	2.6267	Br-103	Bromide ion	-2.4178	—
C-044	X--CX..X	0.1388	2.5	I-104	Iodide ion	-3.1121	—
U-045	Undefined	—	—	U-105	Undefined	—	—
H-046	H ^e attached to C ⁰ (sp ³) no X attached to next C	0.7341	0.8447	S-106	R--SH	0.6146	7.8916
H-047	H ^e attached to C ¹ (sp ³)/C ⁰ (sp ²)	0.6301	0.8939	S-107	R ₂ S/RS--SR	0.5906	7.7935
H-048	H ^e attached to C ² (sp ³)/C ¹ (sp ²)/ C ⁰ (sp)	0.518	0.8005	S-108	R=S	0.8758	9.4338
H-049	H ^e attached to C ³ (sp ³)/C ² (sp ²)/ C ³ (sp ²)/C ³ (sp)	-0.0371	0.832	S-109	R--SO--R	-0.4979	7.7223
H-050	H attached to heteroatom	-0.1036	0.8	S-110	R--SO ₂ --R	-0.3786	5.7558

(Continued)

Table L4 (Continued)

ID	Description	log P	MR	ID	Description	log P	MR
H-051	H attached to alpha-C ^f	0.5234	0.8188	Si-111	>Si<	1.5188	—
H-052	H ^e attached to C ⁰ (sp ³) with 1X attached to next C	0.6666	0.9215	B-112	>B— as in boranes	1.0255	—
H-053	H ^e attached to C ⁰ (sp ³) with 2X attached to next C	0.5372	0.9769	U-113	Undefined	—	—
H-054	H ^e attached to C ⁰ (sp ³) with 3X attached to next C	0.6338	0.7701	U-114	Undefined	—	—
H-055	H ^e attached to C ⁰ (sp ³) with 4X attached to next C	0.362	—	P-115	P ylides	—	—
O-056	Alcohol	−0.3567	1.7646	P-116	R ₃ −P=X	−0.9359	5.5306
O-057	Phenol/enol/carboxyl OH	−0.0127	1.4778	P-117	X ₃ −P=X (phosphate)	−0.1726	5.5152
O-058	=O	−0.0233	1.4429	P-118	PX ₃ (phosphite)	−0.7966	6.836
O-059	Al−O−Al	−0.1541	1.6191	P-119	PR ₃ (phosphine)	0.6705	10.0101
O-060	Al−O−Ar/ Ar−O−Ar/R..O..R/ R−O−C=X	0.0324	1.3502	P-120	C−P(X) ₂ =X (phosphonate)	−0.4801	5.2806

R: any group linked through carbon; X: any electronegative atom (O, N, S, P, Se, halogens); Al and Ar: aliphatic and aromatic groups, respectively; =: a double bond; #: a triple bond; -: an aromatic bond as in benzene or delocalized bonds such as the N−O bond in a nitro group; ..: aromatic single bonds as the C−N bond in pyrrole. Data from [Ghose, Viswanadhan *et al.*, 1998].

^aAs in nitro, N-oxides.

^bPyrrole-type structure.

^cPyridine-type structure.

^dPyridine N-oxide type structure.

^eThe superscript represents the formal oxidation number. The formal oxidation number of a carbon atom equals the sum of the conventional bond orders with electronegative atoms; the C−N bond order in pyridine may be considered as 2 while we have one such bond and 1.5 when we have two such bonds; the C..X bond order in pyrrole or furan may be considered as 1.

^fAn alpha-C may be defined as a C attached through a single bond with −C=X, −C#X, −C−X.

• MLOGP ($\equiv$ Moriguchi model based on structural parameters)

This is a model described by a regression equation based on 13 structural parameters and defined as [Moriguchi, Hirono *et al.*, 1992b, 1994]

$$\log P = -1.014 + 1.244 \cdot (F_{CX})^{0.6} - 1.017 \cdot (N_O + N_N)^{0.9} + 0.406 \cdot F_{PRX} - 0.145 \cdot N_{UNS}^{0.8} \\ + 0.511 \cdot I_{HB} + 0.268 \cdot N_{POL} - 2.215 \cdot F_{AMP} + 0.912 \cdot I_{ALK} - 0.392 \cdot I_{RNG} - 3.684 \cdot F_{QN} \\ + 0.474 \cdot N_{NO_2} + 1.582 \cdot F_{NCS} + 0.773 \cdot I_{BL}$$

$$n = 1230; \quad r^2 = 0.91; \quad s = 0.411; \quad F = 900.4$$

The meaning of the structural parameters and corresponding regression coefficients are reported in Table L5.

Table L5 Regression coefficients of the Moriguchi model.

Symbol	Descriptor	b_i
b_0	Intercept	-1.014
F_{CX}	Summation of number of carbon and halogen atoms weighted by C = 1.0; F = 0.5; Cl = 1.0; Br = 1.5; I = 2.0	1.244
$N_O + N_N$	Total number of nitrogen and oxygen atoms	-1.017
F_{PRX}	Proximity effect of N/O: X - Y = 2; X - A - Y = 1 (X, Y: N/O; A: C, S, or P) with correction -1 for carbox-amide/sulfonamide	0.406
N_{UNS}	Total number of unsaturated bonds (not those in NO ₂)	-0.145
I_{HB}	Dummy variable for the presence of intramolecular H-bonds	0.511
N_{POL}	Number of polar substituents	0.268
F_{AMP}	Amphoteric property: α -amino = 1.0; aminobenzoic acid or pyridinecarboxylic acid = 0.5	-2.215
I_{ALK}	Dummy variable for alkane, alkene, cycloalkane, cycloalkene (hydrocarbons with 0 or 1 double bond)	0.912
I_{RNG}	Dummy variable for the presence of ring structures (not benzene and its condensed rings)	-0.392
F_{QN}	Quaternary nitrogen = 1.0; N-oxide = 0.5	-3.684
N_{NO_2}	Number of nitro groups	0.474
F_{NCS}	-N=C=S group = 1.0; -S-CN group = 0.5	1.582
$I_{\beta L}$	Dummy variable for the presence of β -lactam	0.773

• Moriguchi model based on surface area

This is a model for predicting lipophilicity of compounds based on the $\rightarrow$ *solvent-accessible surface area* SASA generated by a solvent probe of 1.4 Å radius and a set of parameters encoding hydrophilic effects of polar groups [Iwase, Komatsu *et al.*, 1985]:

$$\log P = -1.06 + 1.90 \cdot SASA - 1.00 \cdot \sum_k S_{H_k}$$

$$n = 138; \quad r^2 = 0.99; \quad s = 0.13; \quad F = 7284$$

where S_H are measures of the surface area of polar groups contributing negatively to $\log P$ of the compounds. The latter parameters can be considered as fragmental correction factors whose values are derived separately for polar groups in aliphatic and aromatic systems.

• Dunn model based on surface area

This is a model for predicting $\log P$ values in different solvent systems [Dunn III, Koehler *et al.*, 1987; Koehler, Grigoras *et al.*, 1988], defined by the equation

$$\log P_{solv} = a_{solv} \cdot ISA - b_{solv} \cdot f(HSA)$$

where ISA is the $\rightarrow$ *isotropic surface area*, related to the solute surface accessible to nonspecific solvent interactions, and HSA the solvent-accessible $\rightarrow$ *hydrated surface area*, associated with

hydration of polar functional groups. $f(HSA)$ is the *hydrated fraction surface area*, that is, $HSA/SASA$, encoding the polar component of the lipophilicity as the S_H parameter in the $\rightarrow$ Moriguchi model based on surface area.

• Camilleri model based on surface area

This is a model based on the factorization of the solvent-accessible surface area into 12 contributions relative to 12 molecular fragments [Camilleri, Watts *et al.*, 1988]:

$$\log P = b_0 + \sum_{i=1}^{12} b_i \cdot N_i$$

where b are the regression coefficients associated with the surface area contributions and N_i the number of occurrences of the i th molecular fragment (Table L6).

Table L6 Regression coefficients of the Camilleri model.

Type	Fragment	b_i
b_0	Intercept	−23.9
b_1	Aromatic hydrocarbon	2.49
b_2	Saturated hydrocarbon chains not A_3 , A_6 , A_{10} , A_{12}	2.731
b_3	Single saturated carbon atom attached to a nonhydrocarbon group plus hydrogens	−2.237
b_4	OH group	−1.809
b_5	Oxygen atom of OR group not A_{11}	−0.042
b_6	Hydrocarbon part of OR group not A_{12}	0.963
b_7	Cl atom	3.634
b_8	NH_2 or NH group	−3.197
b_9	$C(=O)R$ group	−0.712
b_{10}	Hydrocarbon chain part in $C(=O)R$ group	0.697
b_{11}	Oxygen atom of OR group in $C(=O)OR$ group	−8.54
b_{12}	Hydrocarbon part of OR group in $C(=O)OR$ group	3.526

• Politzer hydrophobic model

This is a model obtained by applying the $\rightarrow$ GIPF approach (*General Interaction Properties Function approach*) proposed by Politzer and coworkers [Brinck, Murray *et al.*, 1993; Murray, Brinck *et al.*, 1993, 1994] as general method to estimate $\rightarrow$ physico-chemical properties in terms of $\rightarrow$ molecular electrostatic potential (MEP) properties calculated at the $\rightarrow$ molecular surface.

The Politzer hydrophobic model was proposed as the following:

$$\log P = -0.504 + 0.0300 \cdot SA - 0.00472 \cdot (N_N + 2N_O) \cdot \sigma_-^2 - 0.000963 \cdot SA \cdot \Pi$$

$$n = 70; \quad r^2 = 0.97; \quad s = 0.277$$

where SA is the molecular surface area, N_N and N_O are the numbers of nitrogen and oxygen atoms, respectively, σ_-^2 is the variance of the negative regions of the molecular surface potential, Π is the $\rightarrow$ local polarity index.

Improvements of the Politzer hydrophobic model were later proposed using additional $\rightarrow$ quantum-chemical descriptors derived from the molecular electrostatic potential, dipole moment,

and ionization energies. These descriptors were searched for to give the best estimates of the cavity term, polarity/dipolarizability term, and hydrogen-bond parameters defined in $\rightarrow$ *Linear Solvation Energy Relationships* [Haeberlein and Brinck, 1997].

• **KOWWIN** ($\equiv$ *Meylan–Howard hydrophobic model; AFC method*)

The Meylan–Howard hydrophobic model is derived from an atom/fragment contribution method providing 150 hydrophobic atomic and fragmental constants f_i measuring the lipophilic contributions of atoms and fragments in the molecule, together with 250 correction factors [Meylan and Howard, 1995, 1996, 2000; KOWWIN – Syracuse Research Corporation, 2008].

The model is defined as

$$\log P = 0.229 + \sum_i f_i \cdot N_i + \sum_j c_j \cdot N_j$$

$$n = 2351; \quad r^2 = 0.982; \quad s = 0.216$$

where N_i is the number of occurrences of the i th atom-type or fragment, and N_j is the number of occurrences of the j th correction factor c_j .

The hydrophobic constants f_i have been evaluated by a first linear regression analysis of 1120 compounds, without considering correction factors. The correction factors were then derived from a linear regression on additional 1231 compounds, correlating the differences between experimental $\log P$ and the $\log P$ estimated by the first regression model.

• **VLOGP** ($\equiv$ *Gombar hydrophobic model*)

This is a model for the assessment of $\log P$ based on 363 molecular descriptors derived from the molecular topology and obtained from 6675 diverse chemicals, with $r^2 = 0.986$ and $s = 0.20$ [Gombar and Enslein, 1996; Gombar, 1999]. Among the molecular descriptors considered are $\rightarrow$ *molecular weight*, $\rightarrow$ *electrotopological state indices*, $\rightarrow$ *Kier shape descriptors* of order 1–7, and some $\rightarrow$ *symmetry descriptors*. In particular, several descriptors are defined as the sum of the E-states of the atoms involved in the whole molecule or in predefined molecular fragments.

• **BLOGP** ($\equiv$ *Bodor LOGP, Bodor hydrophobic model*)

This is a nonlinear 18-parameter model based on 10 molecular descriptors calculated by semiempirical quantum-chemistry methods, starting from optimized 3D geometries [Bodor, Gabanyi *et al.*, 1989; Bodor and Huang, 1992a, 1994]:

$$\begin{aligned} \log P = & 9.552 + 0.005286 \cdot \text{MW} + 0.08325 \cdot N_C + 1.0392 \cdot I_{\text{alk}} - 0.05726 \cdot \mu \\ & - 7.6661 \cdot O - 5.5961 \cdot O^2 + 2.1059 \cdot O^4 + 0.05984 \cdot SA - 0.0001141 \cdot SA^2 \\ & - 0.2741 \cdot Q - 8.5144 \cdot Q_N + 31.243 \cdot Q_N^2 - 17.377 \cdot Q_N^4 \\ & - 4.6249 \cdot Q_O + 20.346 \cdot Q_O^2 - 5.4195 \cdot Q_O^4 - 5.0040 \cdot Q_{\text{ON}} \end{aligned}$$

$$n = 302, \quad r^2 = 0.96, \quad s = 0.306, \quad F = 368$$

The BLOGP molecular descriptors are shown in Table L7.

Table L7 Molecular descriptors and regression coefficients of the BLOGP model.

Symbol	Descriptor	b_i
b_0	Intercept	9.5524
MW	Molecular weight	0.005 286
N_C	Number of carbon atoms	0.083 25
I_{alk}	Indicator variable for the presence of alkanes	1.0392
μ	Dipole moment	−0.057 26
O	Ovality index	−7.6661
O^2	Second power of the ovality index	−5.5961
O^4	Fourth power of the ovality index	2.1059
SA	van der Waals surface area	0.059 84
SA^2	Second power of the van der Waals surface area	−0.000 1141
Q	Total absolute atomic charge	−0.2741
Q_N	Square root of the sum of the squared charges of nitrogen atoms	−8.5144
Q_N^2	Second power of Q_N	31.243
Q_N^4	Fourth power of Q_N	−17.377
Q_O	Square root of the sum of the squared charges of oxygen atoms	−4.6249
Q_O^2	Second power of Q_O	20.346
Q_O^4	Second power of Q_O	−5.4195
Q_{ON}	Sum of the absolute charges of oxygen and nitrogen atoms	−5.0040

The model shows high correlation between the independent variables and some nonsignificant regression coefficients.

A model based on the same set of $\rightarrow$ *quantum-chemical descriptors* was also proposed for aqueous solubility by a neural network approach [Bodor, Harget *et al.*, 1991; Bodor, Huang *et al.*, 1992, 1994].

• Kantola–Villar–Loew hydrophobic models

This is a lipophilicity model based on atomic charges, surface areas, dipole moments, and a set of adjustable parameters depending only on the atomic number [Kantola, Villar *et al.*, 1991]. The parameter values are determined to reproduce experimental $\log P$ values using the following general model:

$$\log P = \sum_{i=1}^A [\alpha_i \cdot SA_i + \beta_i \cdot SA_i \cdot q_i^2 + \gamma_i \cdot q_i] + \delta \cdot \mu$$

where SA are atomic contributions to the $\rightarrow$ *solvent-accessible surface area*, q are atomic charges, μ the dipole moment. Setting to zero some of the parameters α , β , γ , and δ , different submodels are obtained. The model descriptors are calculated by $\rightarrow$ *computational chemistry* methods, thus resulting in conformationally dependent hydrophobicity values.

• molecular lipophilicity potential model

This is a $\log P$ model based on the $\rightarrow$ *molecular lipophilicity potential* (MLP) defined as

$$\log P = -0.10 + 0.00286 \cdot \sum MLP^+ + 0.00152 \cdot \sum MLP^-$$

$$n = 114, \quad r^2 = 0.94, \quad s = 0.37, \quad F = 926$$

where descriptors ΣMLP^+ and ΣMLP^- are the sum of the positive and negative MLP values, respectively [Gaillard, Carrupt *et al.*, 1994b]. They represent the hydrophobic and polar contributions of the molecule.

The specific expression for MLP used in this model is the following:

$$MLP_i = \sum_k a_k \cdot \frac{1 + \exp(b \cdot c)}{1 + \exp[b \cdot (r_{ik} - c)]}$$

where MLP_i is the molecular lipophilicity potential at the i th grid point, a_k are the $\rightarrow$ Broto–Moreau–Vanduycke hydrophobic atomic constants, r_{ik} the $\rightarrow$ geometric distance between the k th fragment and the i th grid point; b and c are the two parameters defining the shape of the Fermi-type function used to calculate MLP values ($b = 1.33$ and $c = 3.25$). MLP values are calculated using the $\rightarrow$ solvent-accessible surface area as integration space.

• ACD/log P

This is a model for $\log P$ calculation proposed by Petrauskas and Kolovanov [Petrauskas and Kolovanov, 2000] as a modification of the $\rightarrow$ CLOGP, aimed at reducing the large number of correction factors involved in the CLOGP calculation.

For example, H-atoms are never detached from carbon atoms, as CLOGP is. This automatically enlarges a list of fragmental increments, but eliminates the need for many structural correction factors: chain and ring flexibility, chain and group branching, double and triple unsaturation, aromatic ring conjugation in biphenyls and fusion in naphthalenes, and so on.

Using a training set of 3601 compounds, the best-fitted model gave $r^2 = 0.984$ and $s = 0.21$.

• HLOGP model

The HLOGP model, proposed by Viswanadhan *et al.*, [Viswanadhan, Ghose *et al.*, 2000], uses both smaller atom sized and larger fragments encoded into $\rightarrow$ molecular holograms. The model was obtained by generating holograms of various lengths for each molecule in the training set and performing Partial Least Squares (PLS) analysis, followed by the selection of model features leading the least standard error.

Using a training set of 265 compounds, the best-fitted model, with a hologram length of 257 and using 18 PLS components, gave $r^2 = 0.941$ and $S = 0.58$.

• XLOGP

This is a model for $\log P$ calculation based on a $\rightarrow$ group-contribution method proposed by Wang Renxiao *et al.* [Wang, Fu *et al.*, 1997; Wang, Gao *et al.*, 2000]. The model is defined in terms of the $\rightarrow$ solvent accessible surface area and the atomic charges of 76 atom-types, together with five additional correction factors, as

$$\log P = \sum_i a_i \cdot N_i + \sum_j c_j \cdot N_j$$

where a and c are the lipophilic contributions of each atom-type and correction factor, respectively, and N_i and N_j are the number of occurrences of the i th atom-type and j th correction factor, respectively.

Using the SYBYL atomic codes, atom-types are classified as carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorous, and halogens in neutral organic molecules, according to their

hybridization states and their nearest neighboring atoms. Five correction factors were introduced to account for some intramolecular interactions.

Using a training set of 1831 compounds, the best-fitted model gave $r^2 = 0.937$, improving the results obtained without correction factors ($r^2 = 0.908$).

• SLOGP

This is the $\log P$ calculated by a $\rightarrow$ *group-contribution method* proposed by Hou and Xu [Hou and Xu, 2002; Hou and Xu, 2003a; Hou and Xu, 2003b] based on the calculation of $\rightarrow$ *solvent accessible surface area* for 100 atom/group types, together with two additional correction factors, and defined as

$$\log P = \sum_i a_i \cdot N_i + \sum_j c_j \cdot N_j$$

where a and c are the contributions of each atom-type and correction factor, respectively, and N_i and N_j are the number of occurrences of the i th atom-type and j th correction factor, respectively. Using the SMARTS atomic codes, atom-types were defined in the same way as it was for the XLOGP model.

Using a training set of 1850 compounds, containing the same 1831 compounds used to develop the XLOGP model, the best-fitted model gave $r^2 = 0.976$, with $s = 0.368$.

• Duchowitz–Castro $\log P$

This is the $\log P$ calculated by a simple approach based on the contributions of molecule atoms and bonds, as [Duchowicz and Castro, 2000]

$$\log P = c_0 + \sum_i a_i \cdot A_i + \sum_j b_j \cdot B_j$$

where the first summation runs over the atom-types and second one over the bond-types; a and b are the regression coefficients of the different types of atoms (A) (C, H, O, N, etc.) and bonds (B) (C–H, C=O, O–H, etc.). This approach has been tested only on small congeneric data sets.

• lipole

The lipole of a molecule is a measure of the lipophilic distribution and is calculated from atomic lipophilicity values l_i as

$$L = \sum_{i=1}^A r_i \cdot l_i$$

where r_i and l_i are the distance from the $\rightarrow$ *center of mass* and the lipophilicity of the i th atom, respectively; A is the number of atoms in the molecule [TSAR – Oxford Molecular Ltd., 1999].

• topological index of hydrophobicity

This is a topological index based on molecular connectivity defined as

$$\chi_H = {}^1\chi + \sum_k \delta\chi_k$$

where ${}^1\chi$ is the $\rightarrow$ *Randić connectivity index* and $\delta\chi$ are correction factors determined experimentally [Sakhartova and Shatz, 1984; Shatz, Sakhartova *et al.*, 1984]. The corrections are $\delta\chi(\text{alkanes}) = 0$, $\delta\chi(\text{alkylbenzenes}) = -1.597$, $\delta\chi(\text{ketones}) = -3.076$.

• Ferreira–Kiralj hydrophobicity parameters

Two simple structure-based lipophilicity parameters were proposed by Ferreira and Kiralj [Ferreira and Kiralj, 2004] for modeling $\log P$. The first parameter is the fraction w_C of the number of hydrophobic carbon atoms N_C^{hyd} , defined as the number of hydrophobic carbon atoms (all carbon atoms except those in C=O, C–O[−] and CN groups) divided by the number of all nonhydrogen atoms:

$$w_C = \frac{N_C^{\text{hyd}}}{A - N_H}$$

where A is the total number of atoms and N_H the number of hydrogens, respectively.

The second parameter is the surface fraction of hydrophobic carbon atoms S_F , calculated analogously to w_C : instead of atom counts, their CPK atomic surface areas from optimized geometries of compounds (in charged forms at neutral pH) are used.

These parameters were used in addition to other classical $\log P$ descriptors in Partial Least Squares (PLS) regression for modeling biological activities.

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **Li valence vertex degree** → vertex degree
- **L/L quotient matrix** → biodescriptors (⊙ DNA sequences)
- **LMO technique** $\equiv$ *leave-more-out technique* → validation techniques (⊙ cross-validation)
- **loading matrix** → Principal Component Analysis
- **local Balaban index** → connectivity indices
- **Local Chemical Environments** → scoring functions (⊙ MultiLevel Chemical Compatibility)
- **local connectivity indices** → connectivity indices
- **Local Density Of States** → quantum-chemical descriptors (⊙ EIM descriptors)
- **local dipole index** → charge descriptors
- **Local Edge Invariants** → local invariants
- **local ETA index** → ETA indices
- **local functionality index** → ETA indices
- **local hardness** → quantum-chemical descriptors (⊙ hardness indices)
- **local information on distances** $\equiv$ *relative vertex distance complexity* → topological information indices

■ local invariants

These are numerical quantities derived from the molecular topology and used to characterize local properties of a molecule; these numbers are calculated in such a way as to be independent of any arbitrary atom/bond numbering. Local invariants can be distinguished into **Local Vertex Invariants** (LOVIs) and **Local Edge Invariants** (LOEIs), depending on whether they refer to atoms or bonds. They are usually calculated from the → *H-depleted molecular graphs*.

• Local Vertex Invariants (LOVIs)

These are numerical quantities associated with graph vertices independently of any arbitrary vertex numbering, used to characterize local properties in a molecule. They can be either purely

topological if heteroatoms are not distinguished from carbon atoms, or chemical if the heteroatoms are assigned distinct values from carbon atoms, even when these are topologically equivalent [Balaban, 1987, 1994a; Filip, Balaban *et al.*, 1987; Ivanciuc, Balaban *et al.*, 1993b]. To account for the presence of heteroatoms, local vertex invariants can be calculated from molecular graphs where vertices are weighted by physico-chemical $\rightarrow$ *atomic properties*. An ideal set of LOVIs is such that distinct LOVIs are relative to nonequivalent vertices in any graph.

LOVIs of a molecule are usually collected into an A -dimensional vector, A being the number of graph vertices, and, sometimes, as diagonal terms into a $(A \times A)$ diagonal matrix. Examples of matrices collecting LOVIs are the $\rightarrow$ *vertex degree matrix*, $\rightarrow$ *vertex Zagreb matrix*, and $\rightarrow$ *modified vertex Zagreb matrix* and several $\rightarrow$ *augmented matrices*.

Local vertex invariants are used to calculate several molecular $\rightarrow$ *topological indices* by applying different operators such as addition of LOVIs, addition of squares of LOVIs, addition of reciprocal geometric means for any pair of adjacent vertices. Moreover, they can be used to obtain $\rightarrow$ *canonical numbering* of molecular graphs and compare molecules to study $\rightarrow$ *molecular branching* and *centricity*.

The most well-known LOVIs are $\rightarrow$ *vertex degree*, $\rightarrow$ *valence vertex degree*, $\rightarrow$ *bond vertex degree*, and the other several variants of the vertex degree, $\rightarrow$ *atom eccentricity*, $\rightarrow$ *vertex distance degree*, $\rightarrow$ *walk degree*, $\rightarrow$ *atomic path counts*, $\rightarrow$ *atomic ID numbers*, $\rightarrow$ *path degree*, $\rightarrow$ *extended connectivity*, $\rightarrow$ *exponential sum connectivities*, $\rightarrow$ *graph potentials*, LOVIs calculated by $\rightarrow$ *MPR approach* and those applying $\rightarrow$ *centric operator* and $\rightarrow$ *centrocomplexity operator* to $\rightarrow$ *layer matrices*.

A general approach to derive a local vertex invariant from a symmetric $\rightarrow$ *graph-theoretical matrix* ($A \times A$) is to compute the sum of the elements in the i th row, or j th column, of the matrix $\mathbf{M}$:

$$\mathcal{L}_i \equiv VS_i(\mathbf{M}, w) = \sum_{j=1}^A [(\mathbf{M}, w)]_{ij}$$

where VS indicates the $\rightarrow$ *row sum operator* and the resulting LOVI is called **vertex sum**; w is the $\rightarrow$ *weighting scheme* used to calculate the molecular matrix $\mathbf{M}$. $\mathcal{L}$ is here adopted as the general symbol for local vertex invariants. For unsymmetrical graph-theoretical matrices $\mathbf{UM}(w)$, the **vertex double sum**, denoted as VDS_i , was defined as local vertex invariant instead of the vertex sum [Ivanciuc, 1999c]:

$$\mathcal{L}_i \equiv VDS_i(\mathbf{M}, w) = \sum_{j=1}^A [\mathbf{M}(w)]_{ij} + \sum_{j=1}^A [\mathbf{M}(w)]_{ji} - [\mathbf{M}(w)]_{ii}$$

where the diagonal element is subtracted because it is added in both summations.

Another common LOVI derived from a graph-theoretical matrix is the maximum value of the matrix entries in the i th row:

$$\mathcal{L}_i = \max_j ([\mathbf{M}(w)]_{ij})$$

Other typical LOVIs are obtained by adding only matrix entries corresponding to the vertices adjacent to the i th vertex:

$$\mathcal{L}_i = \sum_{j=1}^A a_{ij} \cdot [\mathbf{M}(w)]_{ij}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to one for pairs of adjacent vertices, and zero otherwise. Moreover, an extension of this kind of LOVIs is represented by higher order LOVIs calculated as

$${}^k\mathcal{L}_i = \sum_{j=1}^A [\mathbf{M}(w)]_{ij} \cdot \delta(d_{ij}; k)$$

where $\delta(d_{ij}; k)$ is the Kronecker delta function that is equal to one for pairs of vertices at a topological distance of k , and zero otherwise.

Extended LOVIs, defined by using the same formula as the $\rightarrow$ *extended connectivity*, are generated according to the following expression:

$$\mathcal{L}_i^k = \sum_{j=1}^A a_{ij} \cdot \mathcal{L}_j^{k-1} \quad k = 1, 2, \dots$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix*, being equal to one for pairs of adjacent vertices, and zero otherwise; at the beginning ($k = 1$), any kind of LOVI is simply the $\rightarrow$ *vertex degree* δ , that is, the number of adjacent vertices.

Other LOVIs can be generated by different combinations of the basic LOVIs or any other atomic property, used as the $\rightarrow$ *weighting scheme* w for graph vertices. Let w_1 , w_2 , and w_3 be three vertex weighting schemes, then a generalized LOVI function is here proposed as [Authors, This book]

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[\sum_{j=1}^A a_{ij}^{(k)} \cdot \left(\frac{w_{3j}^\phi}{f(d_{ij}, \gamma)} \right) \right]^\lambda$$

where w_1 , w_2 , and w_3 are the vertex weightings, $a^{(k)}$ are the elements of the k th power of the $\rightarrow$ *adjacency matrix* corresponding to the number of $\rightarrow$ *equipose random walks* of length k (*walk count*) from vertex v_i to vertex v_j , and f is a distance smoothing function used to modulate the role of distances in attenuating contributions from vertices far apart. α , β , γ , ϕ , and λ are user-defined real parameters. The distance smoothing functions are those proposed in the definition of the $\rightarrow$ *interaction graph matrices* and $\rightarrow$ *perturbation graph matrices*:

$$f_1(d_{ij}, \gamma) = d_{ij}^\gamma \quad f_2(d_{ij}, \gamma) = (d_{ij} + 1)^\gamma \quad f_3(d_{ij}, \gamma) = 2^{\gamma \cdot d_{ij}} \quad f_4(d_{ij}, \gamma, x) = (d_{ij} \cdot x^{(d_{ij}-1)})^\gamma$$

Most of the well-known LOVIs are encompassed by this general definition.

According to this scheme, if $\lambda = 0$, a general LOVI for the i th vertex is defined taking into account only properties of the vertex itself:

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta$$

If $\gamma = 0$, any information related to the topological distance between vertices is neglected and the LOVI reduces to

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[\sum_{j=1}^A a_{ij}^{(k)} \cdot w_{3j}^\phi \right]^\lambda$$

Moreover, there are two basic ways to consider the vertex surrounding, depending on the power k of the adjacency matrix. If $k = 1$, only properties of the vertices bonded to the i th vertex are considered:

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[\sum_{j=1}^A a_{ij} \cdot \left(\frac{w_{3j}^\phi}{f(d_{ij}, \gamma)} \right) \right]^\lambda$$

In this case, a_{ij} being the elements of the adjacency matrix, the summation runs over all the pairs of vertices, but the only nonvanishing contributions are from adjacent vertices; moreover, the influence of the smoothing functions is constant, the distance between adjacent vertices being always equal to one. If $k = 0$, properties of all the vertices in the graph take part in LOVI definition, each contributing by a quantity tuned by its separation from the considered vertex:

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[\sum_{j=1}^A \left(\frac{w_{3j}^\phi}{f(d_{ij}, \gamma)} \right) \right]^\lambda$$

The **VTI indices**, proposed by Ivanciuc [Ivanciuc, 1989], which are combinations of $\rightarrow$ topological distances d and $\rightarrow$ vertex degrees δ , are included in this scheme. They are defined as

$$\text{VTI}_i = \delta_i^\beta \cdot \sum_{j=1}^A d_{ij}^\gamma \cdot \delta_j^\phi$$

Eighteen VTI indices were defined by setting $\beta = 0, \pm 1$; $\gamma = \pm 1$; $\phi = 0, \pm 1$ (Table L8). Among these, combination $\beta = 0, \gamma = 1, \phi = 0$ (ID 1) gives the $\rightarrow$ distance sum, $\beta = 1, \gamma = 1, \phi = 0$ (ID 2) gives the $\rightarrow$ vertex degree distance, $\beta = -1, \gamma = 1, \phi = 0$ (ID 3) gives the LOVI t_i used to calculate the $\rightarrow J_i$ index, and $\beta = 0, \gamma = -1, \phi = 0$ (ID 10) gives the $\rightarrow$ reciprocal distance sum.

Other LOVIs similar to VTI indices are derived as the row sums of $\rightarrow$ distance-degree matrices for different combinations of β , γ , and ϕ parameters. These local indices were extensively

Table L8 List of the 18 VTI indices.

ID	γ	β	ϕ	VTI	ID	γ	β	ϕ	VTI
1	1	0	0	$\sum_j d_{ij} = \sigma_i$	10	-1	0	0	$\sum_j d_{ij}^{-1} = RDS_i$
2	1	1	0	$\delta_i \cdot \sum_j d_{ij} = \delta_i \cdot \sigma_i$	11	-1	1	0	$\delta_i \cdot \sum_j d_{ij}^{-1} = \delta_i \cdot RDS_i$
3	1	-1	0	$\delta_i^{-1} \cdot \sum_j d_{ij} = \sigma_i / \delta_i$	12	-1	-1	0	$\delta_i^{-1} \cdot \sum_j d_{ij}^{-1} = RDS_i / \delta_i$
4	1	0	1	$\sum_j d_{ij} \cdot \delta_j$	13	-1	0	1	$\sum_j d_{ij}^{-1} \cdot \delta_j$
5	1	0	-1	$\sum_j d_{ij} \cdot \delta_j^{-1}$	14	-1	0	-1	$\sum_j d_{ij}^{-1} \cdot \delta_j^{-1}$
6	1	1	1	$\delta_i \cdot \sum_j d_{ij} \cdot \delta_j$	15	-1	1	1	$\delta_i \cdot \sum_j d_{ij}^{-1} \cdot \delta_j$
7	1	1	-1	$\delta_i \cdot \sum_j d_{ij} \cdot \delta_j^{-1}$	16	-1	1	-1	
8	1	-1	1	$\delta_i^{-1} \cdot \sum_j d_{ij} \cdot \delta_j$	17	-1	-1	1	$\delta_i^{-1} \cdot \sum_j d_{ij}^{-1} \cdot \delta_j$
9	1	-1	-1	$\delta_i^{-1} \cdot \sum_j d_{ij} \cdot \delta_j^{-1}$	18	-1	-1	-1	$\delta_i^{-1} \cdot \sum_j d_{ij}^{-1} \cdot \delta_j^{-1}$

studied for analyzing $\rightarrow$ molecular branching and used to derive molecular descriptors obtained by setting β , γ , and ϕ equal to a number of values such as $-\infty$, -6 , -5 , -4 , -3 , -2 , -1 , $-1/2$, $-1/3$, $-1/4$, $-1/5$, 0 , $1/5$, $1/4$, $1/3$, $1/2$, 1 , 2 , 3 , 4 [Perdih, 2003; Perdih and Perdih, 2003d, 2004].

Another set of local vertex invariants was proposed by Diudea *et al.* [Diudea, Kacso *et al.*, 1996] using a Randić-like formula as

$$\text{conn}(w)_i = \sum_{j=1}^A a_{ij} \cdot (w_i \cdot w_j)^\alpha$$

where w_i and w_j are atomic weightings associated with vertices v_i and v_j ; the summation runs over all the vertices and accounts only for contributions from vertices adjacent to v_i , a_{ij} being the elements of the adjacency matrix; α is a real exponent usually equal to $-1/2$ and sometimes to $+1/2$. By summing all LOVIs over all atoms, the corresponding molecular $\rightarrow$ graph invariant is obtained. If the weighting scheme is the $\rightarrow$ vertex degree, the obtained LOVIs correspond to twice the first order $\rightarrow$ local connectivity indices. Moreover, the $\rightarrow$ Randić–Razinger index and $\rightarrow$ local Balaban index are obtained when the weighting scheme for vertices is the $\rightarrow$ walk degree and the $\rightarrow$ vertex distance degree, respectively [Diudea, Minailiuc *et al.*, 1997a].

Another set of local vertex invariants, denoted as EFTI_i , was derived from $\rightarrow$ fragment topological indices, when one nonhydrogen atom at a time is considered:

$$\text{EFTI}_i = \text{TI}(\mathcal{G}) - \text{IFTI}(\mathcal{G}') - \text{IFTI}(i)$$

where TI is any topological index, increasing with increase in the number of graph vertices, $\mathcal{G}'$ is the subgraph obtained by erasing the i th vertex with its incident edges, $\text{IFTI}(i)$ is the corresponding topological index calculated for the i th vertex that is often equal to zero or constant (e.g., $\text{IFTI}(i) = 1$ for the $\rightarrow$ Hosoya Z index) [Mekenyan, Bonchev *et al.*, 1988a].

• Local Edge Invariants (LOEIs)

Analogous to the local vertex invariants, local edge invariants are descriptors of the graph edges used to characterize local properties in a molecule; they are numbers associated with graph edges independently of any arbitrary edge numbering. They can be calculated as the local vertex invariants of the first-order $\rightarrow$ line graph corresponding to the molecular graph.

Edge invariants can also be directly obtained by $\rightarrow$ physico-chemical properties of the bonds used as the $\rightarrow$ weighting scheme for graph edges, such as bond dipole moments, $\rightarrow$ bond order indices, and so on, as well as from the $\rightarrow$ edge adjacency matrix and $\rightarrow$ edge distance matrix by applying specific matrix operators such as the $\rightarrow$ row sum operator.

Moreover, local edge invariants can be calculated by some combination of the local vertex invariants or atomic properties of the two incident vertices. Two general formulas to derive edge invariants from vertex invariants $\mathcal{L}$ are

$$\mathcal{L}_{ij} = \frac{\mathcal{L}_i + \mathcal{L}_j}{2} \quad \text{and} \quad \mathcal{L}_{ij} = \sqrt{\mathcal{L}_i \cdot \mathcal{L}_j}$$

corresponding to the arithmetic and geometric mean, respectively, of the local vertex invariants of the two vertices v_i and v_j connected by the edge.

The two previous expressions can be extended taking into account all the vertices bonded to the two vertices v_i and v_j forming the edge i – j :

$$\mathcal{L}_{ij} = \sum_{k=1}^A a_{ik} \cdot \mathcal{L}_k + \sum_{k=1}^A a_{jk} \cdot \mathcal{L}_k \quad \text{and} \quad \mathcal{L}_{ij} = \prod_{k=1}^A (\mathcal{L}_k)^{a_{ik}} \cdot \prod_{k=1}^A (\mathcal{L}_k)^{a_{jk}}$$

Their average values can be also considered, using the corresponding arithmetic and geometric means.

A generalization of the concept of $\rightarrow$ *edge connectivity*, defined in terms of $\rightarrow$ *vertex degrees*, is given by the following expression:

$$\mathcal{L}_{ij} = (VS_i[\mathbf{M}] \cdot VS_j[\mathbf{M}])^\lambda$$

where $\mathbf{M}$ is a $\rightarrow$ *graph-theoretical matrix*, VS the $\rightarrow$ *vertex sum operator*, and λ is a variable parameter.

To obtain a final topological index that gives greater weight to terminal bonds, if the vertex sum VS corresponding to terminal vertices of a graph are smaller than the average vertex sums of the interior vertices, a value of $\lambda = -1/2$ is suggested. On the other hand, when the vertex sums corresponding to terminal vertices of the graph are greater than the average vertex sums of the interior vertices, a value of $\lambda = 1/2$ should be chosen to generate bond contributions [Randić, Balaban *et al.*, 2001]. For $\lambda = -1/2$, the bond contribution above defined is the same as that used in the $\rightarrow$ *Ivanciuc–Balaban operator*.

Moreover, another class of local edge invariants can be obtained as the average value of the $\rightarrow$ *vertex sums* of the two incident vertices raised to a variable exponent λ [Randić, Balaban *et al.*, 2001]:

$$\mathcal{L}_{ij} = \left(\frac{VS_i[\mathbf{M}] + VS_j[\mathbf{M}]}{2} \right)^\lambda$$

Another general class of edge invariants was defined as the harmonic mean of the edge invariants of the edges linked to the edge i – j [Alikhanidi and Takahashi, 2006]:

$$\mathcal{L}_{ij} = \frac{\delta_i + \delta_j - 2}{\sum_{k=1}^A a_{ik} \cdot (1/\mathcal{L}_{ik}) + \sum_{k=1}^A a_{jk} \cdot (1/\mathcal{L}_{jk})} \quad k \neq i \neq j$$

where the numerator is the total number of edges incident to the edge i – j ; δ is the number of edges incident to a vertex, that is, the $\rightarrow$ *vertex degree*. In the denominator, the first summation accounts for contributions from edges incident to the i th vertex, while the second one for contributions from edges incident to the j th vertex.

📖 [Klopman, Raychaudhury *et al.*, 1988; Klopman and Raychaudhury, 1990; Balaban and Balaban, 1991; Balaban, Ciubotariu *et al.*, 1991; Balaban and Balaban, 1992; Balaban, Filip *et al.*, 1992; Bonchev and Kier, 1992; Ivanciuc, Balaban *et al.*, 1992; Kier and Hall, 1992a; Balaban and Diudea, 1993; Bonchev, Kier *et al.*, 1993; Balaban, 1992, 1994c, 1995b; Diudea, Horvath *et al.*, 1995a; Medeleanu and Balaban, 1998]

- **localized effect** $\equiv$ *polar effect* $\rightarrow$ electronic substituent constants
- **local polarity index** $\rightarrow$ electric polarization descriptors
- **local profiles** $\rightarrow$ molecular profiles

- **local quantum-chemical properties** → quantum-chemical descriptors
- **local simple flexibility index** → flexibility indices (⊙ global flexibility index)
- **local softness** → quantum-chemical descriptors (⊙ softness indices)
- **local spectral moment** → edge adjacency matrix
- **local synthetic invariant** → iterated line graph sequence
- **local vertex invariants** → local invariants
- **LOEIs** $\equiv$ *LOcal Edge Invariants* → local invariants
- **LOEL** $\equiv$ *Lowest-Observed-Effect Level* → biological activity indices (⊙ toxicological indices)
- **log D_{pH}** $\equiv$ *octanol–water distribution coefficient* → physico-chemical properties (⊙ partition coefficients)
- **log K_{mw}** $\equiv$ *micelle–water partition coefficient* → physico-chemical properties (⊙ partition coefficients)
- **log K_{oc}** $\equiv$ *soil sorption partition coefficient* → physico-chemical properties (⊙ partition coefficients)
- **log K_{ow}** $\equiv$ *octanol–water partition coefficient* → lipophilicity descriptors
- **log P** $\equiv$ *octanol–water partition coefficient* → lipophilicity descriptors
- **London cohesive energy** → Hildebrand solubility parameter
- **lone-pair electrons index** → electronic descriptors
- **lone-pair electrostatic interaction** → electronic descriptors
- **longest walk connectivity index** → connectivity indices (⊙ walk connectivity indices)
- **long hafnian** → algebraic operators (⊙ determinant)
- **LOO technique** $\equiv$ *leave-one-out technique* → validation techniques (⊙ cross-validation)
- **loops** → graph
- **lopping centric information index** → centric indices
- **Lovasz–Pelikan index** → spectral indices (⊙ eigenvalues of the adjacency matrix)
- **LOVIs** $\equiv$ *LOcal Vertex Invariants* → local invariants
- **Löwdin population analysis** → quantum-chemical descriptors
- **Lowest-Observed-Effect Level** → biological activity indices (⊙ toxicological indices)
- **lowest unoccupied molecular orbital** → quantum-chemical descriptors
- **lowest unoccupied molecular orbital energy** → quantum-chemical descriptors
- **LUDI energy function** → scoring functions
- **Lu index** → hyper-Wiener-type indices
- **luminal over-saturation number** → property filters (⊙ drug-like indices)

M

- **MACC descriptors** $\equiv$ *Maximum Auto-Cross-Correlation descriptors* $\rightarrow$ autocorrelation descriptors
- **MACC-2 transform** $\rightarrow$ grid-based QSAR techniques ($\odot$ GRIND descriptors)
- **MACCS keys** $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **macromolecular graph** $\rightarrow$ biodescriptors ($\odot$ peptide sequences)
- **Madan chemical degree** $\rightarrow$ vertex degree
- **magnetic permittivity** $\rightarrow$ physico-chemical properties ($\odot$ magnetic susceptibility)
- **magnetic susceptibility** $\rightarrow$ physico-chemical properties
- **Mahalanobis distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)
- **Main Distance-Dependent Matrix** $\rightarrow$ 4D Molecular Similarity Analysis
- **Mallows C_p** $\rightarrow$ regression parameters
- **Manhattan distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)
- **map connectivity matrices** $\rightarrow$ biodescriptors ($\odot$ proteomics maps)
- **MaP descriptors** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **map invariants** $\rightarrow$ biodescriptors ($\odot$ proteomics maps)

■ MARCH-INSIDE descriptors

The MARCH-INSIDE (*MARKovian CHemicals "IN Silico" DDesign*) method uses the concepts of Markov's Chain Theory to codify information about the molecular structure [González Díaz, Olazabal *et al.*, 2002; González Díaz, Gia *et al.*, 2003; González Díaz, Torres-Gómez *et al.*, 2005]. This procedure considers as the Markovian states the Pauling's electronegativities of the external electron layers (valence electrons) of any atom core in the molecule. The basic idea underpinning the MARCH-INSIDE approach is that a series of atoms interact to form a molecule at an arbitrary initial time t_0 . Then, after this initial hypothetical situation, electrons start to distribute around cores in discrete intervals of time t_k .

MARCH-INSIDE descriptors are derived from the different k th powers of the **electron-transition stochastic matrix**, denoted as ${}^1\Pi$, which is a $\rightarrow$ *stochastic matrix* of dimension $A \times A$ derived from the $\rightarrow$ *electronegativity-weighted adjacency matrix* ${}^{\chi}\mathbf{A}$, modified by a 3D central chirality factor ω [González Díaz, Sánchez *et al.*, 2003], as

$$[{}^{\chi}\mathbf{A}(\omega)]_{ij} = \begin{cases} \chi_j \cdot e^{\omega_j} & \text{if } (i,j) \in E(\mathcal{G}) \\ \chi_i \cdot e^{\omega_i} & \text{if } i=j \\ 0 & \text{if } (i,j) \notin E(\mathcal{G}) \end{cases} \quad \omega = 0, \pm 1$$

where χ are the atomic Pauling's electronegativities, $\mathcal{E}(\mathcal{G})$ is the set of edges in the molecular graph $\mathcal{G}$, and the variable ω accounts for the spatial configuration of every atom in the molecule: $\omega = +1$, if the atom has R- or axial configuration or E-isomerism, $\omega = 0$, if the atom does not have a specific spatial configuration, and $\omega = -1$ if the atom has S- or equatorial configuration or Z-isomerism. If atom chirality is not taken into account ($\omega = 0$), this matrix coincides with the electronegativity-weighted adjacency matrix ${}^{\chi}\mathbf{A}$.

The electron-transition stochastic matrix ${}^1\Pi$ is then defined as:

$$[{}^1\Pi(\omega)]_{ij} = \begin{cases} \frac{\chi_j e^{\omega_j}}{VS_i({}^{\chi}\mathbf{A}(\omega))} & \text{if } (i,j) \in \mathcal{E}(\mathcal{G}) \\ \frac{\chi_i e^{\omega_i}}{VS_i({}^{\chi}\mathbf{A}(\omega))} & \text{if } i = j \\ 0 & \text{if } (i,j) \notin \mathcal{E}(\mathcal{G}) \end{cases} \quad \omega = 0, \pm 1$$

where VS is the $\rightarrow$ *vertex sum operator* that returns for the i th atom the sum of the electronegativity values for all the atoms bonded to the i th atom, including the i th atom itself by the following:

$$VS_i({}^{\chi}\mathbf{A}(\omega)) = \sum_{j=1}^A [{}^{\chi}\mathbf{A}(\omega)]_{ij} = \chi_i \cdot e^{\omega_i} + \sum_{j=1}^A a_{ij} \cdot (\chi_j \cdot e^{\omega_j})$$

where A is the number of the molecule atoms and a_{ij} the elements of the $\rightarrow$ *adjacency matrix*, equal to one for pairs of bonded atoms, and zero otherwise.

Example M1

Calculation of the electron-transition stochastic matrix ${}^1\Pi$ for the molecule shown below. VS_i and CS_j indicate the matrix row and column sums, respectively; Pauling's electronegativities are $\chi_C = 2.5$, $\chi_N = 3.0$, $\chi_O = 3.5$, and $\chi_F = 4.0$.

$\text{N} \equiv \text{C}_2 - \text{C}_1 \begin{matrix} \text{O} \\ \text{F} \end{matrix}$	Atom	O	F	C ₁	C ₂	N	VS_i
	O	0.583	0	0.417	0	0	1
	F	0	0.615	0.385	0	0	1
	C ₁	0.280	0.320	0.200	0.200	0	1
	C ₂	0	0	0.313	0.313	0.375	1
	N	0	0	0	0.455	0.545	1
	CS_j	0.863	0.935	1.315	0.968	0.920	5

${}^1\Pi =$	Atom	O	F	C ₁	C ₂	N
	O	$\frac{O}{O+C_1}$	0	$\frac{C_1}{O+C_1}$	0	0
	F	0	$\frac{F}{F+C_1}$	$\frac{C_1}{F+C_1}$	0	0
	C ₁	$\frac{O}{C_1+O+F+C_2}$	$\frac{F}{C_1+O+F+C_2}$	$\frac{C_1}{C_1+O+F+C_2}$	$\frac{C_2}{C_1+O+F+C_2}$	0
	C ₂	0	0	$\frac{C_1}{C_2+C_1+N}$	$\frac{C_2}{C_2+C_1+N}$	$\frac{N}{C_2+C_1+N}$
	N	0	0	0	$\frac{C_2}{N+C_2}$	$\frac{N}{N+C_2}$

The k th order electron-transition stochastic matrix ${}^k\Pi(\omega)$ is calculated as the k th power of ${}^1\Pi(\omega)$:

$${}^k\Pi(\omega) = ({}^1\Pi(\omega))^k$$

The elements of this matrix are interpreted as the transition probabilities of electrons of going from the i th atom to the j th atom at different time intervals. The diagonal elements are called **self-return probabilities** by analogy with the $\rightarrow$ *self-returning walks*.

Finally, the MARCH-INSIDE total molecular descriptors are the $\rightarrow$ *stochastic spectral moments* of the k th powers of the electron-transition stochastic matrix defined as

$${}^{\text{SR}}\pi_k(\omega) = \text{tr}[{}^k\Pi(\omega)] = \sum_{i=1}^A [{}^k\Pi(\omega)]_{ii}$$

MARCH-INSIDE atom-type descriptors are analogously calculated, but unlike considering the contributions of all atoms in the molecule, only the diagonal elements of each k th order matrix corresponding to the atom of a given type (e.g., halogens, carbons in aliphatic chains, and so on) are summed up to give the molecular descriptor.

Moreover, to model proteins and peptides, MARCH-INSIDE biodescriptors were derived from the k th powers of an electron-transition stochastic matrix based on the $\rightarrow$ *Electronic Charge Index* used in place of the electronegativity [Ramos de Armas, González Díaz *et al.*, 2005].

MEDNE descriptors (*Markovian Electron Delocalization NEgentropies*) are a set of molecular descriptors that, like the MARCH-INSIDE descriptors, are calculated from the different k th powers of the electron-transition stochastic matrix ${}^k\Pi$ [González Díaz, Marrero *et al.*, 2003; Cruz-Montegudo, González Díaz *et al.*, 2008]. Each k th order matrix is transformed into an A -dimensional vector ${}^A\pi_k$ as the following:

$${}^A\pi_k = {}^A\pi_0^T \cdot {}^k\Pi$$

where T is the transpose and ${}^A\pi_0$ an A -dimensional column vector whose elements ${}^A\pi_0(i)$ are the normalized electronegativities defined as the following:

$${}^A\pi_0(i) = \frac{\chi_i \cdot e^{\omega_i}}{\sum_{j=1}^A \chi_j \cdot e^{\omega_j}}$$

where ω is the chirality factor previously defined. It must be noted that the electronegativity of each i th atom is here normalized by using the sum of the electronegativities of all the atoms in the molecule and not only the values of the bonded atoms as in the MARCH-INSIDE descriptors.

The elements of the A -dimensional vector ${}^A\pi_k$ of the k th order are called *absolute probabilities* ${}^A\pi_k(i)$, and they codify the attraction of each i th atom over any electron in the molecule at any time t_k after traveling along the different walks of length smaller than k .

The k th order MEDNE descriptor Θ_k (called **electronic delocalization entropy** [Ramos de Armas, González Díaz *et al.*, 2004]), which is the sum over all atoms of the entropy $\Theta_k(i)$ involved in the attraction of electrons at least k bonds away from any i th atom in the molecule, and defined as

$$\Theta_k = \sum_{i=1}^A \Theta_k(i) = - \sum_{i=1}^A [{}^A\pi_k(i) \cdot \ln {}^A\pi_k(i)]$$

📖 [González Díaz and Uriarte, 2005; González Díaz, Bonet *et al.*, 2007]

- **mass spectral features** → spectra descriptors
- **matching polynomial** → Hosoya Z-index
- **mathematical representation of molecular descriptors** → molecular descriptors

■ matrices of molecules

Matrices are the most common mathematical tool to encode structural information of molecules. They usually are the starting point for the calculation of many → *molecular descriptors* and → *graph invariants*; moreover, they constitute the mathematical form used as the molecule input in the majority of software packages for calculation of molecular descriptors.

Important kinds of matrices are the → *molecular matrix*, which collects atom spatial coordinates, and all the matrices related to → *molecular geometry*, such as the → *geometry matrix*, whose 3D molecular descriptors are derived from, and computational chemistry approaches are based on, → *WHIM weighted covariances matrices* and the → *molecular influence matrix*. Moreover, derived from computational chemistry, the → *charge density matrix* is another fundamental matrix able to give a deep quantum mechanical description of the molecule.

Other important and very popular matrices are the **graph-theoretical matrices**, a huge number of which were proposed in the last decades to derive topological indices and describe molecules from a topological point of view. Graph-theoretical matrices are matrices derived from a molecular graph G , often from a → *H-depleted molecular graph*. However, in a less restrictive sense, graph-theoretical matrices are all the matrices derived from a molecular graph, even if they encode additional contributions from the molecular geometry or other nontopological quantities. A comprehensive collection of graph-theoretical matrices is reported by Janežič *et al.* [Janežič, Miličević *et al.*, 2007] and extended overviews in [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc and Ivanciuc, 1999].

Graph-theoretical matrices can be either **vertex matrices**, if both rows and columns refer to graph vertices (atoms) and matrix elements encode some property of pairs of vertices, or **edge matrices**, if both rows and columns refer to graph edges (bonds) and matrix elements encode some property of pairs of edges. Vertex matrices are square matrices of dimension $A \times A$, A being the number of graph vertices, whereas edge matrices are square matrices of dimension $B \times B$, B being the number of graph edges. In the book, a vertex matrix is usually referred to by omitting the word “vertex,” whereas for edge matrices, the prefix “edge” is always specified in the matrix name and the superscript E is used in the matrix symbol as $^E\mathbf{M}$.

Vertex matrices are undoubtedly the graph-theoretical matrices most frequently used for characterizing a molecular graph. The matrix entries encode some information about pairs of vertices, such as their connectivities, topological distances, sums of the weights of the vertices along the connecting paths; the diagonal entries can encode chemical information about the vertices. The most important vertex matrices are the → *adjacency matrix* $\mathbf{A}$, which encodes information about vertex connectivities and the → *distance matrix* $\mathbf{D}$, which also encodes information about relative locations of graph vertices.

From vertex matrices a huge number of topological indices were proposed. Edge matrices have been less used to characterize a molecular graph and derive molecular descriptors.

The most important edge matrices are the → *edge adjacency matrix* $^E\mathbf{A}$ and → *the edge distance matrix* $^E\mathbf{D}$. Edge matrices of a molecular graph G are usually calculated from the → *line*

graph of the actual molecular graph G , whose vertices represent edges of G [Gutman and Estrada, 1996]; therefore, an edge matrix of G is simply the corresponding vertex matrix of the line graph of G . For instance, the edge adjacency matrix of G is the adjacency matrix A of the line graph of G . Following this approach, a number of edge matrices were proposed.

It is noteworthy to point out that in literature vertex matrices are sometimes referred to as “edge-matrices,” denoted by M_e (or sometimes as ^eM), when only matrix elements corresponding to the pairs of adjacent vertices are different from zero, or “path-matrices,” denoted by M_p (or sometimes as ^pM), when all off-diagonal elements can be different from zero, meaning that all pairs of vertices in the graph (i.e., paths) are accounted for and not only the pairs of adjacent vertices (i.e., edges). The $\rightarrow$ adjacency matrix A , which is one of the fundamental vertex matrices, is an example of “edge-matrix,” whereas the $\rightarrow$ distance matrix D is the corresponding “path-matrix.” Edge- and path-matrices are usually related, and, specifically, edge-matrices are derived from path-matrices by the following:

$$M_e = M_p \otimes A$$

where A is the adjacency matrix and the symbol $\otimes$ indicates the $\rightarrow$ Hadamard matrix product. Examples of path-matrices and related edge-matrices are $\rightarrow$ path-Cluj matrices and $\rightarrow$ edge-Cluj matrices, $\rightarrow$ path-Wiener matrix and $\rightarrow$ edge-Wiener matrix, $\rightarrow$ path-Szeged matrix and $\rightarrow$ edge-Szeged matrix.

Obviously, the prefix “edge-” in the matrix name is misleading because it can refer to two different kinds of graph-theoretical matrices; therefore, care must be taken when dealing with this terminology. In this book, the original matrix names with the prefix “edge” were retained, although they can be sometimes ambiguous, but with the following artifice: edge matrices ($B \times B$) whose rows and columns refer to graph edges are called edge matrices without the hyphen in the matrix name and denoted as ^eM , whereas edge-matrices ($A \times A$) that are vertex matrices containing information only about pairs of adjacent vertices (i.e., edges) are referred to by using the hyphen in the matrix name and denoted as M_e .

Together with vertex matrices and edge matrices, $\rightarrow$ incidence matrices are other important graph-theoretical matrices used to describe a molecular graph. These are matrices whose rows can represent either vertices or edges and whose columns represent some subgraphs, such as edges, paths, or cycles.

Moreover, matrices can be derived from unweighted or vertex- and/or edge-weighted molecular graphs; in the latter case, several $\rightarrow$ weighted matrices can be obtained depending on the $\rightarrow$ weighting scheme.

Most of the graph-theoretical matrices are symmetrical, whereas some of them are unsymmetrical. Examples of unsymmetrical matrices are $\rightarrow$ Szeged matrices, $\rightarrow$ Cluj matrices, $\rightarrow$ random walk Markov matrix, $\rightarrow$ combined matrices such as the topological distance–detour distance combined matrix, and some weighted adjacency and distance matrices.

From unsymmetrical matrices UM , the corresponding symmetric matrices SM can be obtained as

$$SM = UM^T \otimes UM$$

where $\otimes$ is the $\rightarrow$ Hadamard matrix product, or, alternatively, by adding to the unsymmetrical matrix its transposed matrix as

$$SM = UM + UM^T$$

Given a matrix $\mathbf{M}$, some general classes of matrices can be derived by applying algebraic transformations. They are reported below.

Given a matrix $\mathbf{M}$, **power matrices**, denoted by $\mathbf{M}^k$, are defined as

$$\mathbf{M}^k = \mathbf{M} \cdot \mathbf{M}^{k-1}$$

where k is the matrix power [Ivanciuc, 2000e]. An easy way to calculate a k th power of a matrix is passing through the matrix eigenvalue/eigenvector decomposition as

$$\mathbf{M}^k = \mathbf{L}^T \cdot \mathbf{\Lambda}^k \cdot \mathbf{L}$$

where $\mathbf{L}$ is the matrix collecting the eigenvectors of $\mathbf{M}$ and $\mathbf{\Lambda}^k$ is a diagonal matrix whose elements are the eigenvalues λ^k of $\mathbf{M}$ raised to the k th power.

Several molecular descriptors are calculated from matrices raised to different powers; for example, $\rightarrow$ *walk counts*, $\rightarrow$ *self-returning walk counts*, and $\rightarrow$ *spectral moments* are derived from the different powers of the adjacency matrix $\mathbf{A}$ (i.e., $\mathbf{M} = \mathbf{A}$), $\rightarrow$ $^k\alpha$ *descriptors* from the powers of the $\rightarrow$ *path- χ matrix*, $\rightarrow$ *random walk counts* from the powers of the $\rightarrow$ *random walk Markov matrix*, $\rightarrow$ *spectral moments of the edge adjacency matrix* and $\rightarrow$ *TOMOCOMD descriptors* from the powers of the $\rightarrow$ *edge adjacency matrix*.

Generalized matrices, denoted by $\mathbf{M}^\lambda$, are obtained by using the Hadamard matrix product or, alternatively, by raising to different powers the elements of the matrix $\mathbf{M}$:

$$[\mathbf{M}^\lambda]_{ij} = [\mathbf{M}]_{ij}^\lambda$$

where λ is a real parameter.

Examples of molecular descriptors derived from generalized matrices are $\rightarrow$ *distance distribution moments* and $\rightarrow$ W_λ *indices* from the $\rightarrow$ *distance matrix* and $\rightarrow$ *molecular profiles* from the $\rightarrow$ *geometry matrix*. Moreover, $\rightarrow$ *vertex Zagreb matrix* ($\lambda = 2$) and $\rightarrow$ *modified vertex Zagreb matrix* ($\lambda = -2$) are a generalization of the $\rightarrow$ *vertex degree matrix*, and the $\rightarrow$ *generalized molecular-graph matrix* is a generalization of the distance matrix based on variable parameters.

Generalized reciprocal matrices are a class of generalized matrices obtained by raising the matrix elements to some negative exponent:

$$[\mathbf{M}^{-\lambda}]_{ij} = \begin{cases} [\mathbf{M}]_{ij}^{-\lambda} & \text{if } i \neq j \\ [\mathbf{M}]_{ii} & \text{if } i = j \end{cases}$$

where λ is usually an integer positive parameter. Note that the reciprocal is not applied to diagonal elements. In effect, diagonal elements are equal to zero for simple graphs and for vertex-weighted molecular graphs, they are real values representing atomic properties.

The most popular **reciprocal matrices** are obtained for $\lambda = 1$, such as the $\rightarrow$ *Harary matrix*, $\rightarrow$ *reciprocal geometry matrix*, $\rightarrow$ *reciprocal detour matrix*, $\rightarrow$ *reciprocal Szeged matrix*, $\rightarrow$ *reciprocal Cluj matrix*. The $\rightarrow$ *reciprocal square distance matrix* is derived from the distance matrix by setting $\lambda = 2$.

Other important classes of graph-theoretical matrices are neighborhood matrices and matrices derived by a combination of pairs of matrices, such as the sum matrices, augmented matrices, difference matrices, complement matrices, quotient matrices, combined matrices, and expanded matrices (Table M1).

Table M1 Notations of some molecular matrices.

Matrix	Notation	Matrix	Notation
Power matrix	M^k	Quotient matrix	M_1/M_2
Complement matrix	^cM	Combined matrix	$M_1 \wedge M_2$
Neighborhood matrix	$^N M$	Expanded matrix	$M_1 \_ M_2$
Sum matrix	$M_1 M_2 \Sigma$	Difference matrix	$M_1 M_2 \Delta$

Given two equal-sized graph-theoretical matrices M_1 and M_2 , **sum matrices**, denoted as $M_1 M_2 \Sigma$, are obtained by summing corresponding elements of matrices M_1 and M_2 :

$$M_1 M_2 \Sigma = M_1 + M_2$$

The most popular sum matrices are the $\rightarrow$ *adjacency-plus-distance matrix* obtained by summing the vertex $\rightarrow$ *adjacency matrix* A and the vertex $\rightarrow$ *distance matrix* D and the $\rightarrow$ *edge-adjacency-plus-edge-distance matrix* obtained by summing the $\rightarrow$ *edge adjacency matrix* $^E A$ and the $\rightarrow$ *edge distance matrix* $^E D$. From the adjacency-plus-distance matrix the $\rightarrow$ *Schultz molecular topological index* is derived, whereas from the edge-adjacency-plus-edge-distance matrix is derived the $\rightarrow$ *edge-Schultz index*. Moreover, the determinant of the adjacency-plus-distance matrix, that is, the $\rightarrow$ *det|A + D| index*, was also proposed as a molecular descriptor.

The sum matrix resulting from the $\rightarrow$ *geometry matrix* G and the $\rightarrow$ *bond length-weighted adjacency matrix* $^b A$, where elements corresponding to the pairs of adjacent vertices are $\rightarrow$ *bond distances*, was defined by Mihalić *et al.* [Mihalić, Nikolić *et al.*, 1992]. Its determinant, that is, the $\rightarrow$ *det| $^b A + G$ | index*, and $\rightarrow$ *3D-Schultz index* were proposed and used in QSAR modeling as the molecular descriptors.

Augmented matrices, denoted as $^a M$, are a special case of sum matrices, resulting from the sum of a matrix M plus a diagonal matrix whose diagonal elements are some atomic properties:

$$^a M = M + \alpha \cdot I \quad \text{and} \quad ^a M = M + p \cdot I$$

where I is the $\rightarrow$ *identity matrix*, α a constant, and p a A -dimensional vector containing the considered atomic property. Typical augmented matrices are derived from adjacency and distance matrices, whose diagonal elements equal to zero are replaced with any nonzero value.

Augmented matrices $^a M$ were defined for the calculation of local vertex invariants by $\rightarrow$ *MPR approach*. Moreover, Randić [Randić, 1991b] proposed the $\rightarrow$ *augmented adjacency matrix* by replacing the zero diagonal entries of the adjacency matrix with specific values empirically obtained and by characterizing different atom types in the molecule. The augmented distance matrix was defined by analogy with the augmented adjacency matrix. The $\rightarrow$ *Laplacian matrix* is another augmented matrix, obtained by the combination of the $\rightarrow$ *vertex degree matrix* and the adjacency matrix.

Given two equal-sized graph-theoretical matrices M_1 and M_2 , **difference matrices**, denoted as $M_1 M_2 \Delta$, are obtained by subtracting the corresponding elements of matrices M_1 and M_2 :

$$M_1 M_2 \Delta = M_1 - M_2$$

The most popular difference matrix is the $\rightarrow$ *Laplacian matrix* defined as the difference between the $\rightarrow$ *vertex degree matrix* $\mathbf{V}$ and the $\rightarrow$ *adjacency matrix* $\mathbf{A}$; other difference matrices are the $\rightarrow$ *delta matrix*, $\rightarrow$ *detour-delta matrix*, $\rightarrow$ *Wiener difference matrix*, $\rightarrow$ *Szeged difference matrix*, and $\rightarrow$ *Cluj difference matrix*.

Complement matrices are a special case of difference matrices, resulting from the difference between a matrix with all the off-diagonal elements equal to a constant and a matrix $\mathbf{M}$; they are denoted by ${}^c\mathbf{M}$ and defined as

$$[{}^c\mathbf{M}]_{ij} = \begin{cases} K - [\mathbf{M}]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where K is a constant assumed to be greater than all the $\mathbf{M}$ matrix elements.

Examples of complement matrices are the $\rightarrow$ *distance complement matrix*, $\rightarrow$ *complementary distance matrix*, $\rightarrow$ *reverse Wiener matrix*, $\rightarrow$ *complement Barysz distance matrix*, and $\rightarrow$ *detour complement matrix*.

It must be noted that the value of the elements of the reciprocal and complement matrices decreases when the corresponding value in $\mathbf{M}$ increases; therefore, molecular descriptors calculated from reciprocal or complement matrices have numerical behavior and meaning opposite to molecular descriptors derived from the corresponding original matrix $\mathbf{M}$. For instance, the row sums of the $\rightarrow$ *distance matrix* $\mathbf{D}$ (i.e., distance degrees) are greater for outer vertices than for the core vertices; thus, the value of the $\rightarrow$ *Balaban distance connectivity index*, which is based on the inverse of the distance degrees, is much more determined by the core vertices than the outer ones. On the contrary, the row sums of the reciprocal or complement distance matrix are greater for the core vertices and, therefore, $\rightarrow$ *Balaban-like indices* are much more determined by the outer vertices.

Neighborhood matrices, denoted by ${}^N\mathbf{M}$, are sparse matrices whose entries are the elements of $\mathbf{M}$, which have values smaller than or equal to a predefined threshold t and zero otherwise:

$$[{}^N\mathbf{M}]_{ij} = \begin{cases} [\mathbf{M}]_{ij} & \text{if } [\mathbf{M}]_{ij} \leq t \\ 0 & \text{if } [\mathbf{M}]_{ij} > t \end{cases}$$

The adjacency matrix is an example of neighborhood matrix trivially obtained by applying a threshold $t=1$ on the vertex distance matrix; applying a threshold $t=2$ on the same matrix, only topological distances of 1 and 2 are retained. A threshold less than 1 applied on the $\rightarrow$ *distance/detour quotient matrix* results in a matrix whose elements different from zero are, to some extent, related to cyclic substructures. Applied on the $\rightarrow$ *geometry matrix*, a threshold t equal to a predefined geometric distance produces a sparse geometry matrix where only the pairs of atoms not too far from each other are considered, thus accounting for the most relevant interactions.

Derived from a geometry matrix collecting the Euclidean distances between spots of a proteomics map, the $\rightarrow$ *neighborhood geometry matrix* was originally proposed to calculate descriptors of $\rightarrow$ *proteomics maps* by the additional constraint that the matrix element $i-j$ is set at zero also for nonconnected protein spots [Bajzer, Randić *et al.*, 2003].

Given two graph-theoretical matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of the same size, **quotient matrices** are matrices, denoted by $\mathbf{M}_1/\mathbf{M}_2$, whose elements are given by the ratio of the off-diagonal elements

of M_1 over the corresponding elements of M_2 [Randić, Kleiner *et al.*, 1994]:

$$[M_1/M_2]_{ij} = \begin{cases} \frac{[M_1]_{ij}}{[M_2]_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The most popular quotient matrices are the $\rightarrow$ *distance/distance matrices* obtained by the ratio of two different measures of the separation between molecule atoms. These separation measures can be either 2D if derived from the $\rightarrow$ *molecular graph* or 3D if derived from the $\rightarrow$ *molecular geometry*. The quotient matrices proposed as first are the $\rightarrow$ *topological distance/detour distance quotient matrix*, based on 2D distance measures, and the $\rightarrow$ *geometric distance/topological distance quotient matrix* (or alternatively the $\rightarrow$ *topographic distance/topological distance quotient matrix*), based on a 3D and a 2D distance measure.

Note that if both distances are measured through bonds, then the resulting quotient matrix is not meaningful for acyclic graphs, since all the off-diagonal matrix elements are equal to one. Moreover, matrices obtained by the ratio of a 3D distance (geometric or topographic) over a 2D distance (topological, detour, or resistance) are the same for acyclic structures independently of the chosen 2D distance measure, these 2D measures being all the same. In Table M2, a collection of quotient matrices is reported.

Table M2 List of quotient matrices and their suggested symbols.

ID	M_1	M_2	M_1/M_2	Matrix name	*
1	G	D	G/D	Geometric distance/topological distance quotient matrix	
2	G	Δ	G/ Δ	Geometric distance/detour distance quotient matrix	
3	G	Ω	G/ Ω	Geometric distance/resistance distance quotient matrix	
4	T	D	T/D	Topographic distance/topological distance quotient matrix	
5	T	Δ	T/ Δ	Topographic distance/detour distance quotient matrix	
6	T	Ω	T/ Ω	Topographic distance/resistance distance quotient matrix	
7	D	Δ	D/ Δ	Topological distance/detour distance quotient matrix	*
8	D	Ω	D/ Ω	Topological distance/resistance distance quotient matrix	*
9	Δ	Ω	Δ/Ω	Detour distance/resistance distance quotient matrix	*
10	D	DC	D/DC	Distance/distance complement quotient matrix	
11	D	G	D/G	Topological distance/geometric distance quotient matrix	
12	Δ	G	Δ/G	Detour distance/geometric distance quotient matrix	
13	Ω	G	Ω/G	Resistance distance/geometric distance quotient matrix	
14	D	T	D/T	Topological distance/topographic distance quotient matrix	
15	Δ	T	Δ/T	Detour distance/topographic distance quotient matrix	
16	Ω	T	Ω/T	Resistance distance/topographic distance quotient matrix	
17	Δ	D	Δ/D	Detour distance/topological distance quotient matrix	*
18	Ω	D	Ω/D	Resistance distance/topological distance quotient matrix	*
19	Ω	Δ	Ω/Δ	Resistance distance/detour distance quotient matrix	*
20	DC	D	DC/D	Distance complement/distance quotient matrix	

“*” indicates all the matrices that should be calculated only for cycle-containing structures.

The reciprocal of a quotient matrix is still a quotient matrix, obtained by reversing the role of M_2 and M_1 matrices. In the lower part of Table M2, reciprocal matrices (11–20) of the quotient matrices defined above (1–10) are reported.

Given two graph-theoretical matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of the same size, **combined matrices**, denoted by $\mathbf{M}_1 \wedge \mathbf{M}_2$, are unsymmetrical matrices whose upper matrix elements are the elements of $\mathbf{M}_1$ and lower matrix elements are those of $\mathbf{M}_2$:

$$[\mathbf{M}_1 \wedge \mathbf{M}_2]_{ij} = \begin{cases} [\mathbf{M}_1]_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ [\mathbf{M}_2]_{ij} & \text{if } i > j \end{cases}$$

Also, the transpose of this matrix can be defined as

$$[\mathbf{M}_2 \wedge \mathbf{M}_1]_{ij} = \begin{cases} [\mathbf{M}_2]_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ [\mathbf{M}_1]_{ij} & \text{if } i > j \end{cases}$$

but note that eigenvalues and the sum of all the matrix elements are the same for both combined matrices and their transpose and, accordingly, $\rightarrow$ *spectral indices* and $\rightarrow$ *Wiener-type indices*. However, row sums of combined matrices and the corresponding transposed matrices are different and thus all the related graph invariants. The most popular combined matrices are listed in Table M3.

Table M3 List of combined matrices and their suggested symbols.

ID	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_1 \wedge \mathbf{M}_2$	Matrix name
1	G	D	$\mathbf{G} \wedge \mathbf{D}$	Geometric distance–topological distance combined matrix
2	G	Δ	$\mathbf{G} \wedge \Delta$	Geometric distance–detour distance combined matrix
3	G	Ω	$\mathbf{G} \wedge \Omega$	Geometric distance–resistance distance combined matrix
4	T	D	$\mathbf{T} \wedge \mathbf{D}$	Topographic distance–topological distance combined matrix
5	T	Δ	$\mathbf{T} \wedge \Delta$	Topographic distance–detour distance combined matrix
6	T	Ω	$\mathbf{T} \wedge \Omega$	Topographic distance–resistance distance combined matrix
7	D	Δ	$\mathbf{D} \wedge \Delta$	Topological distance–detour distance combined matrix
8	D	Ω	$\mathbf{D} \wedge \Omega$	Topological distance–resistance distance combined matrix
9	Δ	Ω	$\Delta \wedge \Omega$	Detour distance–resistance distance combined matrix
10	D	G	$\mathbf{D} \wedge \mathbf{G}$	Topological distance–geometric distance combined matrix
11	Δ	G	$\Delta \wedge \mathbf{G}$	Detour distance–geometric distance combined matrix
12	Ω	G	$\Omega \wedge \mathbf{G}$	Resistance distance–geometric distance combined matrix
13	D	T	$\mathbf{D} \wedge \mathbf{T}$	Topological distance–topographic distance combined matrix
14	Δ	T	$\Delta \wedge \mathbf{T}$	Detour distance–topographic distance combined matrix
15	Ω	T	$\Omega \wedge \mathbf{T}$	Resistance distance–topographic distance combined matrix
16	Δ	D	$\Delta \wedge \mathbf{D}$	Detour distance–topological distance combined matrix
17	Ω	D	$\Omega \wedge \mathbf{D}$	Resistance distance–topological distance combined matrix
18	Ω	Δ	$\Omega \wedge \Delta$	Resistance distance–detour distance combined matrix

As for quotient matrices, in the case of acyclic graphs, combining two different D distance measures does not make sense because it results into a combined matrix coincident with the $\rightarrow$ *distance matrix D*.

Expanded matrices, denoted as $\mathbf{M}_1 \_ \mathbf{M}_2$, constitute another class of graph-theoretical matrices resulting from the $\rightarrow$ *Hadamard matrix product* of two equal-sized matrices $\mathbf{M}_1$ and $\mathbf{M}_2$:

$$\mathbf{M}_1 \_ \mathbf{M}_2 = \mathbf{M}_1 \otimes \mathbf{M}_2$$

The most popular expanded matrices are $\rightarrow$ *expanded distance matrices*, $\mathbf{D}_\mathbf{M}$, derived as the Hadamard product between the $\rightarrow$ *distance matrix* $\mathbf{D}$ and some different graph-theoretical matrix $\mathbf{M}$, such as the $\rightarrow$ *Wiener matrix*, $\rightarrow$ *Cluj matrices*, $\rightarrow$ *Szeged matrix*, and $\rightarrow$ *walk matrices*. Moreover, $\rightarrow$ *expanded reciprocal distance matrices*, $\mathbf{D}^{-1}_\mathbf{M}$, were defined by analogy with the expanded distance matrices by using the $\rightarrow$ *reciprocal distance matrix* $\mathbf{D}^{-1}$ instead of the distance matrix in the Hadamard product. Finally, $\rightarrow$ *expanded geometric distance matrices*, $\mathbf{G}_\mathbf{M}$, and $\rightarrow$ *expanded reciprocal geometric distance matrices*, $\mathbf{G}^{-1}_\mathbf{M}$, were also proposed based on the $\rightarrow$ *geometry matrix* $\mathbf{G}$ and its reciprocal matrix, respectively.

Combinatorial matrices, denoted by $\mathbf{M}_\mathbf{B}$, are defined in terms of the binomial coefficient of the elements of a graph-theoretical matrix $\mathbf{M}$. Each entry i - j of the combinatorial matrix is calculated as the following [Diudea, 1996a]:

$$[\mathbf{M}_\mathbf{B}]_{ij} = \binom{[\mathbf{M}]_{ij} + k}{2} \quad [\mathbf{M}_\mathbf{B}]_{ij} = \frac{[\mathbf{M}]_{ij}^2 + [\mathbf{M}]_{ij}}{2} \quad k = 1$$

where k is a constant, usually equal to zero or one.

The most common combinatorial matrices are derived from the $\rightarrow$ *distance matrix* and $\rightarrow$ *detour matrix*; these are the $\rightarrow$ *distance-path matrix*, $\rightarrow$ *detour-path matrix*, $\rightarrow$ *delta matrix*, and $\rightarrow$ *detour-delta matrix*.

Other classes of graph-theoretical matrices are $\rightarrow$ *walk matrices*, $\rightarrow$ *layer matrices*, $\rightarrow$ *distance-degree matrices*, and matrices from which $\rightarrow$ *Schultz-type indices* are derived.

The most important graph-theoretical matrices or classes of graph-theoretical matrices are listed in Table M4.

Table M4 Some matrices used as representation of the molecular structure: symbol, current name, number of rows and columns (A , number of vertices; B , number of edges; C^+ , cyclicity; D , topological diameter; and K , maximum walk length), and type (S , symmetric matrix; U , unsymmetrical matrix; B , symmetric or unsymmetrical matrix; and D , diagonal matrix).

Symbol	Matrix	Rows	Columns	Types
A	Adjacency matrix	A	A	S
Ω^{AP}	All-path matrix	A	A	S
C	Atom connectivity matrix	A	A	S
${}^ZD$	Barysz distance matrix	A	A	S
${}^aE(r)$	Bond distance-weighted edge adjacency matrix	B	B	S
${}^rE$	Bond order-weighted edge adjacency matrix	B	B	S
CT	Charge term matrix	A	A	U
CJ	Cluj matrices	A	A	B
$CJ\Delta$	Cluj-detour matrix	A	A	B
CJD	Cluj-distance matrix	A	A	B
D_Δ	Delta matrix	A	A	S
Δ	Detour matrix	A	A	S
Δ/D	Detour/distance quotient matrix	A	A	U
Δ_P	Detour-path matrix	A	A	S
D	Distance matrix	A	A	S
D/D	Distance/distance matrix	A	A	S

(Continued)

Table M4 (Continued)

Symbol	Matrix	Rows	Columns	Types
D/Δ	Distance/detour quotient matrix	A	A	S
D_P	Distance-path matrix	A	A	S
${}^E A \equiv E$	Edge adjacency matrix	B	B	S
${}^{EC} I$	Edge-cycle incidence matrix	B	C^+	U
${}^b A$	Bond length-weighted adjacency matrix	A	A	S
${}^E D$	Edge distance matrix	B	B	S
$\tilde{D}$	Expanded distance matrix	A	A	S
D_M	Expanded distance matrices	A	A	S
G_M	Expanded geometric distance matrices	A	A	S
H_M	Expanded reciprocal distance matrices	A	A	S
H_G_M	Expanded reciprocal geometric distance matrices	A	A	S
EA	Extended adjacency matrix	A	A	S
M	Galvez matrix	A	A	U
V^λ	Generalized vertex degree matrix	A	A	D
${}^E G$	Geometric edge distance matrix	B	B	S
G	Geometry matrix	A	A	S
I	Incidence matrix	A	B	U
IM	Interaction graph matrices	A	A	B
Z	Hosoya matrix	A	A	S
L	Laplacian matrix	A	A	S
LM	Layer matrices	A	$D + 1$	U
M	Molecular matrix	A	3	U
${}^* D$	Multigraph distance matrix	A	A	S
${}^N M$	Neighborhood matrices	A	A	S
P	P matrix	A	A	S
PM	Perturbation graph matrices	A	A	B
Δ^{-1}	Reciprocal detour matrix	A	A	S
$CJ\Delta^{-1}$	Reciprocal Cluj-detour matrix	A	A	B
CJD^{-1}	Reciprocal Cluj-distance matrix	A	A	B
Δ/D^{-1}	Reciprocal detour/distance matrix	A	A	S
D^{-1}	Reciprocal distance matrix	A	A	S
G^{-1}	Reciprocal geometry matrix	A	A	S
D^{-2}	Reciprocal square distance matrix	A	A	B
SZ^{-1}	Reciprocal Szeged matrix	A	A	B
W^{-1}	Reciprocal Wiener matrix	A	A	S
Ω^{-1}	Conductance matrix	A	A	S
Ω	Resistance matrix	A	A	S
RRW	Restricted random walk matrix	A	A	U
SM	Sequence matrices	A	K	U
SZ	Szeged matrix	A	A	B
$SZ_U P$	Szeged property matrices	A	A	U
T	Topological state matrix	A	A	S
${}^{VC} I$	Vertex-cycle incidence matrix	A	C^+	U
V	Vertex degree matrix	A	A	D
${}^k W_M$	Walk diagonal matrix	A	A	D
$W_{(M_1, M_2, M_3)}$	Walk matrix	A	A	U
W	Wiener matrix	A	A	S
ZM	Zagreb matrices	A	A	S
χ	χ matrix	A	A	S

📖 [Rouvray, 1976; Kunz, 1989; Randić, Guo *et al.*, 1993; Ivanciuc, Ivanciuc *et al.*, 1997]

- **matrix method for canonical ordering** → canonical numbering
- **matrix spectrum operators** → spectral indices
- **matrix sum indices** → Wiener-type indices
- **Matthews correlation index** $\equiv$ *Pearson similarity coefficient* → classification parameters
- **maximal binding energy** → scoring functions (⊙ average binding energy)
- **maximal information content** → information content
- **maximal R indices** → GETAWAY descriptors
- **maximal R total index** → GETAWAY descriptors
- **Maximum Auto-Cross-Correlation descriptors** → autocorrelation descriptors
- **maximum bond length** → resonance descriptors (⊙ RC index)

■ maximum common substructure (MCS)

The maximum common substructure (often “maximal common substructure”) of two compounds is the largest possible substructure that is present in both structures. The recognition of a maximum common substructure depends on the defined matching conditions; for example, two substructures are considered to be identical if all atoms and all bonds (single, double, triple, aromatic) can be matched. A further restriction can be applied concerning the number of hydrogen atoms: two nonhydrogen atoms are considered to be identical only if the number of hydrogens bonded to them is equal.

The MCS is a measure and a description of the similarity of two structures whose numerical value *MCS* is the number of common elements provided by the matching conditions, that is, a measure of the size of the maximum common substructure. It is commonly used in → *similarity searching* [Scsibraný and Varmuza, 1992a].

The MCS of a set of *N* compounds, however, may be very small or may not even exist if an exotic structure is contained in the set. Therefore, the common structural characteristics of a set of structures are better described by a set of MCSs, each of them being the MCS of a pair of structures. Such a set is obtained by determining the MCS for all the $N(N-1)/2$ pairs of compounds; then the number N_i of occurrences of each different MCS is counted in the set. Finally, an ordered set of MCSs is obtained by a ranking function, which considers both frequency and size of the MCS:

$$R_i = (1-k) \cdot \frac{N_i}{N} + k \cdot \frac{A_i}{A^{\max}}$$

where A_i is the number of non-hydrogen atoms in MCS_i and $A^{\max}$ is the maximum number of non-hydrogen atoms in all MCSs. k is a user-adjustable parameter (ranging between 0 and 1), which determines the different influence of the frequency and size of MCS; for $k = 1$, only size is considered in the ranking, while for $k = 0$, only the frequencies [Varmuza, Penchev *et al.*, 1998]. The ordered set of MCSs characterizes common and typical structural properties in the investigated set of compounds.

A measure of similarity obtained by the maximum common substructure between two compounds *s* and *t* is given by

$$SI_{st} = \frac{(A+B)_{MCS}}{(A+B)_s} \cdot \frac{(A+B)_{MCS}}{(A+B)_t}$$

where $(A + B)$ is the sum of atoms and bonds in the maximum common substructure MCS, in the s th compound and t th compound, respectively [Durand, Pasari *et al.*, 1999]. A topological distance between the s th and t th compounds is usually defined as

$$d_{st} = (A + B)_s + (A + B)_t - 2 \cdot (A + B)_{\text{MCS}}$$

Usually the considered MCSs are connected graphs, that is, continuous bonded substructures, but disconnected substructures can also be allowed, using a corrected MCS, as for example,

$$\text{MCS}_{st} = A_{\text{MCS}} - k(N_{\text{FRAG}} - 1)$$

where N_{FRAG} is the number of disconnected fragments, k a penalty function between 0 and 1, and A_{MCS} the atom number of the MCS between the s th and t th compounds.

Therefore, the **highest scoring common substructure (HSCS)** value is a standardized variable proposed to measure the similarity between the two molecules s and t [Sheridan and Miller, 1998], defined as

$$\text{HSCS}_{st} = \frac{\text{MCS}_{st} - \text{Mean}(A_s, A_t)}{\text{Std}(A_s, A_t)}$$

where MCS_{st} is the score of the actual MCS, and Mean and Std the mean expected score and standard deviation of the MCS within a large sample of randomly selected molecules of the same size; they are calculated by regression analysis as

$$\begin{aligned}\text{Mean}(A_s, A_t) &= b_0^{\text{mean}} + b_1^{\text{mean}} \cdot \min(A_s, A_t) \\ \text{Std}(A_s, A_t) &= b_0^{\text{std}} + b_1^{\text{std}} \cdot \min(A_s, A_t)\end{aligned}$$

where $\text{Mean}(A_s, A_t)$ is the number of atoms in the smallest molecule for a pair of randomly selected molecules of the same size as molecules s and t . HSCS values greater than 4.0 can be considered highly significant.

 [Cone, Venkataraghavan *et al.*, 1977; Brint and Willett, 1987a, 1987b; Stahl and Mauser, 2005; Sheridan, Hunt *et al.*, 2006; Gardiner, Gillet *et al.*, 2007]

- **maximum electrophilic superdelocalizability** → quantum-chemical descriptors (⊙ electrophilic superdelocalizability)
- **maximum–minimum path matrix** $\equiv$ *detour-distance combined matrix* → detour matrix
- **maximum–minimum path sum** → detour matrix
- **maximum negative charge** → charge descriptors

■ maximum nuclear repulsion for C–H bond index

A molecular descriptor accounting for nuclear repulsion energy between bonded carbon and hydrogen nuclei. It is defined as

$$E_{nm}^{\text{CH}} = \max_k \left(\frac{Z_{\text{C}} \cdot Z_{\text{H}}}{r_{\text{CH}}} \right)_k$$

where Z are the atomic numbers, r_{CH} the C–H bond length, and k refers to a pair of bonded carbon and hydrogen atoms [Katritzky, Sild *et al.*, 1998a]. It possibly encodes the information

about the hybridization state of carbon atoms because the C–H bond length depends on the carbon hybridization state.

- **maximum nucleophilic superdelocalizability** → quantum-chemical descriptors (⊙ nucleophilic superdelocalizability)
- **maximum path degree sequence** → detour matrix
- **maximum path frequency sequence** → detour matrix
- **maximum path matrix** $\equiv$ *detour matrix*
- **maximum path sum** → detour matrix
- **maximum positive charge** → charge descriptors
- **MCASE** $\equiv$ *MULTICASE* → lipophilicity descriptors (⊙ Klopman hydrophobic models)
- **MCB index** → multiple bond descriptors
- **McClelland resonance energy** → delocalization degree indices
- **McConnaughey similarity coefficient** → similarity/diversity (Table S9)
- **McFarland model** → Hansch analysis
- **Mc Gowan's characteristic volume** → volume descriptors
- **MDDM** $\equiv$ *Main Distance-Dependent Matrices* → 4D-Molecular Similarity Analysis
- **MDE vector** $\equiv$ *molecular distance-edge vector*
- **MDL keys** $\equiv$ *MACCS keys* → substructure descriptors (⊙ structural keys)
- **mean absolute deviation** → statistical indices (⊙ indices of dispersion)
- **mean difference** → statistical indices (⊙ indices of dispersion)
- **mean distance degree deviation** → distance matrix
- **mean extended local information on distances** → topological information indices
- **mean information content** → information content
- **mean information content on the adjacency equality** → topological information indices
- **mean information content on the adjacency magnitude** → topological information indices
- **mean information content on the distance degree equality** → topological information indices
- **mean information content on the distance degree magnitude** → topological information indices
- **mean information content on the distance equality** → topological information indices
- **mean information content on the distance magnitude** → topological information indices
- **mean information content on the edge adjacency equality** → topological information indices
- **mean information content on the edge adjacency magnitude** → topological information indices
- **mean information content on the edge-cycle matrix elements equality** → topological information indices
- **mean information content on the edge-cycle matrix elements magnitude** → topological information indices
- **mean information content on the edge cyclic degree equality** → topological information indices
- **mean information content on the edge cyclic degree magnitude** → topological information indices
- **mean information content on the edge degree equality** → topological information indices

- **mean information content on the edge degree magnitude** → topological information indices
- **mean information content on the edge distance degree equality** → topological information indices
- **mean information content on the edge distance degree magnitude** → topological information indices
- **mean information content on the edge distance equality** → topological information indices
- **mean information content on the edge distance magnitude** → topological information indices
- **mean information content on the edge equality** → information connectivity indices
- **mean information content on the edge magnitude** → information connectivity indices
- **mean information content on the incidence matrix** → incidence matrices (⊙ vertex-edge incidence matrix)
- **mean information content on the leverage magnitude** → GETAWAY descriptors
- **mean information content on the vertex-cycle matrix elements equality** → topological information indices
- **mean information content on the vertex-cycle matrix elements magnitude** → topological information indices
- **mean information content on the vertex cyclic degree equality** → topological information indices
- **mean information content on the vertex cyclic degree magnitude** → topological information indices
- **mean information content on the vertex degree equality** → topological information indices
- **mean information content on the vertex degree magnitude** → topological information indices
- **mean information index on atomic composition** → atomic composition indices
- **mean information index on molecular conformations** → information index on molecular conformations
- **mean local information on distances** $\equiv$ *vertex distance complexity* → topological information indices
- **mean overcrossing number** → polymer descriptors
- **mean polarizability** → electric polarization descriptors
- **mean Randić branching index** $\equiv$ *mean Randić connectivity index* → connectivity indices
- **mean Randić connectivity index** → connectivity indices
- **mean square distance index** → distance matrix
- **mean square error** → regression parameters
- **mean topological charge index** → topological charge indices
- **mean Wiener index** → Wiener index
- **median** → statistical indices (⊙ indices of central tendency)
- **median effective dose** → biological activity indices (⊙ pharmacological indices)
- **median inhibitory concentration** → biological activity indices (⊙ toxicological indices)
- **median lethal dose** → biological activity indices (⊙ toxicological indices)
- **MEDNE descriptors** → MARCH-INSIDE descriptors

■ MEDV-13 descriptor

MEDV, namely, the **Molecular Electronegativity Distance Vector**, is a vectorial molecular descriptor comprising of 91 terms encoding information about relative electronegativities, represented by modified $\rightarrow$ *E-state indices*, and topological distances between all the possible pairs of 13 atom types (MEDV-13) [Liu, Cai *et al.*, 2000; Liu, Yin *et al.*, 2001b, 2002a, 2002b; Sun, Zhou *et al.*, 2004].

To generate MEDV descriptors, first the atoms in the molecule are assigned to one of the 13 defined atom types (Table M5). These are distinguished on the basis of the chemical element of the most occurring atoms in organic molecules, the number of bonded non-hydrogen atoms, that is, the $\rightarrow$ *vertex degree*, which reflects the local topological environment of an atom, and the number of valence electrons of the atom, which is used to distinguish atoms with the same vertex degree but different chemical element.

Table M5 MEDV-13 atom types. δ is the atom vertex degree.

Type	Chemical element	δ	Type	Chemical element	δ	Type	Chemical element	δ
1	C	1	6	N, P	2	10	O, S	2
2	C	2	7	N, P	3	11	S	3
3	C	3	8	P	4	12	S	4
4	C	4	9	O, S	1	13	F, Cl, Br, I	1
5	N, P	1						

Moreover, for each i th atom in the molecule, a modified $\rightarrow$ *E-state index*, used as a measure of relative electronegativity, is calculated as

$$S_i^* = I_i^* + \Delta I_i^* = I_i^* + \sum_{j=1}^A \frac{I_i^* - I_j^*}{d_{ij}^2}$$

where d_{ij} is the $\rightarrow$ *topological distance* between i th and j th atoms and I^* is a modified $\rightarrow$ *intrinsic state* defined as

$$I_i^* = \sqrt{\frac{Z_i^v}{4}} \cdot \frac{(2/L_i)^2 \cdot \delta_i^b + 1}{\delta_i}$$

where Z_i^v is the number of valence electrons and the $\rightarrow$ *valence vertex degree* δ_i^v is replaced by the $\rightarrow$ *bond vertex degree* δ_i^b , which accounts for atom connectedness and bond multiplicity.

On the basis of the values of the vertex degree δ , the bond vertex degree δ^b , and the modified intrinsic state I^* , atoms can be classified into 43 types, called atomic attributes, which are proposed by analogy with the *E-state* atom types of Kier and Hall. In defining the atomic attributes, a conjugated system indicator (CSI) is used instead of the aromatic system indicator to distinguish atoms located at different positions of a conjugated system because they have different effects on the molecule (Table M6).

Table M6 MEDV-13 atomic attributes: δ^b , bond vertex degree; δ , vertex degree; I^* , modified intrinsic state.

No.	Atom type	δ^b	δ	I^*	No.	Atom type	δ^b	δ	I^*
1	—CH ₃	1	1	2.000	23	>N—	3	3	1.491
2	—CH ₂ —	2	2	1.500	24	=NH	2	1	3.354
3	>CH—	3	3	1.333	25	=N—	3	2	2.236
4	>C<	4	4	1.250	26	≡N	3	1	4.472
5	=CH ₂	2	1	3.000	27	aNH	1.5	1	2.795
6	=CH—	3	2	2.000	28	aN—	2.5	2	1.957
7	=C<	4	3	1.667	29	aNa	3	2	2.236
8	=C=	4	2	2.500	30	=N—(=)	5	3	2.236
9	≡CH	3	1	4.000	31	—SH	1	1	1.769
10	≡C—	4	2	2.500	32	—S—	2	2	1.157
11	aCH ₂	1.5	1	2.500	33	=S	2	1	2.313
12	aCH—	2.5	2	1.750	34	>S=	4	3	1.134
13	aC<	3.5	3	1.500	35	≥S≤	6	4	1.123
14	aCHa	3	2	2.000	36	—F	1	1	2.646
15	aCa—	4	3	1.667	37	—Cl	1	1	1.911
16	aaCa	4.5	3	1.833	38	—Br	1	1	1.654
17	—OH	1	1	2.449	39	—I	1	1	1.534
18	=O	2	1	3.674	40	—PH ₂	1	1	1.615
19	—O—	2	2	1.837	41	—PH—	2	2	1.056
20	aO	1.5	1	3.062	42	>P—	3	3	0.870
21	—NH ₂	1	1	2.236	43	≥P<	5	4	0.901
22	—NH—	2	2	1.677					

The symbol “a” refers to a bond in any conjugated system, including aromatic systems. Data from Liu, Yin et al. [Liu, Yin et al., 2002b].

Finally, for each combination of two atom types (u, v), a single molecular descriptor $h(u, v)$ is calculated as

$$h(u, v) = \sum_{i \in u} \sum_{j \in v} \frac{S_i^* \cdot S_j^*}{d_{ij}^2} \quad u, v = 1, 2, \dots, 13$$

where the first sum runs over all the atoms of type u and the second sum on the atoms of type v , S^* is the modified E -state index, and d_{ij} the topological distance between vertices v_i and v_j .

Since 13 different atom types are considered, a total of 91 ((13 × 14)/2) different molecular descriptors result, which constitute the final MEDV-13 vector.

An extension of MEDV-13 descriptor, called **Molecular Holographic Distance Vector** (MHDV) was further proposed to describe the structure of molecules containing several heteroatoms and multiple bonds as well as → *peptide sequences* [Liu, Yin et al., 2001a].

➤ **melting point** → physico-chemical properties

■ Membrane Interaction QSAR analysis (MI-QSAR)

MI-QSAR is a methodology that combines classic molecular descriptors with membrane–solute intermolecular properties of compounds to model chemically and structurally diverse compounds interacting with cellular membranes [Kulkarni, 1999; Kulkarni and Hopfinger, 1999; Iyer, Mishra et al., 2002]. MI-QSAR aims at providing insight into the mechanism of skin

penetration, capturing features of cellular lateral transverse transport involved in the overall skin penetration process of organic compounds.

The MI-QSAR method is receptor based, in effect the assumption made is that the phospholipid regions of a cellular membrane constitute the “receptor.” The receptor is usually constructed as a monolayer from the phospholipids that comprise the cell membrane of the system of interest; for instance, a single dimyristoylphosphatidylcholine (DMPC) molecule is selected as the model phospholipid and an assembly of 25 DMPC molecules ($5 \times 5 \times 1$) in (x , y , z) directions, respectively, is used as the model membrane monolayer.

Molecular dynamics simulations (MDSs) on the model membrane are initially carried out to allow for structural relaxation and distribution of the kinetic energy over the monolayer. Then, each molecule is inserted at three different positions in the monolayer, with the most polar group of the molecule “facing” toward the headgroup region of the monolayer. The three positions are (a) headgroup region, (b) center region of the aliphatic chains, and (c) tail region of the aliphatic chains. Separate MDSs are performed for each compound in each of the three trial positions, and the most favorable orientation and location of the compound in the monolayer is determined. Then, the MI-QSAR descriptors are calculated. These are divided into (a) general **intramolecular solute descriptors**, (b) **intermolecular solute–membrane interaction descriptors**, and (c) **solute aqueous dissolution and solvation descriptors** [Iyer, Tseng *et al.*, 2007]. The first set of solute properties are the molecular descriptors calculated by the program Cerius and examples are given in Table M7. The intermolecular solute–membrane interaction descriptors (set b) are derived from MDS trajectories; these particular intermolecular descriptors are calculated using the most stable solute–membrane geometry realized from MDS sampling of the three initial positions for each compound (Table M8). Solute aqueous dissolution and solvation descriptors (set c), although computed by intramolecular computational methods, are intermolecular properties, the first three relating to solute solvation and the last three to solute dissolution (Table M9).

Table M7 MI-QSAR general intramolecular solute descriptors (set a).

Symbol	General intramolecular solute descriptors
HOMO	Highest occupied molecular orbital energy
LUMO	Lowest unoccupied molecular orbital energy
μ	Dipole moment
V_m	Molecular volume
SA	Molecular surface area
ρ	Density
MW	Molecular weight
MR	Molecular refractivity
HBA	Number of hydrogen-bond acceptors
HBD	Number of hydrogen-bond donors
RBN	Number of rotatable bonds
CPSA	Charged partial surface area descriptors
$^m\chi$ and $^m\kappa$	Kier and Hall connectivity and shape descriptors
R_G	Radius of gyration
I	Principal moment of inertia
PSA	Polar surface area
S_{conf}	Conformational entropy
q_x	Partial atomic charge densities

Table M8 MI-QSAR intermolecular solute–membrane interaction descriptors (set b).

Symbol	Intermolecular solute–membrane interaction descriptors
$\langle F(\text{total}) \rangle$	Average total free energy of interaction of the solute and membrane
$\langle E(\text{total}) \rangle$	Average total interaction energy of the solute and membrane
$E_{\text{INTER}}(\text{total})$	Interaction energy between the solute and the membrane at the total intermolecular system minimum potential energy
$E_{\text{XY}}(Z)_E$	$Z = 1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen-bonding, torsion, and combinations thereof energies at the total intermolecular system minimum potential energy. X, Y can be the solute, S, and/or membrane, M, and if $E = \text{free}$, then $X = Y = S$ and the energies are for the solute not in the membrane, but isolated by itself.
$\Delta E_{\text{XY}}(Z)$	Change in the $Z = 1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen-bonding, torsion, and combinations thereof energies due to the uptake of the solute to the total intermolecular system minimum potential energy
$E_{\text{TT}}(Z)$	$1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen bonding, torsion, and combinations thereof energies of the total (solute and membrane model) intermolecular minimum potential energy
$\Delta E_{\text{TT}}(Z)$	Change in the $Z = 1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen-bonding, and combinations thereof of the total (solute and membrane model) intermolecular minimum potential energy
ΔS	Change in entropy of the membrane due to the uptake of the solute
S	Absolute entropy of the solute–membrane system
$\Delta \rho$	Change in density of the model membrane due to the permeating solute
$\langle d \rangle$	Average depth of the solute molecule from the membrane surface

Table M9 MI-QSAR solute aqueous dissolution and solvation descriptors (set c)

Symbol	Solute aqueous dissolution and solvation descriptors
$F(\text{H}_2\text{O})$	Aqueous solvation free energy
$F(\text{oct})$	1-Octanol solvation free energy
$\log P$	1-Octanol/water partition coefficient
$E(\text{coh})$	Cohesive packing energy of the solute molecules
T_{M}	Hypothetical crystal-melt transition temperature of the solute
T_{G}	Hypothetical glass transition temperature of the solute

 [Kulkarni, Han *et al.*, 2002; Santos-Filho, Hopfinger *et al.*, 2004; Li, Liu *et al.*, 2005]

- **MEP** $\equiv$ *Molecular Electrostatic Potential* $\rightarrow$ quantum-chemical descriptors
- **Merrifield–Simmons bond order** $\rightarrow$ symmetry descriptors ($\odot$ Merrifield–Simmons index)
- **Merrifield–Simmons index** $\rightarrow$ symmetry descriptors
- **mesomeric effect** $\rightarrow$ electronic substituent constants
- **Method of Ideal Symmetry** $\rightarrow$ geometry matrix
- **Meyer anchor sphere volume** $\rightarrow$ size descriptors
- **Meyer–Richards similarity index** $\rightarrow$ quantum similarity
- **Meyer visual descriptor of globularity** $\rightarrow$ shape descriptors

➤ **Meylan–Howard hydrophobic model** $\equiv$ KOWWIN $\rightarrow$ lipophilicity descriptors

■ Mezey 3D shape analysis

This is an approach to shape analysis and comparison of molecules based on algebraic topological methods suitable for algorithmic, nonvisual analysis, and coding of molecular shapes by $\rightarrow$ *computational chemistry* [Mezey, 1985, 1993c]. Several topological methods were proposed for the analysis and coding of molecular shapes, most of them based on the concept of $\rightarrow$ *molecular surface*. In particular, the **Shape Group Method** (SGM) is a topological shape analysis technique of any, almost everywhere twice continuously, differentiable 3D functions (e.g., $\rightarrow$ *electron density*). A detailed description of these methods is given in Mezey [Mezey, 1990c].

In general, topological methods are based on subdividing the molecular surface into domains, according to physical or geometrical conditions [Mezey, 1991b].

For example, if two molecular surfaces of the same molecule are considered to be based on two different physical properties such as the $\rightarrow$ *electron isodensity contour surface* $G_1(m_1)$ and the $\rightarrow$ *molecular electrostatic potential contour surface* $G_2(m_2)$, then the former can be subdivided into domains according to electrostatic potential values. The interpenetration of the two surfaces provides several closed loops on the isodensity contour surface; these loops are sets of points of G_1 with equal value m_2 of $\rightarrow$ *molecular electrostatic potential* (MEP), and define the boundaries of the surface G_1 regions that are characterized by MEP values either greater or lower than the threshold value m_2 for all the points in the region. Using different threshold values of MEP, several different electrostatic potential ranges can be mapped on the isodensity surface; these ranges define a subdivision of $G_1(m_1)$ into domains whose topological interrelations can be characterized by a numerical code, which provides a shape characterization of the molecule.

Applying the same approach to several molecules, the similarity of their shape can be searched for by comparing the topological relations among the corresponding domains on the molecular surfaces.

An alternative approach to domain subdivisions of the molecular surface is based on local curvature properties. It is applicable only to differentiable molecular surfaces such as contour surfaces, for example, the electron isodensity contour surface $G(m)$, m being the threshold value defining the contour surface.

The local curvature properties of the surface $G(m)$ in each point $\mathbf{r}$ of the surface are given by the eigenvalues of the local Hessian matrix. Moreover, for a defined reference curvature b , the number $\mu(\mathbf{r}, b)$ is defined as the number of local canonical curvatures (Hessian matrix eigenvalues) that are less than b . Usually b is chosen equal to zero and therefore the number $\mu(\mathbf{r}, 0)$ can take values 0, 1, or 2 indicating that at the point $\mathbf{r}$ the molecular surface is locally concave, saddle type, or convex, respectively. The three disjoint subsets A_0 , A_1 , and A_2 are the collections of the surface points at which the molecular surface is locally concave, saddle type, or convex, respectively; the maximum connected components of these subsets A_0 , A_1 , and A_2 are the surface domains denoted by $D_{0,k}$, $D_{1,k}$, and $D_{2,k}$ where the index k refers to an ordering of these domains, usually according to decreasing surface size (Figure M1).

The mutual arrangement of the domains along the molecular surface can be represented by the topological neighborhood relation, two domains being neighbors if they have a common boundary line. Therefore, a symmetric square *shape matrix* can be built where the rows and columns represent the surface domains; the off-diagonal entries of the shape matrix can be equal to 1 if the considered domains are adjacent and 0 otherwise, the diagonal entries are equal to the number μ (usually 0, 1, or 2) depending on the domain type. If the surface domains are

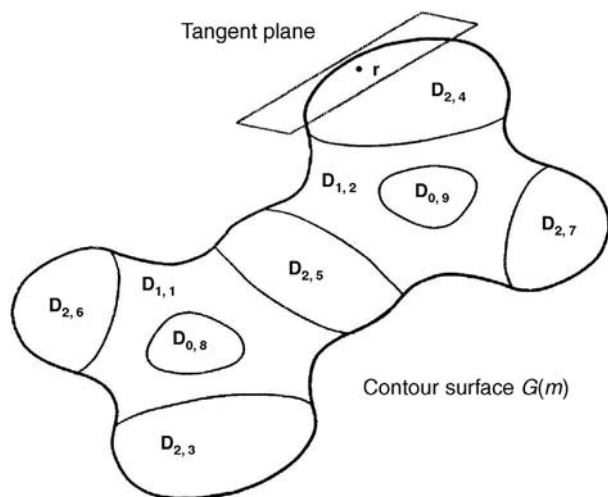


Figure M1 A subdivision of a molecular contour surface $G(m)$, based on local curvature properties. The contour surface is subdivided into locally concave ($D_{0,k}$), saddle-type ($D_{1,k}$), and locally convex ($D_{2,k}$) domains with respect to the local tangent plane in each point r . The second index k refers to an ordering of these domains according to decreasing surface area.

listed in increasing order according to the size of their surface areas, then the shape matrix encodes both shape and size information. Moreover, from the shape matrix, the corresponding *shape graph* can be derived as additional tool for shape characterization of the molecular surface with respect to the reference curvature b ; in fact, shape matrix and shape graph are topological descriptors of molecular surface shape regarded as 3D topological shape codes. They are particularly useful in quantifying the similarity of shapes of the different molecules; the comparison of molecular shape is reduced to a comparison of shape matrices or graphs.

It is worth noting that the topological relations among the domains are invariant within a given range of different molecular conformations. In effect, change in molecular conformation can lead to change in size, location, and even in domain existence, but for certain conformational changes, the existence and the mutual neighborhood relations of the domains remain invariant.

Considering a finite number of threshold values m , a set of contour surfaces $G(m)$ is studied for each molecule combined with a set of reference curvature values b . Therefore, for each pair (m, b) of parameters, the curvature domains $D_0(m, b)$, $D_1(m, b)$, and $D_2(m, b)$ are computed and the truncation of contour surfaces $G(m)$ is performed by removing all curvature domains D_μ of specified type μ (in most applications $\mu = 2$) from the contour surface, thus obtaining a truncated surface $G(m, \mu)$ for each (m, b) pair. For most small changes of the parameter values, the truncated surfaces remain topologically equivalent and only a finite number of equivalence classes are obtained for the entire range of a and b values.

In the next step, the *shape groups* of the molecular surface, that is, the zero-, one-, and two-dimensional algebraic homology groups of the truncated surfaces, are computed. The zero-, one-, and two-dimensional **Betti numbers** $b_\mu^0(m, b)$, $b_\mu^1(m, b)$, and $b_\mu^2(m, b)$ are the ranks of these zero-, one-, and two-dimensional homology groups, that is, the shape groups. They are a list of

topologically invariant numbers and represent a numerical shape code of the molecule, providing a detailed shape characterization of the distribution of the property used to define the molecule surface. In practice, the Betti numbers of 1D shape groups give the most important chemical information and the analysis is often restricted to this class of Betti numbers.

📖 [Mezey, 1987a, 1987b, 1988a, 1988b, 1988c, 1989, 1990b, 1991c, 1992, 1993a, 1993b, 1993d, 1994, 1996, 1997a, 1999; Arteca and Mezey, 1987, 1988a, 1988b, 1989; Arteca, Jammal *et al.*, 1988a, 1988b; Walker, Maggiora *et al.*, 1995]

- **ME-MFP descriptors** → substructure descriptors (⊙ structural keys)
- **MFP descriptors** ≡ *Mini-FingerPrints* → substructure descriptors (⊙ structural keys)
- **MHDV** ≡ *molecular holographic distance vector* → MEDV-13 descriptor
- **micelle–water partition coefficient** → physico-chemical properties (⊙ partition coefficients)
- **MIC index** ≡ *Modified Information Content index* → indices of neighborhood symmetry
- **Migration Index** → chromatographic descriptors (⊙ capacity factor)
- **MINBID** → ID numbers
- **MINCID** → ID numbers
- **Mini-FingerPrints** → substructure descriptors (⊙ structural keys)
- **minimal spanning tree** → graph
- **minimal steric difference** → minimal topological difference

■ Minimal Topological Difference (MTD)

Among the → *hyperstructure-based QSAR techniques*, the MTD method is based on the approximate atom-per-atom superimposition of the n molecules of a → *data set* to build a → *hypermolecule* (i.e., 3D → *hyperstructure*): hydrogen atoms, small differences in atomic positions, bond lengths and bond angles are neglected. The S vertices of the hypermolecule correspond to the positions of the data set molecule atoms [Simon, Dragomir *et al.*, 1973; Simon and Szabadai, 1973b; Simon, Chiriac *et al.*, 1984].

The basic idea is that the geometry of the hypermolecule is related to the geometry of the receptor binding site, and molecule steric affinity to the binding site is obtained by comparing geometry of the molecule and that of the hypermolecule. Moreover, to represent the active regions of the hypermolecule, vertices within the → *binding site cavity (cavity vertices)* are labeled with a parameter $\epsilon = -1$, vertices in the cavity walls (*wall vertices*) with $\epsilon = +1$, and vertices in the external part of the cavity, that is, in the aqueous solution, with $\epsilon = 0$.

In this way, **MTD descriptors** measuring the → *steric misfit* between the binding site cavity and the considered molecules are calculated as

$$MTD_i = c + \sum_{s=1}^S \epsilon_s \cdot I_{is}$$

where the subscript i refers to the considered molecule; c is the total number of cavity vertices of the hypermolecule and should be a measure of the volume of the binding site cavity; S is the total number of hypermolecule vertices; I_{is} is a binary variable for the s th hypermolecule vertex equal to 1 if the i th molecule occupies the s th vertex with an atom, otherwise it equals zero.

The MTD_i descriptor is a measure of steric misfit of the i th molecule with respect to the receptor cavity and is equal to the number of unoccupied cavity vertices plus the number of

occupied wall vertices; it can be considered both a $\rightarrow$ *steric descriptor* and a $\rightarrow$ *differential descriptor*.

For a molecule coincident with the active region of the hypermolecule, it can be observed that c atoms are located at the hypermolecule cavity vertices ($\epsilon = -1$), and no atoms coincide with the hypermolecule wall vertices, that is, there is no steric misfit and $MTD = 0$.

The **MTD model** is obtained by including MTD descriptors in the $\rightarrow$ *Hansch model*:

$$\hat{y}_i = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij} - b' \cdot MTD_i$$

where Φ are selected $\rightarrow$ *physico-chemical properties* of a Hansch-type model and the sign minus of the MTD coefficient indicates the detrimental contribution to the activity due to steric misfit.

The final optimal set of $\epsilon_1, \epsilon_2, \dots, \epsilon_S$ values (*receptor map*) is searched for according to an optimization procedure, starting from random or arbitrary assignment of hypermolecule vertices to cavity, wall, or external regions. To each set of ϵ_s values corresponds a set of MTD descriptors and a set of calculated responses. The optimization procedure is based on the maximization of the correlation of the biological responses calculated by the MTD model with the experimental responses.

The **MTD-MC method** is a modified version of the MTD method, accounting for the existence of several low-energy conformations of molecules used to derive the hypermolecule by overlap. Each molecule is described by a vector of binary variables $I_{i(k)s}$ equal to one if the s th vertex of the hypermolecule is occupied by the i th molecule in the k th conformation. If more than one low-energy conformation is allowed for a molecule, the conformations considered will be the one that best fits the binding site cavity, that is, the one with the lowest MTD value:

$$MTD_i = \min_k \left(c + \sum_{s=1}^S \epsilon_s \cdot I_{i(k)s} \right)$$

where the minimum is chosen over all of the considered conformations of the i th molecule.

A modification of the MTD approach that also acts on biological responses, called the **MTD-ADJ method**, was proposed to improve the performance of the modeling power of the method, accounting for the relative contribution of the active conformation to the activity of each compound [Sulea, Kurunczi *et al.*, 1998].

Let C_A be the concentration of active conformation A, the following relationships hold

$$C_A = \alpha_A \cdot C \quad y^{\text{exp}} = -\log C \quad y^{\text{adj}} = -\log(\alpha_A C) = y^{\text{exp}} - \log \alpha_A$$

where C is the total concentration, y^{exp} and y^{adj} the experimental and adjusted biological responses. The factor α_A for each conformation is calculated by the Boltzmann distribution:

$$\alpha_A = \frac{g_A \cdot \exp^{-E_A/RT}}{\sum_{k=1}^N g_k \cdot \exp^{-E_k/RT}}$$

where E are the calculated total conformational energies, g the degeneration degree of the conformational energy levels, and R and T the gas constant and the absolute temperature, respectively.

For each conformation of each compound, the corresponding *MTD* value and adjusted biological response are calculated. By using the optimization and $\rightarrow$ *validation techniques* for *MTD* model, each compound's conformation that best fits the adjusted response is retained and should be considered the active conformation of the compound. If more than one conformation is selected for the same compound, all these conformers have the same *MTD* values, while the adjusted response is calculated considering the conformer contributions to the total population, that is, $\Sigma\alpha$.

The *MTD-ADJ* method provides additional information concerning the active conformations of the compounds.

The **Minimal Steric Difference** (*MSD*) method is the first version of the *MTD* approach, based on the comparison of each molecule with the molecule with the highest biological activity in the data set, taken as the $\rightarrow$ *reference structure* instead of the hypermolecule. The assumption is that the most active molecule fits into the binding site best [Simon and Szabadai, 1973b].

MSD_i is a descriptor of steric misfit defined as the number of nonoverlapping non-hydrogen atoms for the maximal superimposition of the *i*th molecule to the reference molecule. *MSD* coincides with *MTD* when simple minimal steric differences are calculated with respect to the most active compound, no external atoms are considered in the hypermolecule, and the number of hypermolecule cavity vertices corresponds to the atoms of the reference molecule.

The **Monte Carlo version of MTD** (*MCD*) is a modification of the *MSD* approach, where the MCD_i descriptor is calculated as the nonoverlapping volume (NOV_i) of the *i*th molecule with respect to the reference molecule and the reference molecule with respect to the *i*th molecule [Motoc, Holban *et al.*, 1977].

The two superimposed molecules are included within a cube with volume *V* and a large number of points *N* is randomly dispersed within the cube. The **nonoverlapping volume** NOV_i of the *i*th molecule is calculated as

$$NOV_i = V \cdot \left(\frac{N_{REF} + N_i}{N} \right)$$

where N_{REF} is the number of points falling into the $\rightarrow$ *van der Waals molecular surface* of the reference molecule but not into that of the *i*th molecule, and N_i is the number of points falling into the van der Waals envelope of the *i*th molecule but not into that of the reference molecule; *N* is the total number of points randomly distributed throughout volume *V* of the cube.

A further modification of the *MTD* approach is called **Steric Interactions in Biological Systems** (*SIBIS*), where **attractive steric effects** (*SMDC*) and **repulsive steric effects** (*SMDW*) are considered separately [Motoc and Dragomir, 1981; Motoc, 1984a, 1984b]. Moreover, the optimization procedure searching for the receptor map, that is, the optimal set of ϵ_s values, is modified by the introduction of connectivity restrictions, where all the cavity vertices have to form a single topological connected network, that is, the receptor cavity is not fragmented into several subcavities.

The two steric contributions are defined as

$$SMDC_i = \sum_{s=1}^{S_{cav}} b'_{is} \cdot I_{is} \quad SMDW_i = \sum_{s=1}^{S_{wall}} b''_{is} \cdot I_{is}$$

where b_{is} are correction factors, accounting for the size of the atom of the i th molecule in the s th position of the hypermolecule; I_{is} is a binary variable for the s th hypermolecule vertex equal to 1 if the i th molecule occupies the s th vertex with an atom, and equal to zero otherwise. The first descriptor $SMDC$ is a sum running over all hypermolecule cavity positions S_{cav} ($\epsilon_s = -1$) and the second descriptor $SMDW$ a sum over all the hypermolecule cavity wall positions S_{wall} ($\epsilon_s = +1$).

Therefore, the **SIBIS model** is defined as

$$\hat{y}_i = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij} + b' \cdot SMDC_i - b'' \cdot SMDW_i$$

where Φ are selected physico-chemical properties of the Hansch model; the plus sign of the $SMDC$ coefficient indicates a favorable contribution (attractive steric effects) to activity and sign minus of the $SMDW$ coefficient a detrimental contribution (repulsive steric effects).

Moreover, similar to the $\rightarrow$ *Molecular Field Topology Analysis* (MFTA), a further variant of the MTD approach was proposed with the name **MTD-PLS method** [Oprea, Kurunczi *et al.*, 2001; Kurunczi, Olah *et al.*, 2002], intended as a simple (topologically based) $\rightarrow$ *scoring function* that could be useful in the absence of relevant receptor information. In particular, a “chemically intuitive” function is obtained forcing MTD - PLS coefficients to assume only negative (or zero) values for fragmental volume descriptors and positive (or zero) values for fragmental hydrophobicity descriptors.

📖 [Simon and Szabadai, 1973a; Simon, 1974, 1993; Simon, Holban *et al.*, 1976; Simon, Chiriac *et al.*, 1976; Simon, Badilescu *et al.*, 1977; Popoviciu, Holban *et al.*, 1978; Balaban, Chiriac *et al.*, 1980; Motoc, 1983b; Balaban *et al.*, 1985; Balaban, Niculescu-Duvaz *et al.*, 1987; Magee, 1991; Simon and Bohl, 1992; Ciubotariu, Deretey *et al.*, 1993; Oprea, Ciubotariu *et al.*, 1993; Fabian, Timofei *et al.*, 1995; Muresan, Bologa *et al.*, 1995; Sulea, Kurunczi *et al.*, 1995; Mracec, Mracec *et al.*, 1996; Timofei, Kurunczi *et al.*, 1996; Oprea, Kurunczi *et al.*, 1997; Polanski, 1997; Hadaruga, Muresan *et al.*, 1999; Ciubotariu, Grozav *et al.*, 2001; Ciubotariu, Gogonea *et al.*, 2001; Minailiuc and Diudea, 2001; Timofei, Kurunczi *et al.*, 2001; Thakur, Thakur *et al.*, 2004b; Mracec, Juchel *et al.*, 2006]

- **minimum–maximum path matrix** $\equiv$ *distance-detour combined matrix* $\rightarrow$ detour matrix
- **minimum path matrix** $\equiv$ *distance matrix*
- **Minkowski distance** $\rightarrow$ similarity/diversity (☉ Table S7)
- **Minoli–Bonchev complexity index** $\rightarrow$ molecular complexity
- **Minoli complexity index** $\rightarrow$ molecular complexity
- **MI-QSAR** $\equiv$ *Membrane Interaction QSAR analysis*
- **misclassification risk** $\rightarrow$ classification parameters
- **MIS indices** $\rightarrow$ molecular geometry
- **mixed CoMFA approach** $\rightarrow$ comparative molecular field analysis
- **mixed CoMFA model** $\rightarrow$ comparative molecular field analysis
- **MLOGP** $\rightarrow$ lipophilicity descriptors
- **MLSER** $\equiv$ *modified LSER* $\rightarrow$ Linear Solvation Energy Relationships
- **MmPS topological index** $\equiv$ *detour-Wiener combined index* $\rightarrow$ detour matrix
- **M/M quotient matrix** $\rightarrow$ biodescriptors (☉ DNA sequences)

- **MNA descriptors** $\equiv$ *Multilevel Neighborhoods of Atoms descriptors* $\rightarrow$ substructure descriptors ($\odot$ fingerprints)
- **MobyDigs software** $\rightarrow$ DRAGON descriptors
- **mode** $\rightarrow$ statistical indices ($\odot$ indices of central tendency)

■ model complexity

Model complexity is an important parameter to compare different QSAR/QSPR models. Moreover, the prediction power of a model is inversely related to its complexity, when complexity is unnecessary increased.

In general, model complexity is related to the number of variables selected for modeling purposes. Let $\mathbf{I}$ be the vector of length p , where p is the total number of variables, constituted of p binary variables. Each variable takes a value equal to zero ($I_j = 0$) if the j th variable is not in the model and a value equal to one ($I_j = 1$) if the j th variable is in the model.

The general problem of searching for the best set of variables ($\mathbf{I}^*$ vector) can be faced with by two different approaches: methods for $\rightarrow$ *variable reduction* and methods for $\rightarrow$ *variable selection*. The first group of methods allows selection of variables by inner relationships among the p variables $\mathbf{x}_j$:

$$\mathbf{I} = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$$

while the second group of methods by considering the relationships among the variables $\mathbf{x}_j$ and the response y to be modeled:

$$\mathbf{I} = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p; y)$$

In the first case, attention is paid to excluding variables carrying low or redundant information, in the second, to excluding variables, which are not functionally related to the studied response. In the latter, besides the exclusion of specific variables, one can condense the information from all the original variables into a few significant latent variables (linear combinations) by methods such as *Principal Component Regression* and *Partial Least Squares regression*.

The main measures of model complexity are reported below.

• number of terms in the model

The simplest definition of model complexity is based on the number of terms in the model or, in other words, the model complexity is made up by the number of model variables from Ordinary Least Squares regression ($cp\mathbf{x} = p$), the number M of significant principal components from Principal Component Regression ($cp\mathbf{x} = M$), and the number of significant latent variables from Partial Least Squares regression ($cp\mathbf{x} = M$).

• standardized regression coefficients sum

Model complexity is defined as the sum of standardized regression coefficients:

$$cp\mathbf{x} = \sum_{j=1}^p |b'_j| = \sum_{j=1}^p \left| \frac{b_j \cdot s_j}{s_y} \right|$$

where b'_j is the j th standardized regression coefficient, b_j the ordinary regression coefficient, s_y and s_j the standard errors of the response and j th variable, respectively, and p the total number of model variables.

• **information content ratio**

Model complexity is defined as the ratio between the **multivariate entropy** S_X of the data matrix $\mathbf{X}$ (n objects and p variables) of the model and $\rightarrow$ *Shannon's entropy* H_Y of the response vector $\mathbf{y}$, thus also accounting for the information content of the response $\mathbf{y}$ [Todeschini and Consonni, 2000]:

$$cpx = \frac{S_X}{H_Y} \quad 0 \leq cpx \leq \frac{p \cdot \log_2 n}{H_Y} \leq p$$

where H_Y and S_X are defined as

$$H_Y = - \sum_k \frac{n_Y}{n} \log_2 \frac{n_Y}{n} \quad S_X = [1 + (p-1)(1-K)] \cdot \frac{\sum_{j=1}^p H_j}{p}$$

where k runs on the different equivalence classes for $\mathbf{y}$ and n_Y is the number of equal y values; H_j is the Shannon entropy of the j th variable; p is the total number of variables and n is the total number of objects. K is the $\rightarrow$ *multivariate K correlation index*.

When all the y and x values (for each j th variable) are different (i.e., the Shannon's entropy of each variable is $\log_2 n$), the model complexity depends only on the total number p of model variables and the K correlation in the matrix $\mathbf{X}$:

$$cpx = \frac{[1 + (p-1)(1-K)] \cdot \frac{1}{p}(p \log_2 n)}{\log_2 n} = 1 + (p-1)(1-K)$$

Then, $cpx = p$ indicates the presence in the model of p perfectly uncorrelated x -variables, while cpx values lower than 1 indicate insufficient information in the $\mathbf{X}$ -block to completely model the $\mathbf{y}$ response.

- **model of the frontier steric effect** $\rightarrow$ steric descriptors ($\odot$ Taft steric constant)
- **model sum of squares** $\rightarrow$ regression parameters

■ **Mode of Action ($\equiv$ MOA)**

In toxicology, it is accepted that reliable QSARs can be attained only if toxicants are considered separately, depending on their mechanism of action (MOA), that is, only those chemicals showing a similar mode of action can be used together to search for a QSAR [Bradbury and Lipnick, 1990; Bradbury, 1994; Schüürmann, Segner *et al.*, 1997; Freidig and Hermens, 2000; Nendza and Müller, 2000].

For chemicals with the same MOA, similar structural features can be searched for assuming that they give rise to similar reactivity mechanisms. Then, the basic QSAR strategy provides for identifying critical structural elements responsible for activity via a hypothetical shared mode of action and then constructing QSAR models able to classify different modes of action.

Moreover, the mode of action is very important for mixtures of chemicals because the biological effect of mixtures can give rise to different kinds of toxicological response (antagonism, less than additive response, additive response, and synergism), depending on the toxicological mode of action and the chemical interactions of the substances involved [Calamari and Vighi, 1992; Gramatica, Vighi *et al.*, 2001]. As a consequence, in several cases, $\rightarrow$ *biological*

activity indices need to be studied for different classes of compounds, each having a different mode of action.

📖 [Lipnick, 1991; Gramatica, Vighi *et al.*, 2001; Musumarra, Condorelli *et al.*, 2001; Aptula, Netzeva *et al.*, 2002; Ren, 2002d; Pino, Giuliani *et al.*, 2003; Ren, 2003f; Schüürmann, Aptula *et al.*, 2003; Öberg, 2004a; Spycher, Nendza *et al.*, 2004; Chakraborty and Devakumar, 2005; Papa, Villa *et al.*, 2005; Spycher, Pellegrini *et al.*, 2005; Basak, Gute *et al.*, 2006]

- **modified edge-weighted Harary index** → weighted matrices (⊙ weighted adjacency matrices)
- **modified edge-weighted Harary matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **modified edge-Zagreb matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **modified Free–Wilson analysis** → Free–Wilson analysis
- **modified Hosoya index** → delocalization degree indices (⊙ Hosoya resonance energy)
- **modified Hosoya index** $\equiv$ *stability index* → characteristic polynomial-based descriptors
- **modified Hosoya index** $\equiv$ *Z* index* → Hosoya Z index
- **Modified Information Content index** → indices of neighborhood symmetry
- **modified LEACH index** → environmental indices (⊙ leaching indices)
- **Modified LSER** → Linear Solvation Energy Relationships
- **modified LUDI energy function** → scoring functions (⊙ LUDI energy function)
- **modified partial equalization of orbital electronegativities** → electronegativity
- **modified Randić index** → connectivity indices
- **modified spectrum-like descriptors** → spectrum-like descriptors
- **modified total adjacency index** → Zagreb indices
- **modified vertex Zagreb matrix** → vertex degree
- **modified weighted Tanimoto coefficient** → similarity/diversity
- **modified Wiener index** → Wiener index
- **modified Wiener indices** $\equiv$ *generalized Wiener indices* → Wiener index
- **modified Zagreb indices** → Zagreb indices
- **modulo compression algorithm** → substructure descriptors
- **modulo-L descriptors** → spectra descriptors
- **MOE descriptors** $\equiv$ *QuaSAR descriptors*
- **Mohar indices** → Laplacian matrix
- **molar polarization** → electric polarization descriptors
- **molar refractivity** → physico-chemical properties
- **molar refractivity partition index** → physico-chemical properties (⊙ molar refractivity)
- **molar volume** → volume descriptors
- **MolBlaster descriptors** → substructure descriptors (⊙ structural keys)

■ MolConn-Z descriptors

MolConn-Z is a software for the calculation of more than 400 molecular descriptors of different kinds [MolConn-Hall Associates Consulting, 1991; Faulon, Visco *et al.*, 2003].

MolConn-Z descriptors include valence, path, cluster, path/cluster, and chain molecular connectivity indices, kappa molecular shape indices, topological and electrotopological state indices, differential connectivity indices, Wiener and Platt indices, information

indices, counts of different vertices, and counts of paths and edges between different types of vertices, fragment counts, hydrogen-bonding descriptors, as well as a small set of 3D-descriptors.

📖 [Zheng and Tropsha, 2000; Tropsha and Zheng, 2002; Xiao, Xiao *et al.*, 2002; Bergström, Norinder *et al.*, 2003; Balaban, Basak *et al.*, 2004; Basak, Gute *et al.*, 2004; Kovatcheva, Golbraikh *et al.*, 2004; Medina-Franco, Golbraikh *et al.*, 2005; Wang, Li *et al.*, 2005]

- **MolDiA descriptors** → substructure descriptors (⊙ structural keys)
- **molecular branching** → molecular complexity
- **molecular centrality** → molecular complexity

■ molecular complexity

The concept of *molecular complexity* was introduced into chemistry only quite recently and is mainly based on the → *information content* of molecules. Several different measures of complexity can be obtained according to the diversity of the considered structural elements such as atom types, bonds, connections, cycles, and so on. The first attempts to quantify molecular complexity were based on the elemental composition of molecules; later other molecular characteristics were considered, such as the symmetry of molecular graphs, molecular branching, molecular cyclicity, and centrality [Bonchev, 1990; Bonchev and Seitz, 1996; Rücker and Rücker, 2000; Randić and Plavšić, 2002; Bonchev and Rouvray, 2003, 2007; Newman, 2004; Rücker, Rücker *et al.*, 2004; Bonchev and Buck, 2005].

A composite hierarchical concept of molecular complexity was proposed by Bonchev–Polansky [Bonchev and Polansky, 1987] according to their *general complexity scheme*, which begins with molecule size and proceeds through topology; molecules of the same size and topology are distinguished by their atom and bond types; moreover, a further discrimination is provided by geometric → *interatomic distances* and molecular symmetry.

In particular, topological complexity is hierarchically defined and the main features are *molecular branching*, *molecular cyclicity*, and *molecular centrality*.

Some reviews about molecular complexity are [Bonchev, 1999, 2003b; Barone and Chanon, 2001; Hann, Leach *et al.*, 2001; Cerruti, 2005; Schuffenhauer, Brown *et al.*, 2006].

• molecular branching

This is a molecule property comprising several structural variables such as number of branching, valence, distances apart, distances from the → *graph center*, and length of branches [Kirby, 1994]. Given this multifaceted definition of branching, its quantification is not an easy task. However, operational definitions of branching can be given by selected molecular indices, called **branching indices**, which, to some extent, reflect the branching of molecules as intended in an intuitive way [Randić, 1975d; Gutman and Randić, 1977; Bonchev, von Knop *et al.*, 1979; Barysz, von Knop *et al.*, 1985; Bertz, 1988; Rouvray, 1988b; Bonchev, 1995; Klein and Babic, 1997; Randić, 1997b; Gutman and Vidović, 2002b; Perdih, 2003].

The → *Wiener index* was the first proposed index of molecular branching [Bonchev and Trinajstić, 1977]; it is a function, inversely related to branching, of the number, length, and position of branches as well as of the number of atoms. For an isomeric series, it can be considered mainly dependent on molecular branching. Other specific molecular descriptors

proposed as measures of branching are the $\rightarrow$ Lovász–Pelikan index, $\rightarrow$ ramification index, $\rightarrow$ Zagreb indices, $\rightarrow$ $\lambda\lambda_1$ branching index, and $\rightarrow$ branching ETA index. The $\rightarrow$ Balaban centric index, the $\rightarrow$ Randić connectivity index, the $\rightarrow$ mean information content of the distance equality [Bonchev and Trinajstić, 1978], and the $\rightarrow$ Merrifield–Simmons index can also be used to discriminate among isomeric molecules with different branching patterns. Moreover, several branching indices were proposed by specific pairs of m and n values of $\rightarrow$ $v^m d^n$ matrices or a, b, c values of the $\rightarrow$ general distance–degree matrix $G(a, b, c)$ [Perdih, 2003], by summing all the off-diagonal elements, by calculating the largest eigenvalues and by the product of all the off-diagonal elements. In particular, the largest eigenvalue obtained for $a = b = -1/4$ and $c = -1$ was proposed as branching index.

The simplest but effective index of molecular branching is the **Bertz branching index**, defined as [Bertz, 1988; Ivanciuc, Ivanciuc *et al.*, 2000c]

$$BI = \frac{1}{2} \cdot \sum_{i=1}^A \delta_i \cdot (\delta_i - 1)$$

where the sum runs over all the atoms and δ_i is the $\rightarrow$ vertex degree of the i th atom; as it can be noted, contribution to branching of terminal atoms ($\delta_i = 1$) is zeroed. Note that this index can also be calculated from the $\rightarrow$ edge adjacency matrix as the $\rightarrow$ connection number.

• molecular cyclicity

It is another important feature of molecular complexity, defined in terms of number of molecule cycles and the manner in which the cycles are connected. It was first characterized by Bonchev [Bonchev, Mekenyan *et al.*, 1980b; Bonchev and Mekenyan, 1983] by a system of rules based on the number of atoms, the number of cycles, the number of atoms in a cycle, the number of cycles having a common edge, and so on. Moreover, a number of molecular descriptors, usually called $\rightarrow$ ring descriptors, were proposed to describe molecular cyclicity, accounting for the presence of cycles in molecules.

The $\rightarrow$ Wiener index was initially chosen as a single-valued molecular descriptor related to the cyclicity degree of isomeric molecules; it decreases for molecules with cyclic structures more complex than in molecules of the same size but with fewer rings. The $\rightarrow$ Harary index and $\rightarrow$ Kirchhoff number were also proposed as discriminating cyclicity indices [Bonchev, Balaban *et al.*, 1994]. Moreover, Randić [Randić, 1997a] defined the cyclicity of a molecule in terms of the cyclicity of the corresponding molecular graph “... as the departure of the cyclic character of the graph from that of the monocyclic graph relative to the departure of the complete graph from the monocyclic graph.” The cyclic character of the graph is given by the $\rightarrow$ D/Δ index calculated from the $\rightarrow$ distance/detour quotient matrix. In practice, the degree of cyclicity of a molecule with A atoms is calculated by the comparison of the corresponding molecular graph with the two extreme graphs of the same size (monocycle and $\rightarrow$ complete graph). The $\rightarrow$ cyclicity index γ is a quantitative measure of molecular cyclicity as defined by Randić.

Other cyclicity indices are the $\rightarrow$ total edge cyclicity, the $\rightarrow$ global cyclicity indices, the $\rightarrow$ spanning tree number, the $\rightarrow$ ring ID number, the $\rightarrow$ ring degree–distance index, and all the indices derived from $\rightarrow$ quotient matrices defined in terms of two different graph distance matrices.

- **molecular centrality**

It is considered less important than branching and cyclicity, but it contributes to the quantification of molecular complexity by distinguishing between molecular structures organized differently with respect to their centers [Bonchev, 1997]. → *Centric indices* are topological descriptors related to molecular centrality. Moreover, the → *centric topological index* and → *centrocomplexity topological index* calculated from the → *branching layer matrix* of an → *iterated line graph sequence* were proposed to measure molecular centrality [Diudea, Horvath *et al.*, 1992].

Molecular complexity indices are mainly based on → *information indices* defined to account for molecule complexity. These indices may be broadly divided into topological complexity indices and chemical complexity indices [Basak, 1987]. The former are calculated as the → *information content* of molecular graphs where atoms are not distinguished; among these, only those able to account for multiplicity in the graph are used to measure molecular complexity. The latter accounts for the chemical nature of the individual atoms in terms of bonding topology of weighted graphs or through the use of the → *physico-chemical properties* of the atoms in the molecule.

Some complexity indices are defined below [Nikolić, Trinajstić *et al.*, 2003]. Other molecular descriptors that give information about molecular complexity are reported elsewhere; these are the → *total walk count*, → *path counts*, → *spanning tree number* → *indices of neighborhood symmetry*, and the → *total adjacency index*. The latter was proposed by Bonchev and Polansky [Bonchev and Polansky, 1987] as a simple measure of topological complexity, being a measure of the degree of connectedness of molecular graph. Moreover, some among the → *GETAWAY descriptors* have been proposed to account at varying extents for molecular complexity.

- **Bertz complexity index (I_{CPX})**

The most popular complexity index was introduced by Bertz (Bertz, 1981, 1983a, 1983b) taking into account both the variety of bond connectivities and atom types of a → *H-depleted molecular graph*.

A general form of a molecular complexity index I_{CPX} is

$$I_{CPX} = I_{CPB} + I_{CPA}$$

where I_{CPB} and I_{CPA} are the → *information contents* related to the bond connectivity and the atom-type diversity, respectively. The term I_{CPB} was originally defined as

$$I_{CPB} \equiv C(TI) = 2 \cdot TI \cdot \log_2 TI - \sum_g TI_g \cdot \log_2 TI_g$$

where TI is any → *graph invariant*, and TI_g is the number of equivalent elements forming the graph invariant TI . The choice of the graph invariant should be based on the assumption that molecular complexity increases with size, branching, vertex and edge weights, and so on. The → *connection number* N_2 (also called → *Bertz branching index*, BI), that is, the number of bond pairs, was proposed by Bertz as a good choice for evaluating molecular complexity as it measures both the size and symmetry of the graph. Therefore, the two terms of the Bertz complexity index are defined as

$$I_{CPB} = 2 \cdot N_2 \cdot \log_2 N_2 - \sum_g (N_2)_g \cdot \log_2 (N_2)_g = N_2 \cdot \log_2 N_2 + {}^{CONN}I_{ORB}$$

$$I_{CPA} \equiv I_{AC} = A \log_2 A - \sum_g A_g \log_2 A_g$$

where $(N_2)_g$ is the number of symmetrically identical connections of type g ; A is the total number of atoms (hydrogen excluded); A_g is the number of atoms of the g th element, and ${}^{CONN}I_{ORB}$ is the $\rightarrow$ *total connection orbital information content*. The term I_{CPB} measures the complexity of a molecule given by the partition of equivalent connections sensitive to branching, rings, and multiple bonds of the molecule; when all the connections are the same, the bond complexity term is equal to $N_2 \log_2 N_2$ to take into account the size of the molecule, together with its symmetry in terms of bond connectivity. The atom complexity term I_{CPA} accounts for the presence of heteroatoms in the molecule and corresponds to the $\rightarrow$ *total information index on atomic composition* calculated for H-depleted molecular graphs.

📖 [Nikolić and Trinajstić, 2000]

- **Rashevsky complexity index** ($\bar{I}_{RASH}$)

This is a quantitative measure of graph complexity per vertex based on the sum of a chemical and a topological term:

$$\bar{I}_{RASH} = \bar{I}_{AC} + \bar{I}_{TOP}$$

where the two terms are the $\rightarrow$ *mean information index on atomic composition* $\bar{I}_{AC}$ and the $\rightarrow$ *topological information content* $\bar{I}_{TOP}$, respectively [Rashevsky, 1955]. Note that the topological information content proposed by Rashevsky is not based on graph orbits as is the most general topological information content later proposed by Trucco [Trucco, 1956a, 1956b]. In effect, two vertices v_i and v_j are considered topologically equivalent if for each k th neighboring vertex (k ranging between 1 and the $\rightarrow$ *atom eccentricity*) of vertex v_i , there exists a k th neighboring vertex of the same degree for vertex v_j .

The Rashevsky complexity index was further developed by Mowshowitz to obtain a measure of relative complexity of undirected and directed graphs [Mowshowitz, 1968a, 1968b, 1968c, 1968d].

- **Dosmorov complexity index** (I_{DOSM})

This is a molecular complexity index defined as a linear combination of five single information indices [Dosmorov, 1982]:

$$I_{DOSM} = I_{AC} + I_{at} + I_B + I_{SYM} + I_{CONF}$$

where I_{AC} is the $\rightarrow$ *total information index on atomic composition* calculated for H-depleted molecular graphs, I_{at} the $\rightarrow$ *atomic information content*, I_B the $\rightarrow$ *information bond index*, I_{SYM} the $\rightarrow$ *information index on molecular symmetry*, and I_{CONF} the $\rightarrow$ *information index on molecular conformations*. This combination of indices was proposed as a general index of molecular complexity accounting for the chemical nature of atoms, molecular size, number, and kind of molecular bonds, symmetry, and conformations. Moreover, by incorporating the atomic information content the Dosmorov index becomes more discriminating than the Bertz index in the case of different substituent atoms of the same valence [Bonchev, 1983].

- **Bonchev complexity information index** (I_{BONC})

This is a molecular descriptor defined by analogy with the Dosmorov complexity index, also accounting for electronic properties of molecules [Bonchev, 1983]:

$$I_{BONC} = I_{IC} + I_{NUCL} + I_{EL} + I_{Topology} + I_{SYM} + I_{CONF}$$

where I_{IC} is the $\rightarrow$ information index on isotopic composition, I_{EL} is one of the $\rightarrow$ electronic information indices; $I_{Topology}$ can be any topological information index accounting for structural complexity; I_{SYM} is the $\rightarrow$ information index on molecular symmetry, I_{CONF} the $\rightarrow$ information index on molecular conformations, and I_{NUCL} the $\rightarrow$ nuclear information content.

• Minoli complexity index

This is a measure of complexity of a molecular graph monotonically increasing on the number of vertices and edges, and reflecting the degree of connectedness of the graph [Minoli, 1976]. It is defined for undirect graphs, with no loops and multiple edges as

$$\chi = \frac{A \cdot B}{A + B} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A P_{ij} = \frac{AB}{A + B} \cdot \sum_{m=1}^L {}^mP$$

where A is the number of graph vertices, B the number of graph edges, P_{ij} the number of paths of any length between vertices v_i and v_j , ${}^mP$ the total number of paths of length m in the graph, and L the length of the longest path in the graph. ${}^mP$ are the elements of the $\rightarrow$ molecular path code.

The Bonchev variant [Bonchev, 1990] of the Minoli index, called **Minoli–Bonchev complexity index**, was defined by replacing the number of paths ${}^mP$ of length m with the products of the total number of paths by their length as

$$\chi_{MB} = \frac{A \cdot B}{A + B} \cdot \sum_{m=1}^L {}^mP \cdot m = \frac{A \cdot B}{A + B} \cdot W^{AP}$$

where W^{AP} is the $\rightarrow$ all-path Wiener index.

• Bertz–Herndon relative complexity index (C_{BH})

This is a simple measure of structural complexity of a molecule based on its graph representation G compared with the parent $\rightarrow$ complete graph $K(G)$, that is, the complete graph with the same number of vertices. It is defined as

$$C_{BH} = \frac{K_G}{K_{K(G)}}$$

where K is the total number of connected subgraphs in G and $K(G)$, respectively [Bertz and Herndon, 1986; Bonchev, 1997].

• Bonchev topological complexity indices ($\equiv$ topological complexity indices, TC, or overall topological indices, OI)

These are derived from the graph representation of molecules. A general overall topological index is formulated as [Bonchev, 1999, 2000, 2001a, 2001b; Bonchev and Trinajstić, 2001]

$$OI = \sum_{k=1}^K \mathcal{D}(G_k) = \sum_{k=1}^K f(\mathcal{L}_i(i \in G_k))$$

where the summation runs over all the connected subgraphs G_k of the molecular graph G , K being the total number of connected subgraphs of G , and $\mathcal{D}(G_k)$ is any $\rightarrow$ graph invariant derived from each k th subgraph and defined as some function f of the $\rightarrow$ local vertex invariants $\mathcal{L}_i$ of all vertices belonging to the k th subgraph.

The overall topological index OI is defined in two versions, depending on whether local vertex invariants L_i are those of the entire graph G or those of the subgraph G_k ; for the latter, the symbol OI1 was suggested instead of OI .

The m th order overall topological index, denoted by ${}^m\text{OI}$, is defined as the sum of the invariants $\mathcal{D}({}^mG_j)$ of all ${}^mK$ subgraphs ${}^mG_j$, which have m edges:

$${}^m\text{OI} = \sum_{j=1}^{{}^mK} \mathcal{D}({}^mG_j)$$

The overall topological vector $\text{OI}(G)$ of any graph G having B edges is the sequence of all ${}^m\text{OI}$ indices listed in ascending order:

$$\text{OI}(G) = ({}^0\text{OI}, {}^1\text{OI}, {}^2\text{OI}, \dots, {}^B\text{OI})$$

The m th order overall topological indices ${}^m\text{OI}_t$ of the t th class of subgraphs, t standing for subgraph type such as path, cluster, path cluster, cycles, and so on, were also defined by limiting the summation to the subgraphs of type t .

Moreover, the complexity vector $\mathbf{K}(G)$ of a graph G was analogously defined as

$$\mathbf{K}(G) = ({}^0K, {}^1K, {}^2K, \dots, {}^BK), \quad K = \sum_{m=0}^B {}^mK$$

where ${}^mK$ is the number of subgraphs having m edges.

Overall connectivity indices are overall topological indices derived from the $\rightarrow$ adjacency matrix of the molecular graph. The overall connectivity, denoted by TC , is defined as [Bonchev, 1999, 2000, 2001a, 2001b; Bonchev and Trinajstić, 2001]

$$TC = \sum_{m=0}^B {}^mTC = \sum_{k=1}^K A_V(G_k) = \sum_{k=1}^K \sum_{i=1}^{N_k} \delta_i (i \in G_k)$$

where $A_V(G_k)$ is the $\rightarrow$ total adjacency index of the k th subgraph G_k ; N_k is the number of vertices in the k th connected subgraph, and δ_i is the $\rightarrow$ vertex degree of the i th vertex of the subgraph. In practice, the vertex degree is assigned to each vertex in the molecular graph, and then for each connected subgraph the degrees of all subgraph vertices are added; this quantity is summed up over all subgraphs of the same order to generate the partial m th order overall connectivity ${}^mTC$, and, finally, all these partial overall connectivities are summed up to give the total overall connectivity TC . Note that the overall connectivity of zero order coincides with the total adjacency index of the entire molecular graph:

$${}^0TC \equiv A_V(G) = \sum_{k=1}^A \delta_k (k \in G)$$

A different definition of overall connectivities $TC1$ was given by considering the vertex degrees as they are in isolated subgraphs. The inequality $TC > TC1$ always holds. Moreover, **valence overall connectivity indices** TC^v were defined by using the $\rightarrow$ valence vertex degrees δ^v in place of the simple vertex degrees:

$$TC^v = \sum_{k=1}^K \sum_{i=1}^{N_k} \delta_i^v (i \in G_k)$$

Overall connectivity indices were proposed as a meaningful measure of topological complexity of molecules, since they satisfy two fundamental requirements to a complexity measure: to increase with both the number of structural elements and their interconnectedness; the basic idea is that “*The higher the connectivity of molecular graph and its connected subgraphs, the more complex the molecule*” [Bonchev and Trinajstić, 1977].

Overall Zagreb indices, denoted by OM2, were defined as an extension of $\rightarrow$ *Zagreb indices* as [Bonchev, 1997; Nikolić, Tolić *et al.*, 2000]

$$OM2 = \sum_{m=0}^B {}^mOM2$$

where ${}^mOM2$ is the m th order overall Zagreb index, which can be calculated either generally ${}^mOM2$ or separately for each type of subgraphs ${}^mOM2_t$:

$${}^mOM2 = \sum_{j=1}^{mK} \prod_{i=1}^{N_j} \delta_i (i \in G_j) \quad {}^mOM2_t = \sum_{j=1}^{mK_t} \prod_{i=1}^{N_j} \delta_i (i \in G_j)$$

where δ_i is the $\rightarrow$ *vertex degree* of the i th vertex belonging to the j th subgraph, having N_j vertices; ${}^mK$ is the total number of connected subgraphs having m edges, and ${}^mK_t$ is the total number of connected subgraphs of type t having m edges. Note that the overall Zagreb index of zero order is simply the sum of all the vertex degrees, thus resulting equivalent to the overall connectivity of zero order and hence the total adjacency index:

$${}^0OM2 \equiv {}^0TC \equiv A_v(G) = \sum_{k=1}^A \delta_k (k \in G)$$

Moreover, the first order overall Zagreb index coincides with the second Zagreb index M_2 :

$${}^1OM2 \equiv M_2 = \sum_{k=1}^B (\delta_i \cdot \delta_j)_k$$

Other overall indices, similar to overall Zagreb indices, were defined by using a different function that has terms inverse to those of OM2; these indices were denoted by ON and are defined as [Bonchev, 1999, 2000, 2001a, 2001b; Bonchev and Trinajstić, 2001]

$$ON = \sum_{m=0}^B {}^mON \quad {}^mON = \sum_{j=1}^{mK} \prod_{i=1}^{N_j} \delta_i^{-1} (i \in G_j) \quad {}^mON_t = \sum_{j=1}^{mK_t} \prod_{i=1}^{N_j} \delta_i^{-1} (i \in G_j)$$

where the difference from overall Zagreb indices is given by the use of reciprocal vertex degrees. The zero order overall index ON is

$${}^0ON = \sum_{k=1}^A \delta_k^{-1} (k \in G)$$

whereas the overall index of first order is the $\rightarrow$ *modified Zagreb index* ${}^mM_2$:

$${}^1\text{ON} \equiv {}^mM_2 = \sum_{k=1}^B (\delta_i \cdot \delta_j)_k^{-1}$$

The **overall Wiener indices** were analogously defined with the aim of extending the $\rightarrow$ *Wiener index* to its most complete version [Bonchev, 2001b]. They are defined as

$$\text{OW} = \sum_{m=0}^B {}^m\text{OW} \quad {}^m\text{OW} = \sum_{j=1}^{mK} \prod_{i=1}^{N_j} \sigma_i(i \in G_j) \quad {}^m\text{OW}_t = \sum_{j=1}^{mK_t} \prod_{i=1}^{N_j} \sigma_i(i \in G_j)$$

where σ_i is the $\rightarrow$ *vertex distance degree*, OW is the total overall Wiener index, ${}^m\text{OW}$ the m th order overall Wiener index, and ${}^m\text{OW}_t$ the m th order overall Wiener index restricted to those subgraphs having m edges and of type t . Note that the overall Wiener index of maximal order is the Wiener index: ${}^B\text{OW} = W$; the zero-order overall Wiener index equals zero: ${}^0\text{OW} = 0$; and the first-order overall Wiener index is equal to the number of graph edges: ${}^1\text{OW} = B$.

• **Bonchev-Trinajstić complexity index (BT)**

This is a complexity index defined in terms of $\rightarrow$ *information content*. The equivalence classes are defined collecting equal topological distances in the $\rightarrow$ *H-depleted molecular graph* [Bonchev and Trinajstić, 1977]. It is defined as

$$BT = \frac{A \cdot (A-1)}{2} \cdot \log_2 \frac{A \cdot (A-1)}{2} - \sum_{k=1}^D {}^kf \cdot \log_2 {}^kf$$

where A is the number of graph vertices, ${}^kf$ is the $\rightarrow$ *graph distance count* of k th order, and D the topological diameter.

• **Randić–Plavšić complexity index (ξ)**

The Randić–Plavšić complexity index or **ξ index** is based on the concept of the $\rightarrow$ *augmented valence*. The augmented valence AV of the i th vertex is obtained by adding to the $\rightarrow$ *vertex degree* the vertex degrees of all the other vertices in the graph, each weighted by a quantity that decreases as the distance from the vertex v_i increases [Randić, 2001b; Randić and Plavšić, 2002, 2003].

The Randić–Plavšić complexity index ξ is defined as the sum of augmented valences of all mutually nonequivalent vertices in the graph:

$$\xi = \sum_{i=1}^{A'} \text{AV}_i$$

where the sum is restricted to nonsymmetrical atoms A' .

- **molecular complexity indices** $\rightarrow$ molecular complexity
- **Molecular Connectivity Indices** $\rightarrow$ connectivity indices
- **molecular connectivity topochemical index** $\equiv$ *atomic molecular connectivity index* $\rightarrow$ connectivity indices
- **molecular cyclicity** $\rightarrow$ molecular complexity

- **molecular cyclized degree** → ring descriptors
- **Molecular Descriptor Family** → graph invariants
- **molecular descriptor properties** → molecular descriptors

■ **molecular descriptors** ($\equiv$ *chemical descriptors*)

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, and health researches, as well as in quality control, being the way molecules, thought of as real bodies, are *transformed* into numbers, allowing some mathematical treatment of the chemical information contained in the molecule [Todeschini and Consonni, 2000]. Molecular descriptors allow to find → *structure/response correlations* and perform → *similarity searching*, → *substructure searching*, and → *drug design*.

Therefore, molecular descriptors are formally mathematical representations of a molecule obtained by a well-specified algorithm applied to a defined *molecular representation* or a well-specified experimental procedure: *the molecular descriptor is the final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*.

The term “useful” has a double meaning: it means that the number can give more insight into the interpretation of the molecular properties and/or is able to take a part in a model for the prediction of some interesting molecular properties. Even if the interpretation of a molecular descriptor can be weak, provisional, or even completely lacking, it could be strongly correlated to some molecular properties to give models with high prediction power. On the other hand, descriptors with poor prediction power can be usefully retained in models when they are well theoretically founded and interpretable due to their ability to encode structural chemical information.

Although several molecular quantities were defined from the beginning of the quantum-chemistry and the graph theory, the term “molecular descriptor” has become popular with the development of structure–property correlation models. The → *Platt number* [Platt, 1947] and → *Wiener index* [Wiener, 1947c] defined in 1947 are sometimes referred to as the first molecular descriptors.

By the definition given above, the molecular descriptors are divided into two main classes: → *experimental measurements*, such as → *log P*, → *molar refractivity*, → *dipole moment*, *polarizability*, and, in general, → *physico-chemical properties*, and **theoretical molecular descriptors**, which are derived from a symbolic representation of the molecule and can be further classified according to the different types of *molecular representation*.

Mathematics and statistics, graph theory, computational chemistry, and molecular modeling techniques enable the definition of a large number of theoretical descriptors characterizing physico-chemical and biological properties, reactivity, shape, steric hindrance, and so on of the whole molecule, molecular fragments, and substituents.

The fundamental difference between theoretical descriptors and experimentally measured ones is that theoretical descriptors contain no statistical error due to experimental noise, which is not the case for experimental measurements.

However, the assumptions needed to facilitate calculation and numerical approximation are themselves associated with an inherent error, although in most cases the direction, but not the magnitude, of the error is known. Moreover, within a series of related compounds, the error term is usually considered to be approximately constant. All kinds of error are absent only for the

most simple theoretical descriptors such as count descriptors or for descriptors directly derived from exact mathematical theories such as $\rightarrow$ *graph invariants*.

Theoretical descriptors derived from physical and physico-chemical theories show some natural overlap with experimental measurements. Several $\rightarrow$ *quantum-chemical descriptors*, surface areas, and $\rightarrow$ *volume descriptors* are examples of such descriptors also having an experimental counterpart.

With respect to the experimental measurements, the greatest recognized advantages of the theoretical descriptors are usually (but not always) in terms of cost, time, and availability.

The **molecular representation** is the way that a molecule, that is, a phenomenological real body, is symbolically represented by a specific formal procedure and conventional rules. The quantity of chemical information, which is transferred to the molecule symbolic representation, depends on the kind of representation [Testa and Kier, 1991; Jurs, Dixon *et al.*, 1995].

The simplest molecular representation is the **chemical formula** (or **molecular formula**), which is the list of the different atom types, each accompanied by a subscript representing the number of occurrences of the atoms in the molecule. For example, the chemical formula of *p*-chlorotoluene is C_7H_7Cl , indicating the presence in the molecule of $A=8$ (number of atoms, hydrogen excluded), $N_C=7$, $N_H=7$, and $N_{Cl}=1$ (the subscript "1" is usually omitted in the chemical formula).

This representation is independent of any knowledge concerning the molecular structure, and hence molecular descriptors obtained from the chemical formula can be called **0D descriptors**. Examples are the $\rightarrow$ *atom number A*, $\rightarrow$ *molecular weight MW*, $\rightarrow$ *atom-type count N_x* , and, in general, $\rightarrow$ *constitutional descriptors* and any function of the $\rightarrow$ *atomic properties*.

The atomic properties are usually the $\rightarrow$ *weighting schemes* used to characterize molecule atoms; the most common atomic properties are atomic mass, $\rightarrow$ *atomic charge*, $\rightarrow$ *van der Waals radius*, $\rightarrow$ *atomic polarizability*, and $\rightarrow$ *atom electronegativity*. Atoms can also be characterized by the $\rightarrow$ *local vertex invariants* (LOVIs) derived from graph theory.

The **substructure list representation** can be considered as a one-dimensional representation of a molecule and consists of a list of structural fragments of a molecule; the list can be only a partial list of fragments, functional groups, or substituents of interest present in the molecule, thus not requiring a complete knowledge of the molecule structure. The descriptors derived by this representation can be called **1D-descriptors** and are typically used in $\rightarrow$ *substructural analysis* and $\rightarrow$ *substructure searching*.

The two-dimensional representation of a molecule considers how the atoms are connected, that is, it defines the connectivity of atoms in the molecule in terms of the presence and nature of chemical bonds. Approaches based on the $\rightarrow$ *molecular graph* allow a two-dimensional representation of a molecule, usually known as **topological representation**. Molecular descriptors derived from the algorithms applied to a topological representation are called **2D-descriptors**, that is, they are the so-called $\rightarrow$ *graph invariants*. In the last few years, several efforts have been made to formalize the several formulas and algorithms dealing with molecular graph information: "a graph operator applies a mathematical equation to compute a whole class of related molecular graph descriptors, using different molecular matrices and various weighting schemes. ... In this way, molecular graph operators introduce a systematization of topological indices and graph invariants by assembling together all descriptors computed with the same mathematical formula or algorithm, but with different parameters or molecular matrices" [Ivanciuc, 2000i].

Two-dimensional representations alternative to the molecular graph are the **linear notation systems**, for example, **Wiswesser Line Notation system** (WLN) [Smith and Baker, 1975], **SMILES** [Weininger, 1988, 1990, 2003; Weininger, Weininger *et al.*, 1989; Convard, Dubost *et al.*, 1994; Hinze and Welz, 1996], and **SMARTS** (SMART – Daylight Chemical Information Systems, 2004). **CAST** (*CAnonical representation of STereochemistry*) is a method that gives a linear notation that canonically represents stereochemistry around a specific site in a molecule [Sato, Koshino *et al.*, 2000, 2001, 2002].

The three-dimensional representation views a molecule as a rigid geometrical object in space and allows not only a representation of the nature and connectivity of the atoms, but also the overall spatial configuration of the molecule. This representation of a molecule is called **geometrical representation** and molecular descriptors derived from this representation are called **3D-descriptors**. Examples of 3D descriptors are the → *geometrical descriptors*, several → *steric descriptors*, and → *size descriptors*.

Several molecular descriptors derive from multiple molecular representations and can then be classified with difficulty. For example, graph invariants derived from a molecular graph weighted by properties obtained by → *computational chemistry* are both 2D and 3D descriptors.

The **bulk representation** of a molecule describes the molecule in terms of a physical object with 3D attributes such as bulk and steric properties, surface area, and volume.

The **stereoelectronic representation** (or **lattice representation**) of a molecule is a molecular description related to those molecular properties arising from electron distribution, interaction of the molecule with probes characterizing the space surrounding them (e.g., → *molecular interaction fields*). This representation is typical of the → *grid-based QSAR techniques*. Descriptors at this level can be considered **4D-descriptors**, being characterized by a scalar field, that is, a lattice of scalar numbers, associated with the 3D → *molecular geometry*.

Finally, the **stereodynamic representation** of a molecule is a time-dependent representation, which adds structural properties to the 3D representations, such as flexibility, conformational behavior, transport properties, and so on. → *Dynamic QSAR*, → *4D-Molecular Similarity Analysis*, and → *4D-QSAR Analysis* are examples of a multiconformational approach.

Within the two main classes of descriptors, experimental measurements and theoretical descriptors, several other subclasses of molecular descriptors can be recognized on the basis of a rational analysis of the molecular descriptor properties.

The main properties of the descriptors can be represented by a four-level taxonomy. Together with the first-level classification based on the molecular representation, as defined above, the other three levels are summarized below.

- **mathematical representation of molecular descriptors**

The descriptors can be represented by a scalar value, a vector, a two-way matrix, a tensor, or a scalar field, which can be discretized into a lattice of grid points.

- **invariance properties of molecular descriptors**

The invariance properties of molecular descriptors can be defined as the ability of the algorithm for their calculation to give a descriptor value that is independent of the particular characteristics of the molecular representation, such as atom numbering or labeling, spatial reference frame,

molecular conformations, and so on. Invariance to molecular numbering or labeling is assumed as a minimal basic requirement for any descriptor. **Chemical invariance** of a molecular descriptor means that its values are independent of the atom types and multiple bonds, that is, the descriptor is not able to account for heteroatoms and $\rightarrow$ *bond multiplicity* in the molecules. Such invariance is considered explicitly in classifying topological indices as $\rightarrow$ *topostructural indices* and $\rightarrow$ *topochemical indices*.

Two other important invariance properties, **translational invariance** and **rotational invariance**, are the invariance of a descriptor value to any translation or rotation of the molecules in the chosen reference frame. Molecular descriptors being invariant to translation and rotation of a molecule are referred to as **TRI descriptors**. These invariance properties have to be considered when dealing with descriptors derived from $\rightarrow$ *molecular geometry*. For all descriptors based on $\rightarrow$ *internal coordinates*, rototranslational invariance is naturally guaranteed. For descriptors based on spatial atomic coordinates, translational invariance is usually easily attained by centering the atomic coordinates; rotational invariance may be satisfied by using, as the reference frame, an univocally defined frame such as the principal axes of each molecule. In some QSAR methods, such as grid-based QSAR techniques, the problem of invariance to rotation is, at least in principle, overcome by adopting $\rightarrow$ *alignment rules*.

Conformational invariance means that molecular descriptor values are independent of the conformational changes in molecules. *Conformations* of molecules are the different atom dispositions in the 3D space, that is, configurations that flexible molecules can assume without any change to their connectivity. Usually interest in different conformations of a molecule is related to those conformations for which the total energy is relatively close to the minimum energy, that is, within a cutoff energy value of some kcal/mol.

Molecular descriptors can be distinguished according to their conformational invariance degree in four classes, as suggested by Charton [Charton, 1983]:

- (a) **No Conformational Dependence (NCD descriptors)**: This is typical of all descriptors, which do not depend on 3D molecular geometry, such as $\rightarrow$ *molecular weight*, $\rightarrow$ *count descriptors*, and $\rightarrow$ *topological indices*.
- (b) **Low Conformational Dependence (LCD descriptors)**: This is the case of molecular descriptors whose values show small variations only in the presence of relevant conformational changes, such as *cis/trans* configurations. Examples are $\rightarrow$ *cis/trans descriptors* and usually $\rightarrow$ *charge descriptors*.
- (c) **Intermediate Conformational Dependence (ICD descriptors)**: These are molecular descriptors whose values show small variations in the presence of any conformational changes. Typical descriptors of this class are $\rightarrow$ *EVA descriptors* and descriptors based on mass distribution, for example, $\rightarrow$ *radius of gyration*.
- (d) **High Conformational Dependence (HCD descriptors)**: This is the case of descriptors with values very sensitive to any conformational change in the molecule. Typical descriptors of this class are $\rightarrow$ *interaction energy values* obtained from $\rightarrow$ *molecular interaction fields*, $\rightarrow$ *3D-MoRSE descriptors*, $\rightarrow$ *WHIM descriptors*, $\rightarrow$ *G-WHIM descriptors*, $\rightarrow$ *spectrum-like descriptors*, $\rightarrow$ *shape descriptors* based on molecular geometries, and so on.

Among the $\rightarrow$ *quantum-chemical descriptors*, descriptors of different kinds of conformational dependence can be found: $\rightarrow$ *ionization potential*, $\rightarrow$ *electron affinity*, and molecular orbital

energies are often LCD or ICD descriptors, whereas molecular energies are usually HCD descriptors.

To quantify the conformational sensitivity of molecular descriptors, the **conformational pairwise sensitivity** (*CPS*) was proposed as the difference between conformationally dependent physico-chemical properties *P*, such as, for example, dipole moments, polar surface area, or 3D calculated log*P*, over the difference between geometry-dependent molecular descriptors *D* [Vistoli, Pedretti *et al.*, 2005]. This quantity is defined as

$$CPS_{st}(P, D) = \frac{|P_s - P_t|}{|D_s - D_t|}$$

where *s* and *t* are two different molecular conformations. The **conformational global sensitivity** of a descriptor (*CGS*) is then defined as the average of the pairwise sensitivities for all the possible pair combinations of *N* conformers:

$$CGS(P, D) = \frac{\sum_{st} CPS_{st}(P, D)}{N \cdot (N-1)}$$

It should be noted that some invariance properties such as invariance to atom numbering and rototranslations are mandatory for molecular descriptors used in QSAR/QSPR modeling; in several cases, chemical invariance is required, particularly when dealing with a series of compounds with different substituents; moreover, conformational invariance is closely dependent on the considered problem.

• degeneracy of molecular descriptors

This property refers to the ability of a descriptor to avoid equal values for different molecules. In this sense, descriptors can show no degeneracy at all (N), low (L), intermediate (I), or high (H) degeneracy. The degree of degeneracy of a descriptor can naturally be measured by $\rightarrow$ *Shannon's entropy*. Moreover, the degree of degeneracy depends on the molecules present in the considered data set. Suitable measures of molecular descriptor degeneracy can be provided by using a data set consisting of an extended hydrocarbon series as well as heteroatoms and cycles.

Information content and Shannon's entropy of molecular descriptors were extensively studied by Bajorath, Godden, and coworkers in several papers [Godden, Stahura *et al.*, 2000; Godden and Bajorath, 2000, 2002, 2003].

Degeneracy of 735 molecular descriptors as well as their pairwise correlations were estimated on the NCI database for 221,860 compounds and made available on a software module called Molecular Descriptor Correlations (MDC) [MDC – Milano Chemometrics, 2006].

Degeneracy is considered an undesirable characteristic for all molecular descriptors that are used for the characterization of molecules in store and retrieval database systems; however, in QSAR modeling, degenerate properties are better modeled by molecular descriptors showing analogous degeneracy [Todeschini, Consonni *et al.*, 1998].

Based on the previous criteria, examples of an indicative classification of molecular descriptors are shown in Table M10.

Table M10 Mathematical properties of some molecular descriptors.

Descriptors	Molecular representation	Mathematical representation	Invariance properties	Degeneracy
Molecular weight	0D	Scalar	NCD	H
Atom-type counts	0D	Scalar	NCD	H
Fragment counts	1D	Scalar	NCD	H
Topological information indices	2D	Scalar	NCD	L/I
Molecular profiles	2D	Vector	NCD	N
2D autocorrelation descriptors	2D	Vector	NCD	N/L
3D autocorrelation descriptors	3D	Vector	MCD	N
Substituent constants	3D	Scalar	NCD/LCD	L/I
WHIM descriptors	3D	Vector	HCD	N
3D-MoRSE descriptors	3D	Vector	LCD/MCD	N
GETAWAY descriptors	3D	Vector	MCD	N
Surface/volume descriptors	3D	Scalar	HCD/MCD	L
Quantum-chemical descriptors	3D	Scalar	MCD/HCD	N/L
Compass descriptors	3D	Vector	HCD	N
Interaction energy values	4D	Lattice	HCD/RD	N
GRIND descriptors	4D	Vector	HCD	N

Suitable molecular descriptors, besides the trivial invariance properties, should satisfy some basic requirements. The list of desirable requirements of chemical descriptors suggested by Randić [Randić, 1996a] is shown in Table M11.

Table M11 List of desirable requirements for molecular descriptors.

#	Descriptors
1	Should have structural interpretation
2	Should have good correlation with at least one property
3	Should preferably discriminate among isomers
4	Should be possible to apply to local structure
5	Should be possible to generalize to “higher” descriptors
6	Descriptors should be preferably independent
7	Should be simple
8	Should not be based on properties
9	Should not be trivially related to other descriptors
10	Should be possible to construct efficiently
11	Should use familiar structural concepts
12	Should have the correct size dependence
13	Should change gradually with gradual change in structures

• transformations of molecular descriptors

Some formal definitions of molecular descriptors are only suitable for small molecules, but for big molecules, they take so large values that they are not algorithmically manageable. In these cases, a proper transformation should be applied; the most common transformations are

logarithmic transformations, such as

$$\mathcal{D}' = \log(\mathcal{D}) \quad \text{or} \quad \mathcal{D}' = \ln(\mathcal{D}) \quad \text{or} \quad \mathcal{D}' = \log(1 + \mathcal{D}) \quad \text{or} \quad \mathcal{D}' = \ln(1 + \mathcal{D})$$

Another common transformation of molecular descriptors is their conversion into binary descriptors [Xue, Godden *et al.*, 1999a, 2003b], that is, each descriptor represented by continuous or integer numbers is reduced to a single binary value (0 or 1) or, in some cases, a series of binary values. A simple way to assign 0 or 1 values to the descriptor in place of the actual descriptor value is by comparing the descriptor value with a cutoff value; 0 or 1 values are assigned to those descriptor values that fall either below or above the cutoff value. The cutoff can be, for example, the median of the descriptor estimated from the data set or the chemical library.

Transformations of a set of molecular descriptors are often performed when there is the need of a $\rightarrow$ *variable reduction* or the need to modify binary vectors, such as site and substituent-oriented variables, into real-valued variable vectors. The milestone of these techniques is the $\rightarrow$ *Principal Component Analysis* (PCA), but also $\rightarrow$ *Fourier analysis* and $\rightarrow$ *Wavelet analysis* are often used, especially for $\rightarrow$ *spectra descriptors* compression.

Fourier analysis was, for instance, applied to change site and substituent-oriented binary variables in the $\rightarrow$ *Free-Wilson analysis*, into a few latently dependent real coefficients [Holik and Halamek, 2002].

Wavelet analysis was also proposed for variable reduction problems and, in particular, the wavelet coefficients obtained from *discrete wavelet transforms* (DWT) were proposed as a molecular representation in $\rightarrow$ *PEST descriptor methodology* and their sums as molecular descriptors [Breneman, Sundling *et al.*, 2003; Lavine, Davidson *et al.*, 2003].

Another interesting tool to obtain linear combinations of descriptors is that based on the **Andrews' curves** [Andrews, 1972]. The Andrews' curves are a pictorial way to represent and compare multivariate objects. In this representation of a p -dimensional space, each i th object is represented by a mathematical function as

$$f_i(t) = x_{i1}/\sqrt{2} + x_{i2} \cdot \sin(t) + x_{i3} \cdot \cos(t) + x_{i4} \cdot \sin(2t) + x_{i5} \cdot \cos(2t) + \dots$$

plotting this function over the range $-\pi \leq t \leq \pi$. The length of the $f_i(t)$ vector depends on the resolution by which the range is spanned by the t parameter.

A set of molecules will thus appear as a set of curves drawn across the plot (Figure M2). This transformation preserves the means, the variances and the Euclidean distances calculated from the original variables x . However, the curves are not invariant with respect to the order of the variables: in general, the low frequencies (x_1, x_2, x_3) are distinguished more readily on the plot than the high frequencies (x_p, x_{p-1}, x_{p-2}). For this reason, it is better to associate the most important variables with the low frequencies [Todeschini, Consonni *et al.*, 1998].

• susceptibility of molecular descriptors

The susceptibility of a molecular descriptor $\mathcal{D}$ is defined according to the following expression [Perdih and Perdih, 2002e, 2003d]:

$$S_{a,b} = \frac{\mathcal{D}_b}{\mathcal{D}_a} - 1$$

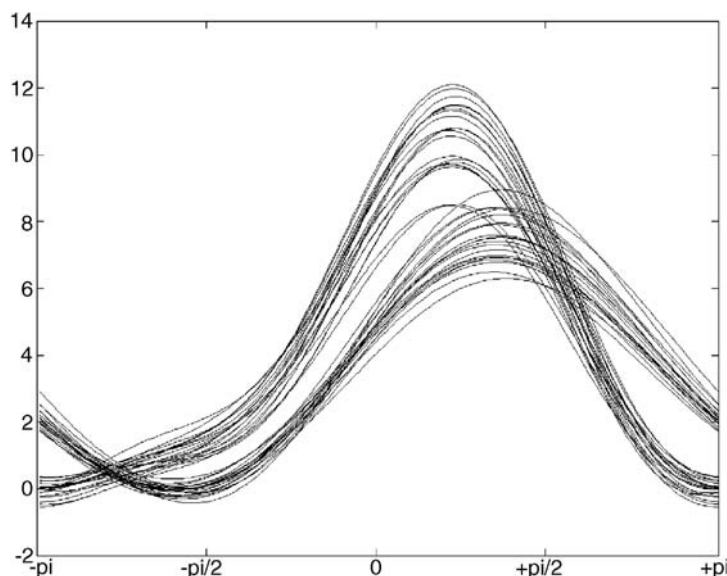


Figure M2 Andrews' plot of 40 samples described by four variables.

where subscript *a* refers to a reference structure with respect to some molecule structural feature, whereas *b* refers to a reference structure with “opposite” characteristic with respect to the studied structural feature.

Susceptibilities were defined for alkanes for the increase in carbon number and branching; moreover, differential susceptibilities were proposed to highlight the contribution of the number of branches, the position of branches, the separation between branches, and the change of the substituent from methyl to ethyl [Perdih and Perdih, 2003d].

Susceptibilities can be evaluated also for the influence of heteroatoms assuming benzene (or *n*-alkane) as the reference molecule and heteroatom-substituted benzenes (or heteroatom-substituted alkanes).

• comparisons of molecular descriptors

Comparisons among molecular descriptors and among different classes of descriptors are important for at least two reasons: (a) to enhance the comprehension of the chemical meaning of complex descriptors by comparing them with other more interpretable descriptors; (b) to evaluate their different prediction ability relatively to the different kinds of response to be modeled.

Together with the comparison of the → *degeneracy of molecular descriptors*, several papers contain pairwise correlation tables of molecular descriptors as well as extended discussions about different classes of descriptors. These are: [Duperray, Chastrette *et al.*, 1976a; Hall and Kier, 1978a; Todeschini, Cazar *et al.*, 1992; Moriguchi, Hirono *et al.*, 1994; Basak, Gute *et al.*, 1996c; Mannhold and Dross, 1996; Das, Dömötör *et al.*, 1997; Dearden and Ghafourian, 1999; Viswanadhan, Ghose *et al.*, 2000; Clare, 2002; Consonni, Todeschini *et al.*, 2002b; Cruciani, Pastor *et al.*, 2002; Zissimos, Abraham *et al.*, 2002c; Doweyko, 2004; Asikainen, Ruuskanen

et al., 2005; Fechner, Paetz *et al.*, 2005; Hollas, 2005b, 2005c; Perdih, 2000b, 2000c; Quigley and Nauhton, 2002; Gedeck, Rohde *et al.*, 2006].

Molecular descriptors are usually classified into several classes by a mixed taxonomy based on different points of view. For example, descriptors are often distinguished by their *physico-chemical meaning* such as $\rightarrow$ *electronic descriptors*, $\rightarrow$ *steric descriptors*, $\rightarrow$ *lipophilicity descriptors*, $\rightarrow$ *hydrogen-bonding descriptors*, $\rightarrow$ *shape descriptors*, $\rightarrow$ *charge descriptors*, $\rightarrow$ *electric polarization descriptors*, and $\rightarrow$ *reactivity descriptors*; moreover, on the basis of the specific mathematical tool used for the calculation of the molecular descriptors, $\rightarrow$ *autocorrelation descriptors*, $\rightarrow$ *spectral indices*, $\rightarrow$ *determinant-based descriptors*, $\rightarrow$ *Wiener-type indices*, $\rightarrow$ *Schultz-type indices*, $\rightarrow$ *characteristic polynomial-based descriptors*, $\rightarrow$ *connectivity-like indices*, and $\rightarrow$ *Balaban-like indices* can be distinguished.

■ [Duperray, Chastrette *et al.*, 1976a; Jurs, Chou *et al.*, 1979; Jurs, Stouch *et al.*, 1985; Lavenhar and Maczka, 1985; Mekenyan and Bonchev, 1986; Jurs, Hasan *et al.*, 1988; Dearden, 1990; Govers, 1990; Randić, 1990b, 1991a, 1991c; Silipo and Vittoria, 1990; Weininger and Weininger, 1990; Ash, Warr *et al.*, 1991; Cronin, 1992; Horvath, 1992; Bonchev, Mountain *et al.*, 1993; Katritzky and Gordeeva, 1993; Randić and Trinajstić, 1993b; Rücker and Rücker, 1993; Dearden, Cronin *et al.*, 1995b; Basak, Gute *et al.*, 1996c, 1997, 1998b; Karelson, Lobanov *et al.*, 1996; Balaban, 1997a; Basak, Grunwald *et al.*, 1997; Klein and Babić, 1997; Matter, 1997; Gasteiger, 1998; Lee, Park *et al.*, 1998; Baumann, 1999; Jalali-Heravi and Parastar, 1999; Andersson, Sjöström *et al.*, 2000; Godden, Stahura *et al.*, 2000; Karelson, 2000; Randić and Basak, 2000b; Todeschini and Consonni, 2000; Vidal, Thormann *et al.*, 2005; Todeschini, 2006]

■ molecular distance-edge vector

The Molecular Distance-Edge vector (or **MDE vector**), denoted by λ , was proposed as a molecular 10-dimensional vectorial descriptor based on the geometric means of the topological distances between carbon atoms of predefined type [Liu, Cao *et al.*, 1998; Liu, Liu *et al.*, 1999]; the four types of carbon atoms were classified simply as primary carbon C_1 (three bonded hydrogens), secondary C_2 (two bonded hydrogens), tertiary C_3 (one bonded hydrogens), and quaternary C_4 (no bonded hydrogens). The single elements are defined as

$$\lambda_{uv} = \frac{n_{uv}}{\bar{d}_{uv}^2} \quad u = 1, 2, 3, 4; \quad v \geq u$$

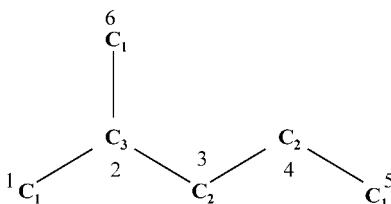
where

$$\bar{d}_{uv} = \prod_{u \leq v} (d_{i(u)j(v)})^{1/(2n_{uv})}$$

The geometric mean takes into account all the $\rightarrow$ *topological distances* between carbon atoms i and j of types u and v ; n_{uv} is the number of possible atom pairs for a fixed combination of carbon types. λ_{uv} is set at zero by definition if no atom pairs with types u and v are present in the molecular graph.

Example M2

MDE vector for 2-methylpentane.



C ₁ C ₁	C ₁ C ₂	C ₁ C ₃	C ₁ C ₄	C ₂ C ₂	C ₂ C ₃	C ₂ C ₄	C ₃ C ₃	C ₃ C ₄	C ₄ C ₄
$d_{16} = 2$	$d_{13} = 2$	$d_{12} = 1$	—	$d_{34} = 1$	$d_{32} = 1$	—	—	—	—
$d_{15} = 4$	$d_{14} = 3$	$d_{62} = 1$			$d_{42} = 2$				
$d_{65} = 4$	$d_{63} = 2$	$d_{52} = 3$							
	$d_{64} = 3$								
	$d_{53} = 2$								
	$d_{54} = 1$								

$$\begin{aligned} \bar{d}_{C_1C_1} &= (2 \cdot 4 \cdot 4)^{1/(2 \cdot 3)} = 32^{1/6} = 1.7818 & \lambda_{C_1C_1} &= 3/3.1748 = 0.9449 & \lambda_{C_1C_2} &= 6/2.0398 = 2.9415 \\ \bar{d}_{C_1C_2} &= (2 \cdot 3 \cdot 2 \cdot 3 \cdot 2 \cdot 1)^{1/(2 \cdot 6)} = 72^{1/12} = 1.4282 & \lambda_{C_1C_3} &= 3/1.4422 = 2.0802 & \lambda_{C_1C_4} &= 0 \\ \bar{d}_{C_1C_3} &= (1 \cdot 1 \cdot 3)^{1/(2 \cdot 3)} = 3^{1/6} = 1.2009 & \lambda_{C_2C_2} &= 1/1 = 1 & \lambda_{C_2C_3} &= 2/1.4142 = 1.4142 \\ \bar{d}_{C_2C_2} &= 1^{1/2} = 1 & \lambda_{C_2C_4} &= 0 & \lambda_{C_3C_3} &= 0 \\ \bar{d}_{C_2C_3} &= (1 \cdot 2)^{1/(2 \cdot 2)} = 2^{1/4} = 1.1892 & \lambda_{C_3C_4} &= 0 & \lambda_{C_4C_4} &= 0 \end{aligned}$$

$$\lambda = (0.9449, 2.9451, 2.0802, 0, 1, 1.4142, 0, 0, 0, 0)$$

- **Molecular Diversity Analysis descriptor** $\equiv$ *MolDiA descriptor* $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **molecular eccentricity** $\rightarrow$ shape descriptors
- **molecular electronegativity** $\rightarrow$ quantum-chemical descriptors ($\odot$ electronic chemical potential)
- **Molecular Electronegativity Distance Vector** $\rightarrow$ MEDV-13 descriptor
- **Molecular Electronegativity Edge Vector** $\rightarrow$ autocorrelation descriptors
- **Molecular Electrostatic Potential** $\rightarrow$ quantum-chemical descriptors
- **molecular electrostatic potential contour surface** $\rightarrow$ molecular surface
- **molecular energies** $\rightarrow$ quantum-chemical descriptors

■ Molecular Field Topology Analysis (MFTA)

The method of Molecular Field Topology Analysis is among the $\rightarrow$ *hyperstructure-based QSAR techniques*. It was proposed as a “topological analogue” of the $\rightarrow$ *CoMFA* method because it is based on topological rather than spatial alignment of structures [Zefirov, Palyulin *et al.*, 1997; Palyulin, Radchenko *et al.*, 2000; Melnikov, Palyulin *et al.*, 2007].

The quantitative description of structural features is provided by local physico-chemical parameters. First, for a set of structures of known activity (a training set), the so-called $\rightarrow$ *molecular supergraph* (MSG) is automatically constructed. The MSG is a certain graph, such that each training set structure can be represented as its subgraph. It enables the construction of $\rightarrow$

uniform-length descriptors for all structures in the set. To build each descriptor vector, the MSG vertices and edges corresponding, respectively, to the atoms and bonds of a given structure are assigned the values of local descriptors (e.g., atomic charge q and $\rightarrow$ van der Waals radius R^{vdw}) and the remaining vertices and edges are labeled with neutral descriptor values that provide a reasonable simulation of properties in an unoccupied region of space. The descriptor vector formation is illustrated in Figure M3.

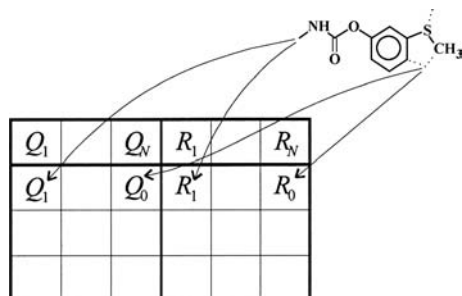


Figure M3 Descriptor vector formation in MFTA.

The following local descriptors are currently calculated: Gasteiger's $\rightarrow$ atomic charge q estimated by the $\rightarrow$ partial equalization of orbital electronegativity approach, Sanderson's $\rightarrow$ electronegativity χ^S , Bondi's van der Waals radius R^{vdw} , atomic contribution to the $\rightarrow$ van der Waals molecular surface SA^{vdw} , relative steric accessibility defined as $Ac = SA^{vdw}/SA_{free}$ (where SA_{free} is the van der Waals surface of the "free" (isolated) atom of the same type), $\rightarrow$ electrotopological state S , atomic lipophilicity contribution l taking into account the environment of an atom, and group lipophilicity L_g defined as a sum of contributions for both a non-hydrogen atom and attached hydrogens, the ability of an atom in a given environment to be a donor and acceptor of a hydrogen bond characterized by the binding constants, local stereochemical indicator variables, and the site occupancy factors for atoms I_a and bonds I_b (which have the value 1, if a given feature is present in the structure and 0 otherwise). This set of local descriptors provides sufficient coverage of major interaction types that are important for the interaction of a ligand with a biological target. However, the set is open and can be easily extended to account for the specific features of the problem.

Since the number of descriptors is rather large, partial least squares (PLS) regression is used to analyze the descriptor-activity relationships. As a result, the quantitative characteristic of the influence on activity of each descriptor in each position, including common structural fragments, can be determined (Figure M4). Such characteristics provide a basis for designing new, potentially more active structures as well as being anchor points for spatial structure alignment.

MFTA often gives models that are comparable in quality of description and prediction to models based on the widely used classical QSAR methods and 3D approaches.

- **molecular fingerprints** $\rightarrow$ substructure descriptors
- **molecular flexibility** $\rightarrow$ flexibility indices
- **molecular flexibility number** $\rightarrow$ flexibility indices

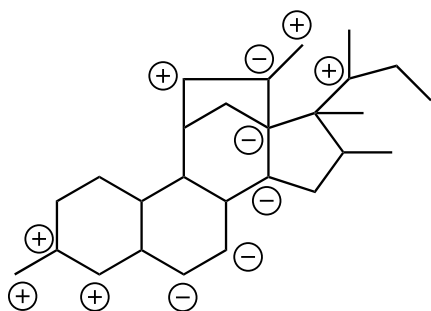


Figure M4 Molecular supergraph of steroid data set with major atomic charge contributions into the CBG affinity.

- **molecular formula** $\equiv$ *chemical formula* $\rightarrow$ molecular descriptors
- **molecular fragments** $\rightarrow$ count descriptors

■ molecular geometry

A molecule is the smallest fundamental group of atoms of a chemical compound that can take part in a chemical reaction. The atoms of the molecule are organized in a 3D structure; the **molecular matrix**, denoted by **M**, is a rectangular matrix $A \times 3$ whose rows represent the molecule atoms and the columns the atom **Cartesian coordinates** (x, y, z) with respect to any rectangular coordinate system with axes X, Y, Z . The Cartesian coordinates of a molecule usually correspond to some optimized molecular geometry obtained by the methods of $\rightarrow$ *computational chemistry*. The molecular geometry can also be obtained from crystallographic coordinates or 2D–3D automatic converters.

The **connectivity table** of a molecule is a rectangular table whose rows represent atoms and row entries the labels of all the bonded atoms.

Since the molecular matrix does not contain information on atom adjacencies, it usually is given as an augmented matrix **M'**, obtained by union of the molecular matrix and the connectivity table, where the first column denotes the atom type (e.g., carbon, hydrogen, chlorine atoms) and the last four columns contain the labels of the atoms connected to the i th atom:

$$\mathbf{M} = \begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \dots & \dots & \dots \\ x_A & y_A & z_A \end{vmatrix} \quad \mathbf{M}' = \begin{vmatrix} \text{at. 1} & x_1 & y_1 & z_1 & c_{11} & c_{12} & c_{13} & c_{14} \\ \text{at. 2} & x_2 & y_2 & z_2 & c_{21} & c_{22} & c_{23} & c_{24} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{at. A} & x_A & y_A & z_A & c_{A1} & c_{A2} & c_{A3} & c_{A4} \end{vmatrix}$$

Note that the last four columns of the **M'** matrix constitute the connectivity table of the molecules.

An alternative to the molecular matrix representation of a molecule is that of **internal coordinates**, where the relative position of each atom to the other atoms in the molecule is given: these coordinates are bond distances, bond angles, and torsion angles. **Bond distances** r_{st} are the interatomic distances between bonded atoms (usually expressed in Ångström); **bond angles** ϑ_{stv} are plane angles among triples of connected atoms (s, t, v) within the molecule; **torsion angles** ω_{stvz} are dihedral angles among quadruples of connected atoms (s, t, v, z) (Figure M5). Note that

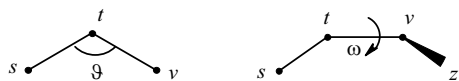


Figure M5 Bond and torsion angles.

bond distances and bond angles are less sensitive to conformational change than interatomic distances and torsion angles.

Internal coordinates are collected in the so-called **Z-matrix**, which is a rectangular matrix, whose rows are the atoms, defined as

$$\mathbf{Z} = \begin{array}{cccccc} \text{at. 1} & & & 0 & 0 & 0 \\ \text{at. 2} & r_{12} & & 1 & 0 & 0 \\ \text{at. 3} & r_{23} & \vartheta_{321} & 2 & 1 & 0 \\ \text{at. 4} & r_{34} & \vartheta_{432} & \omega_{4321} & 3 & 2 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{at. A} & r_{As} & \vartheta_{Ast} & \omega_{Astv} & s & t & v \end{array}$$

where r , ϑ , and ω are the molecular internal coordinates considered among the $\rightarrow$ *geometrical descriptors*. The last three columns contain the labels of atoms involved in bonds, bond angles, and torsion angles.

Other simple geometrical descriptors are **interatomic distances** r_{st} between pairs of atoms s and t . Interatomic distances are distinguished into **intramolecular interatomic distances**, that is, distances between any pair of atoms (s , t) within the molecule and **intermolecular interatomic distances**, that is, distances between atoms of a molecule and atoms of a receptor structure, a reference compound or another molecule. While classical computational chemistry describes molecular geometry in terms of three-dimensional Cartesian coordinates or internal coordinates, the $\rightarrow$ *distance geometry* (DG) method takes the interatomic distances as the fundamental coordinates of molecules, exploiting their close relationship to experimental quantities and molecular energies. Moreover, for series of congeneric compounds, bond distances, optimized by quantum-chemistry approaches and selected by genetic algorithms, were directly used as molecular descriptors in QSAR studies [Smith and Popelier, 2004].

The molecular matrix **M** and the matrix **Z** are the natural starting point for the calculation of several 3D atomic and molecular descriptors, such as $\rightarrow$ *quantum-chemical descriptors*, $\rightarrow$ *molecular interaction fields*, $\rightarrow$ *EVA descriptors*, $\rightarrow$ *WHIM descriptors*, $\rightarrow$ *GETAWAY descriptors*, $\rightarrow$ *CoMMA descriptors*, $\rightarrow$ *Compass descriptors*, and $\rightarrow$ *molecular surface descriptors*.

Another common source of geometrical descriptors is the geometry matrix.

The **geometry matrix** (or **geometric distance matrix**) of a molecule, denoted by **G**, obtained from the molecular matrix **M**, is a square symmetric matrix $A \times A$, where each entry r_{st} is the **geometric distance** calculated as the $\rightarrow$ *Euclidean distance* between the atoms s and t :

$$\mathbf{G} \equiv \begin{array}{cccc} 0 & r_{12} & \dots & r_{1A} \\ r_{21} & 0 & \dots & r_{2A} \\ \dots & \dots & \dots & \dots \\ r_{A1} & r_{A2} & \dots & 0 \end{array}$$

Diagonal entries are always zero. Geometric distances are intramolecular interatomic distances.

Like the molecular matrix, the geometry matrix contains information about molecular configurations and conformations; however, the geometry matrix does not contain information about atom connectivity. Thus, for several applications, it is accompanied by a connectivity table where, for each atom, there is listed the identification number of the atoms bonded to it. The geometry matrix can also be calculated on geometry-based standardized bond lengths and bond angles and derived by embedding a graph on a regular two-dimensional or three-dimensional grid; in these cases, the geometry matrix is often referred to as the **topographic matrix** **T** and the interatomic distance to the **topographic distance** [Balaban, 1997a]. Depending on the kind of grid used for graph embedding, different topographic matrices can be obtained.

The **bond length-weighted adjacency matrix** is obtained from the geometry matrix **G** as [Mihalić, Nikolić *et al.*, 1992]

$${}^b\mathbf{A} = \mathbf{G} \otimes \mathbf{A}$$

where $\otimes$ indicates the $\rightarrow$ *Hadamard matrix product* and **A** is the $\rightarrow$ *adjacency matrix*.

From the geometry matrix used to represent a $\rightarrow$ *molecular graph*, a number of $\rightarrow$ *local vertex invariants* and related $\rightarrow$ *graph invariants*, called **topographic indices**, can be derived [Randić and Wilkins, 1979b; Randić, 1988a; Randić, Jerman-Blazic *et al.*, 1990; Diudea, Horvath *et al.*, 1995b; Randić and Razingar, 1995a; Balaban, 1997b].

Analogously to the $\rightarrow$ *vertex distance degree*, the *i*th row sum of the geometry matrix is called **geometric distance degree** ${}^G\sigma_i$ (or **Euclidean degree**) [Balasubramanian, 1995b]:

$${}^G\sigma_i = \sum_{j=1}^A r_{ij}$$

This is a local vertex invariant used, for example, in the definition of the $\rightarrow$ *3D-connectivity indices* $\chi\chi$ and the $\rightarrow$ *Euclidean connectivity index*. In general, the row sum of this matrix represents a measure of the centrality of an atom; atoms that are close to the $\rightarrow$ *center of the molecule* have smaller atomic sums, whereas those far from the center have large atomic sums. The smallest and the largest row sums give the extreme values of the first eigenvalue of the geometry matrix; therefore, when all the atoms are equivalent, that is, the distance degrees are all the same, the geometric distance degree yields exactly the first eigenvalue. The average sum of all geometric distance degrees is a molecular invariant called **average geometric distance degree**, that is,

$${}^G\bar{\sigma} = \frac{1}{A} \cdot \sum_{i=1}^A {}^G\sigma_i = \frac{1}{A} \cdot \sum_{i=1}^A \sum_{j=1}^A r_{ij}$$

whereas the half sum of all geometric distance degrees is another molecular descriptor called **3D-Wiener index** by analogy with the $\rightarrow$ *Wiener index* calculated from the topological distance matrix. The 3D Wiener index is calculated as

$${}^{3D}W_H \equiv Wi(\mathbf{G}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A r_{ij}$$

where r_{ij} is the interatomic distance between the *i*th and *j*th atom [Mekenyan, Peitchev *et al.*, 1986a, 1986b; Bogdanov, Nikolić *et al.*, 1989, 1990; Randić, Jerman-Blazic *et al.*, 1990]. This index is obviously more discriminant than the 2D Wiener index as it accounts for spatial molecular geometry; it shows different values for different molecular conformations, the largest values corresponding to the most extended conformations, the smallest to the most compact

conformations. Therefore, it is considered among $\rightarrow$ *shape descriptors* since it decreases with increasing sphericity of a structure [Nikolić, Trinajstić *et al.*, 1991]. The 3D Wiener index can be calculated both considering ${}^3D W_H$ and not considering ${}^3D W$ hydrogen atoms [Basak, Gute *et al.*, 1999a]. Moreover, a strictly related molecular descriptor is the $\rightarrow$ *bond length-weighted Wiener index* calculated by using as the distance between two atoms the sum of the bond lengths along the shortest path.

A 3D local vertex invariant based on the geometric distance was proposed as [Toropov, Toropova *et al.*, 1998; Krenkel, Castro *et al.*, 2002]

$$3DW_i = \sum_{j=1}^A (1 - a_{ij}) \cdot \exp(r_{ij}^{-2}) \quad j \neq i$$

where the summation accounts only for contributions from the pairs of nonadjacent atoms, a_{ij} being the elements of the adjacency matrix equal to one only for pairs of adjacent atoms, and zero otherwise. The exponential form of the distance was chosen from a series of terms approximating the attracting interatomic potentials.

From these local invariants, $\rightarrow$ *Zagreb indices*, $\rightarrow$ *connectivity-like indices*, and $\rightarrow$ *Wiener-type indices* were derived as

$$\begin{aligned} 3DM_1 &= \sum_{i=1}^A 3DW_i & 3DM_2 &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (3DW_i \cdot 3DW_j) \\ 3D^0\chi &= \sum_{i=1}^A (3DW_i)^{-1/2} & 3D^1\chi &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (3DW_i \cdot 3DW_j)^{-1/2} \\ 3DW_i &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A \exp(r_{ij}) \end{aligned}$$

where A is the number of molecule atoms.

Molecular descriptors based on this kind of local vertex invariant are called **MIS indices**, being defined in the framework of the **Method of Ideal Symmetry** (MIS), based on a partial optimization procedure of the molecular geometry, where bond lengths and bond angles are kept fixed and only free rotations around C–C bonds are varied [Toropov, Toropova *et al.*, 1994].

The maximum value entry in the i th row of the geometry matrix is a local descriptor called **geometric eccentricity** ${}^G\eta_i$ representing the longest geometric distance from the i th atom to any other atom in the molecule:

$${}^G\eta_i = \max_j(r_{ij})$$

From the eccentricity definition, **geometric radius** ${}^GR$ and **geometric diameter** ${}^GD$ can immediately characterize a molecule. The radius of a molecule is defined as the minimum geometric eccentricity and the diameter is defined as the maximum geometric eccentricity in the molecule, according to the following:

$${}^GR = \min_i({}^G\eta_i) \quad \text{and} \quad {}^GD = \max_i({}^G\eta_i)$$

These parameters are $\rightarrow$ *size descriptors* also depending on the molecular shape ($\rightarrow$ *Petitjean shape indices*), such as their topological counterpart, that is, $\rightarrow$ *topological radius* and $\rightarrow$ *topological diameter*.

Derived from the geometry matrix, the **neighborhood geometry matrix** (or **neighborhood Euclidean matrix**), denoted as ${}^N\mathbf{G}$, was also proposed as [Bajzer, Randić *et al.*, 2003]

$$[{}^N\mathbf{G}]_{ij} = \begin{cases} r_{ij} & \text{if } r_{ij} \leq R_t \\ 0 & \text{if } r_{ij} > R_t \end{cases}$$

where R_t is a user-defined distance threshold. This matrix was used to calculate descriptors of $\rightarrow$ *proteomics maps* by the additional constraint that the matrix element i - j is set at zero also for nonconnected protein spots.

The **reciprocal geometry matrix**, denoted as $\mathbf{G}^{-1}$, is obtained from the geometry matrix as the following:

$$[\mathbf{G}^{-1}]_{ij} = \begin{cases} r_{ij}^{-1} & i \neq j \\ 0 & i = j \end{cases}$$

In the same way, the **reciprocal topographic matrix**, denoted as $\mathbf{T}^{-1}$, is defined in terms of the reciprocal of topographic distances instead of the reciprocal of geometric distances.

Example M3

Geometry matrix $\mathbf{G}$, geometric distance degrees ${}^G\sigma$, eccentricities ${}^G\eta$, geometric radius ${}^G R$ and diameter ${}^G D$ for 2-methylpentane. ${}^G\bar{\sigma}$ and ${}^{3D}W$ are the average geometric distance degree and the 3D-Wiener index, respectively.

Atom	1	2	3	4	5	6	${}^G\sigma$	${}^G\eta$
1	0	1.519	2.504	3.856	5.014	2.498	15.391	5.014
2	1.519	0	1.530	2.521	3.864	1.521	10.955	3.864
3	2.504	1.530	0	1.521	2.509	2.507	10.571	2.509
4	3.856	2.521	1.521	0	1.511	3.038	12.447	3.856
5	5.014	3.864	1.511	1.511	0	4.348	17.246	5.014
6	2.498	1.521	3.038	3.038	4.348	0	13.912	4.348

$${}^G R = \min_i({}^G\eta_i) = 2.509 \quad {}^G\bar{\sigma} = \frac{1}{6} \cdot (15.391 + 10.955 + 10.571 + 12.447 + 17.246 + 13.912) = 13.420$$

$${}^G D = \max_i({}^G\eta_i) = 5.014 \quad {}^{3D}W = \frac{1}{6} \cdot (15.391 + 10.955 + 10.571 + 12.447 + 17.246 + 13.912) = 40.261$$

From the geometry matrix, the usual $\rightarrow$ *graph invariants* can be calculated such as $\rightarrow$ *characteristic polynomial*, $\rightarrow$ *spectral indices*, $\rightarrow$ *ID numbers*, $\rightarrow$ *3D-Balaban index*, $\rightarrow$ *3D-Schultz index*, and so forth [Randić, 1988b; Nikolić, Trinajstić *et al.*, 1991]. It is noteworthy that all these indices despite their topological counterparts are sensitive to molecular geometry. Moreover, geometry matrix is used for the calculation of $\rightarrow$ *size descriptors* and $\rightarrow$ *3D-MoRSE descriptors*.

Important derived matrices are the powers of the geometry matrix, used to define $\rightarrow$ *molecular profiles* descriptors. Moreover, **distance/distance matrices**, denoted as $\mathbf{D}/\mathbf{D}$, were defined as $\rightarrow$ *quotient matrices* in terms of geometric r_{ij} or topographic distances t_{ij} and $\rightarrow$

topological distances d_{ij} to unify 2D and 3D information about the structure of molecules [Randić, 1994, 1999]:

$$\begin{aligned} [\mathbf{G}/\mathbf{D}]_{ij} &= \begin{cases} \frac{r_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} & [\mathbf{T}/\mathbf{D}]_{ij} &= \begin{cases} \frac{t_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \\ [\mathbf{D}/\mathbf{G}]_{ij} &= \begin{cases} \frac{d_{ij}}{r_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} & [\mathbf{D}/\mathbf{T}]_{ij} &= \begin{cases} \frac{d_{ij}}{t_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \end{aligned}$$

The **geometric distance/topological distance quotient matrix**, denoted as $\mathbf{G}/\mathbf{D}$, is a square symmetric matrix $A \times A$, A being the number of molecule atoms, whose entries are the quotient of the corresponding elements of the molecular geometry matrix $\mathbf{G}$ and the graph $\rightarrow$ distance matrix $\mathbf{D}$. An alternative to the geometric distance/topological distance quotient matrix is the **topographic distance/topological distance quotient matrix** ($\mathbf{T}/\mathbf{D}$), derived by using the $\rightarrow$ topographic matrix $\mathbf{T}$ instead of the geometry matrix. Note that in the original papers, both these matrices were indifferently referred to as distance/distance matrix and denoted by $\mathbf{DD}$.

The **topological distance/geometric distance quotient matrix**, denoted by $\mathbf{D}/\mathbf{G}$, is the reciprocal matrix of the $\mathbf{G}/\mathbf{D}$ matrix, and the **topological distance/topographic distance quotient matrix**, denoted by $\mathbf{D}/\mathbf{T}$, the reciprocal matrix of the $\mathbf{T}/\mathbf{D}$ matrix.

The row sums of these matrices contain information on the molecular folding; in effect, in highly folded structures, they tend to be relatively small as the interatomic distances are small while the topological distances increase as the size of the structure increases. Therefore, the average row sum is a molecular invariant called **average distance/distance degree**, that is,

$$ADDD = \frac{1}{A} \cdot \sum_{i=1}^A \sum_{j=1}^A \frac{r_{ij}}{d_{ij}} \quad j \neq i$$

while the half sum of all distance/distance matrix entries is another molecular descriptor called **D/D index**, that is,

$$D/D \equiv Wi(\mathbf{G}/\mathbf{D}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \frac{r_{ij}}{d_{ij}} \quad j \neq i$$

where Wi is the $\rightarrow$ Wiener operator.

From the largest eigenvalue of the distance/distance matrix a $\rightarrow$ folding degree index was also defined.

Other matrices that combine topological and geometrical information are **distance–distance combined matrices**, which are defined in terms of geometric (r_{ij}) or topographic (t_{ij}) distances as [Janežič, 2007]:

$$\begin{aligned} [\mathbf{G} \wedge \mathbf{D}]_{ij} &= \begin{cases} r_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij} & \text{if } i > j \end{cases} & [\mathbf{T} \wedge \mathbf{D}]_{ij} &= \begin{cases} t_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij} & \text{if } i > j \end{cases} \\ [\mathbf{D} \wedge \mathbf{G}]_{ij} &= \begin{cases} d_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ r_{ij} & \text{if } i > j \end{cases} & [\mathbf{D} \wedge \mathbf{T}]_{ij} &= \begin{cases} d_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ t_{ij} & \text{if } i > j \end{cases} \end{aligned}$$

where $\mathbf{D}$ is the topological distance matrix and d_{ij} the topological distance between vertices v_i and v_j ; $\mathbf{G} \wedge \mathbf{D}$ is the **geometric distance–topological distance combined matrix**, $\mathbf{T} \wedge \mathbf{D}$ is the **topographic distance–topological distance combined matrix**, $\mathbf{D} \wedge \mathbf{G}$ is the **topological distance–geometric distance combined matrix**, and $\mathbf{D} \wedge \mathbf{T}$ is the **topological distance–topographic distance combined matrix**. Note that $\mathbf{D} \wedge \mathbf{G}$ and $\mathbf{D} \wedge \mathbf{T}$ are the transpose matrices of $\mathbf{G} \wedge \mathbf{D}$ and $\mathbf{T} \wedge \mathbf{D}$.

📖 [Turro, 1986; Mihalić and Trinajstić, 1991; Mihalić, Nikolić *et al.*, 1992; Kunz, 1993, 1994; Warthen, Schmidt *et al.*, 1993; Balasubramanian, 1995b; Estrada and Ramirez, 1996; Zhu and Klein, 1996; Laidboeur, Cabrol-Bass *et al.*, 1997; Randić and Razinger, 1997; Ivanciuc, Ivanciuc *et al.*, 1998b; Ivanciuc and Ivanciuc, 1999; Tao and Lu, 1999; Blatova, Blatov *et al.*, 2001, 2002; Imre, Veress *et al.*, 2003; Todeschini and Consonni, 2003; Wisniewski, 2003; Wang, Wang *et al.*, 2006]

■ **molecular graph** ($\equiv$ *structural graph, constitutional graph*)

It is a nondirected connected $\rightarrow$ *graph* G , which represents a chemical compound, that is, a graph where vertices and edges are chemically interpreted as atoms and covalent bonds [Harary, 1969a, 1969b; Balaban and Harary, 1976; Rouvray and Balaban, 1979; Rouvray, 1990a; Bonchev and Rouvray, 1991; Trinajstić, 1992]. A molecular graph obtained excluding all the hydrogen atoms is called **H-depleted molecular graph** (or **hydrogen-depleted molecular graph** or **Labeled Hydrogen-Suppressed molecular Graph**, LHSG), whereas a molecular graph where also hydrogens are graph vertices is called **H-filled molecular graph** (or **hydrogen-included molecular graph** or **hydrogen-filled molecular graph** or **Labeled Hydrogen-Filled molecular Graph**, LHFG).

Such a graph depicts the connectivity of atoms in a molecule irrespective of the metric parameters such as equilibrium $\rightarrow$ *interatomic distances* between nuclei, $\rightarrow$ *bond angles*, and $\rightarrow$ *torsion angles*, representing the 3D $\rightarrow$ *molecular geometry*. Thus, a molecular graph is a $\rightarrow$ *topological representation* of the molecule, and it is from this that several $\rightarrow$ *molecular descriptors* are derived (Figure M6).

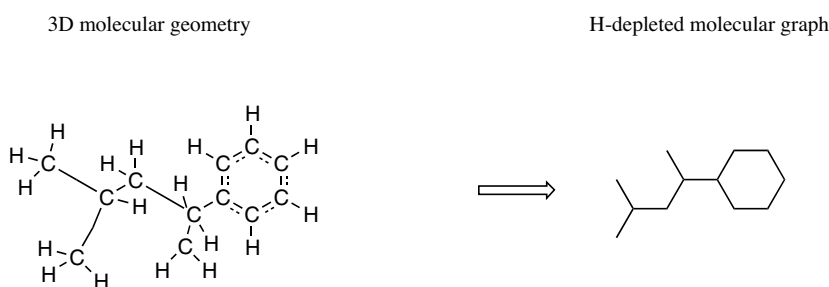


Figure M6 The transition from 3D geometry to 2D topology.

Some vertices of the H-depleted molecular graph can be more precisely defined as the **hydride group**, which is a heavy atom plus its bonded hydrogens. For example, hydride groups are $-\text{CH}_3$, $-\text{CH}_2-$, $=\text{NH}$, $-\text{NH}_2$, and $-\text{OH}$.

A **molecular subgraph** is a subset of atoms and related bonds, which is in itself a valid graph usually representing molecular fragments and functional groups.

There are four commonly used subgraph types: **path subgraph**, **cluster subgraph**, **path-cluster subgraph**, and **chain subgraph** (or *Ring*), emphasizing different aspects of atom connectivity within the molecule. They are defined according to the following rules: (1) if the subgraph contains a cycle, it is of type Chain (CH); otherwise, (2) if all $\rightarrow$ *vertex degrees* in the subgraph (not in the whole graph) are either greater than 2 or equal to 1, the subgraph is of type Cluster (C); otherwise, (3) if all vertex degrees in the subgraph are either equal to 2 or 1, the subgraph is of type Path (P); otherwise, (4) the subgraph is of type Path-Cluster (PC) (Figure M7).

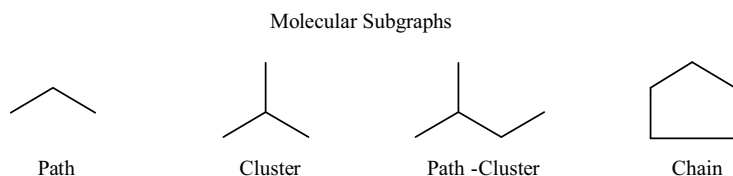


Figure M7 Elementary molecular subgraphs.

The **order of a subgraph** is the number of edges within it. Note that subgraphs of order 0 are considered of type Path, subgraphs of order 1 and 2 are only of type Path, and subgraphs of order 3 can be of type Path, Cluster, or Chain only.

Referring to the subgraph order, r th order indices can be defined as $\rightarrow$ *count descriptors*, that is, the number of r th order subgraphs in the graph G . The zero order index is simply the $\rightarrow$ *atom number* A , that is, the number of graph vertices; first order index is the $\rightarrow$ *bond number* B , that is, the number of graph edges; second order index is the $\rightarrow$ *connection number*, that is, the second order $\rightarrow$ *molecular path count*; third order indices are the number of paths of length 3, the number of three-edge clusters, and the number of three-edges cycles.

The total number K of connected subgraphs of a molecular graph G is a very simple measure of $\rightarrow$ *molecular complexity*, obviously referring only to structural complexity of the molecule; it is called **total subgraph count**.

The simplest form to represent the chemical information contained in a molecular graph is by $\rightarrow$ *graph-theoretical matrices*. Examples are $\rightarrow$ *adjacency matrix* A , $\rightarrow$ *edge adjacency matrix* E , vertex $\rightarrow$ *distance matrix* D , $\rightarrow$ *edge distance matrix* E^D , $\rightarrow$ *incidence matrix* I , $\rightarrow$ *Wiener matrix* W , $\rightarrow$ *Hosoya Z-matrix* Z , $\rightarrow$ *Cluj matrices* CJ , $\rightarrow$ *detour matrix* Δ , $\rightarrow$ *Szeged matrix* SZ , $\rightarrow$ *geometric distance/topological distance quotient matrix* G/D , and $\rightarrow$ *detour-distance combined matrix* $\Delta \wedge D$.

A **reduced graph** is a molecule representation aimed at the storage and retrieval of generic chemical structures. The internal representation of the molecule is hierarchically tree structured as a topological graph, the chemical nature of the various parts of the generic structure being represented in the vertices of the graph and the information about their connections and relationships in its edges. Since information on the chemical nature of each part is predominantly based on conventional $\rightarrow$ *connectivity tables*, the whole is a sort of superconnection table or connection table of connection tables, and is referred to as an **Extended Connection Table Representation (ECTR)**. Within the ECTR, each vertex is called a partial structure and each edge a gate [Barnard, Lynch *et al.*, 1982; Gillet, Downs *et al.*, 1991; Gillet, Willett *et al.*, 2003].

Because in a reduced graph, groups of atoms within the structure are collapsed together to form single vertices and smaller graphs are obtained, more quick searching can be performed on large molecule data sets, using any of the conventional methods [Barnard, 1993].

The reduced graph may contain vertices representing the cyclic and acyclic portions of the molecule or contiguous groups of carbon or heteroatoms.

Common structural features between molecules are searched for by means of graph-theoretical approaches, such as the determination of the maximal $\rightarrow$ *cliques* of the docking graph. A clique in the docking graph corresponds to a grouping of functional groups in the original reduced graphs, where all the intragrouping distances are the same in both original graphs.

An example of a reduced graph is a molecular graph composed of weighted edges and with vertex number equal to the number of functional groups perceived by using predefined rules [Takahashi, Sukekawa *et al.*, 1992].

Once all the functional atomic groups are perceived, the interrelations between them are checked. The relationships evaluated are described in terms of matrix expression, and they are divided into the following two cases:

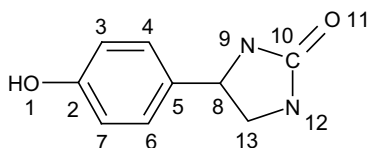
- (1) The case in which two functional groups are partially overlapping (overlap matrix).
- (2) The case in which one of the functional atomic groups is completely included by another one (inclusion matrix).

For the former case, the relationship is described in the overlap matrix, and for the latter, the relationship is described in the inclusion matrix to avoid the duplication of vertices in the reduced graph representation. They are used to determine the $\rightarrow$ *topological distance* between functional groups.

Therefore, this reduced graph representation is a graph constituted by a number of vertices equal to the number of perceived functional groups and weighted edges, whose weights correspond to the shortest topological distance between the different functional groups. If the structure has ring(s), there are several possible paths that can be drawn simultaneously between functional groups. In such a case, some of the edges are weighted with multiple values.

Example M4

Derivation of a reduced graph.



- A : Hydroxy {1}
 B : Benzene {2, 3, 4, 5, 6, 7}
 C : Amido {9, 10, 11}
 D : Amido {12, 10, 11}
 E : Urea {9, 10, 11, 12}

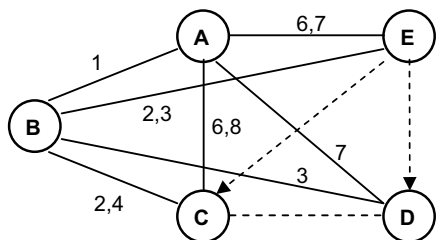
Overlap matrix

	A	B	C	D	E
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	1	1
D	0	0	1	1	1
E	0	0	1	1	1

Inclusion matrix

	A	B	C	D	E
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	1	1	1

The reduced graph obtained by defining the functional groups outlined above can be drawn as the following:



Other approaches to generate and/or manage reduced graphs were proposed, such as the SwiFT index method, created to identify redundant subtrees [Fischer and Rarey, 2007], and homeomorphically reduced graph formed by deleting all atoms of connectivity 2 [Cringean and Lynch, 1989].

The **feature tree** is a representation of a molecule similar to a reduced graph. A feature tree represents hydrophobic fragments and functional groups of the molecule and the way these groups are linked together [Rarey and Dixon, 1998; Rarey and Stahl, 2001]. The vertices of the feature tree are molecular fragments and edges connect vertices that represent fragments that are connected in the simple molecular graph. Moreover, each vertex in the tree is associated with a set of features representing chemical properties of the molecular fragment corresponding to the vertex.

Feature trees are used in $\rightarrow$ *similarity/diversity* analysis of compounds: the comparison of the feature trees of two compounds is based on matching subtrees of the two feature trees onto each other.

📖 [Mason, 1943; Gutman and Trinajstić, 1972; Gutman *et al.*, 1975; Gutman, Ruscic *et al.*, 1975; Randić, 1975b; Balaban, 1976d, 1978a, 1985a, 1985b, 1987, 1991, 1993d, 1995a; Polanski and Rouvray, 1976a, 1976b; Balaban and Rouvray, 1980; Mekenyan, Bonchev *et al.*, 1980, 1981; Trinajstić, Jericevic *et al.*, 1983; Hosoya, 1986; Trinajstić, Klein *et al.*, 1986; Hansen and Jurs, 1988a; Dias, 1993; Klein, 1997; Bytautas and Klein, 1998, 1999; John, Mallion *et al.*, 1998; Bonchev and Rouvray, 2000; Kruja, Marks *et al.*, 2002; Ivanciuc, 2003f; Kerber, Laue *et al.*, 2004; García-Domenech, Gálvez *et al.*, 2008]

➤ **molecular graphics and modeling descriptors** $\rightarrow$ molecular graphics descriptors

■ molecular graphics descriptors

These are descriptors derived from high-quality 2D projections of molecules or molecular aggregates obtained by current molecular graphic techniques, which can be an extensive source of quantitative information on molecular properties [Kiralj and Ferreira, 2003a].

In general, quantities directly “measured” from pictures using some digital or analogue technique can be 1D (such as molecular dimensions), 2D (such as surface areas), or 3D (such as molecular volumes).

Combination of these descriptors with some structural information from other sources (such as data from experimental structure determination or other descriptors useful in molecular modeling) yields composite functions, which are called **molecular graphics–structural descriptors** and **molecular graphics and modeling descriptors**, respectively. Both classes of descriptors can be global (describing the entire molecule) or local (being related to some molecular fragment).

- **molecular graphics–structural descriptors** → molecular graphics descriptors
- **molecular holographic distance vector** → MEDV-13 descriptor
- **molecular holograms** → substructure descriptors (⊙ fingerprints)
- **molecular ID numbers** → ID numbers
- **molecular influence matrix** → GETAWAY descriptors

■ **molecular interaction fields** ($\equiv$ *interaction fields*)

A molecular interaction field is a scalar field of → *interaction energy values* between a molecule, whose → *molecular geometry* is known, and a → *probe* [Wade, 1993; Andrews, 1993; Leach, 1996]. For QSAR studies, molecular interaction fields are calculated using one or more probes for a number of compounds previously aligned by specific → *alignment rules* and embedded in the same fixed → *grid*, that is, a regular 3D array of N_G points, each point **p** being characterized by grid coordinates (x, y, z). The interaction energy values are calculated by moving the probe in each grid point.

Depending on the selected probe and the defined potential energy function, several molecular interaction fields can be calculated. The most common are *steric fields* and *electrostatic fields*, sometimes referred to as **CoMFA fields** because originally implemented in → *CoMFA*. Several interaction fields are actually calculated in → *GRID method*.

Derived from a topological approach → *E-state fields* and → *HE-fields* were defined.

The **enthalpic fields** are all of the molecular interaction fields accounting for enthalpic contributions to the free energy of ligand–receptor binding, such as *steric fields* and *electrostatic fields*. On the other hand, the **entropic fields** are all of the molecular interaction fields accounting for entropic contributions to the free energy of ligand–receptor binding. The entropy of binding is related to hydrophobic interactions between nonpolar ligand and receptor lipophilic chemical groups after the release of water molecules formerly structured around the receptor groups, and to the loss of conformational freedom due to ligand immobilization at the binding site. The entropy of binding is mainly modeled by *hydrophobic fields* or *hydrogen bonding fields*; however, sometimes also the degrees of torsional freedom in the molecule were considered to account for the entropy change resulting from the reduced conformational freedom of the ligand in the receptor complex [Greco, Novellino *et al.*, 1997].

Some molecular interaction fields are listed below.

• **steric interaction fields** ($\equiv$ *van der Waals interaction fields*)

A steric interaction field is obtained calculating the van der Waals interaction energy E_{vdw} between probe and target in each grid point [Kim, 1992b]. Different potential energy functions were proposed to model van der Waals interactions between atoms. The most common are *Lennard–Jones potential*, *Buckingham potential*, and *Hill potential* [Leach, 1996].

The general formula of **Lennard–Jones 6–12 potential function** [Lennard–Jones, 1924, 1929] is

$$E_{vdw} = \sum_s \sum_t \left(\frac{A_{st}}{r_{st}^{12}} - \frac{C_{st}}{r_{st}^6} \right)$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule; r_{st} is the interatomic distance between the s th atom of the probe and the t th atom of the target; A and C are two functions defined as

$$A_{st} = \sqrt{\epsilon_s \cdot \epsilon_t} \cdot (R_s + R_t)^{12} \quad \text{and} \quad C_{st} = 2 \cdot \sqrt{\epsilon_s \cdot \epsilon_t} \cdot (R_s + R_t)^6$$

where ϵ is the well depth and R one half the separation at which the energy passes through a minimum (i.e., the $\rightarrow$ *van der Waals radius*). The Lennard–Jones potential is characterized by an attractive component that varies as r^{-6} and a repulsive component that varies as r^{-12} . The energy function modeling the steric repulsion between pairs of atoms becomes large and positive at interatomic distances r less than the sum of the van der Waals radii of the probe atom and the target atom.

The **Buckingham potential function** [Buckingham, 1938] is defined in an exponential form as

$$E_{vdw} = \sum_s \sum_t \left(A \cdot \exp^{-B \cdot r_{st}} - \frac{C}{r_{st}^6} \right)$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule; r_{st} is the interatomic distance between the s th atom of the probe and the t th atom of the target; A , B , and C are functions of the well depth ϵ , the van der Waals radius R , and an adjustable parameter α . The exponential energy function is commonly used for small molecules, the Lennard–Jones 12–6 function for macromolecules.

The **Hill potential function** is an exponential function defined as

$$E_{vdw} = \sum_s \sum_t \left[-2.25 \cdot \sqrt{\epsilon_s \cdot \epsilon_t} \cdot \left(\frac{R_s + R_t}{r_{st}} \right)^6 + 8.28 \cdot 10^5 \cdot \sqrt{\epsilon_s \cdot \epsilon_t} \cdot \exp \left(-\frac{r_{st}}{0.073 \cdot (R_s + R_t)} \right) \right]$$

where ϵ is the well depth and R the van der Waals radius. The coefficients were determined by fitting them to data for the rare gases [Hill, 1948].

• electrostatic interaction fields

These are molecular interaction fields obtained by calculating electrostatic interaction energy E_{el} between probe and target in each grid point. Besides the $\rightarrow$ *molecular electrostatic potential* (MEP), the most common energy function for electrostatic interactions is the **Coulomb potential energy function** defined as

$$E_{el} = \sum_s \sum_t \frac{q_s \cdot q_t}{4\pi \cdot \epsilon_0 \cdot \epsilon_m \cdot r_{st}}$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule; r_{st} is the interatomic distance between the s th atom of the probe and the t th atom of the target; q is the $\rightarrow$ *partial atomic charge*; and ϵ_0 is the permittivity of the free space and ϵ_m is the relative dielectric constant of the surrounding medium.

The **GRID electrostatic energy function** was proposed to account for the dielectric discontinuity between a solute and the solvent as [Goodford, 1985]

$$E_{el} = \sum_s \sum_t \frac{q_s \cdot q_t}{K \cdot \zeta} \cdot \left(\frac{1}{r_{st}} + \frac{(\zeta - \epsilon)/(\zeta + \epsilon)}{\sqrt{(r_{st}^2 + 4 \cdot s_s \cdot s_t)}} \right)$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule; r_{st} is the interatomic distance between the s th atom of the probe and the t th atom of the target; q is the partial atomic charge; K is a constant; ζ and ϵ are the relative dielectric constants of the protein and the target solution phases, respectively; s_s and s_t are the nominal depths at which the probe atom and the target atom are respectively buried in the target phase. These depths are calculated by counting the number of neighboring target atoms within a distance of 4 Å and translating this into an equivalent depth using a calibrated scale.

• molecular orbital fields

These are fields restricted to the regions occupied by selected molecular orbitals; of particular interest are fields related to the $\rightarrow$ *highest occupied molecular orbital* (HOMO) and to the $\rightarrow$ *lowest unoccupied molecular orbital* (LUMO) [Navajas, Poso *et al.*, 1996; Oprea and Waller, 1997; Durst, 1998].

Molecular orbital fields are descriptors particularly useful when an ionic or charge transfer reaction is part of the ligand–receptor interaction; in this case, electrostatic fields are not able to fully represent the electronic characteristics of molecules.

To calculate a molecular orbital field, semiempirical single-point calculations are performed on the molecule-optimized geometry and the electron density at each grid point in the region of the selected orbital is determined.

• hydrophobic fields

These are molecular descriptors based on hydrophobic interaction energy between nonpolar surfaces of ligand and receptor. The energy of hydrophobic interactions derives from the disruption of the water structure around nonpolar surfaces resulting in a gain of entropy [Abraham and Kellogg, 1993].

Kellogg and Abraham interaction field, also called **Hydropathic Interactions** (HINT), is a hydrophobic field calculated by $\rightarrow$ *Leo–Hansch hydrophobic fragmental constants* scaled by surface area and a distance-dependent function [Kellogg, Semus *et al.*, 1991; Kellogg and Abraham, 1992; Abraham and Kellogg, 1993]. **Hydropathy** is a term used in structural molecular biology to represent the hydrophobicity of amino acid side chains.

The hydropathic field in each grid point is calculated as

$$E_{\text{hy}} = \sum_s \sum_t (SA_s \cdot h_s \cdot SA_t \cdot h_t \cdot R_{st} + R'_{st})$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule; SA is the atomic $\rightarrow$ *solvent-accessible surface area*, h the hydropathic atom constant, and R_{st} and R'_{st} are functions of the interatomic distance r_{st} between the s th atom of the probe and the t th atom. The function R_{st} scales the product between solvent-accessible surface area and hydropathic constant with a distance usually defined as

$$R_{st} = I_{st} \cdot \exp^{-r_{st}}$$

where I_{st} is a sign-flip function recognizing acid–base interactions. The function R'_{st} is a Lennard–Jones-type potential accounting for close contacts of atoms by van der Waals radius term:

$$R'_{st} = A \cdot \epsilon_{st} \cdot \left[\left(\frac{R_s + R_t}{r_{st}} \right)^{12} - 2 \cdot \left(\frac{R_s + R_t}{r_{st}} \right)^6 \right]$$

where A is a scaling factor, ϵ_{si} is the depth of the Lennard–Jones potential well, and R is the van der Waals radius of the considered atoms. The probe is usually taken as a single atom and its parameters are set to unity.

Hydropathic atom constants h are derived from Leo–Hansch hydrophobic fragmental constants in such a way that

- the sum of hydropathic atom constants in a group is consistent with the group fragmental constant;
- frontier atoms in a group are more important than shielded atoms;
- bond, chain, branch, and proximity factors are applied in an additive scheme, the former three to all eligible atoms, the last to the central atoms of polar groups.

Positive hydropathic constants indicate hydrophobic atoms, whereas negative constants indicate hydrophilic atoms.

The **Molecular Lipophilicity Potential** (MLP) describes the combined lipophilic effect of all fragments in a molecule on its environment and can be calculated at any point in space around the molecule [Audry, Dubost *et al.*, 1986, 1992; Fauchère, Quarendon *et al.*, 1988; Furet, Sele *et al.*, 1988; Audry, Dallet *et al.*, 1989]. It is defined by considering a molecule surrounded by nonpolar or low polarity organic solvent molecules, and assuming that the solvent molecule distribution around the considered molecule depends on the fragmental or atomic contributions to $\log P$ and the distances at which the solvent molecules are from the target molecule. Therefore, the molecular lipophilicity potential at each k th grid point is calculated as

$$\text{MLP}_k = \sum_{i=1}^A \frac{a_i}{1 + r_{ki}}$$

where the sum runs over all atoms (or fragments) of the target molecule; a_i are the $\rightarrow$ Ghose–Crippen hydrophobic atomic constants for the i th atom (or fragments) in the target molecule, and r_{ki} is the distance between the considered atom (or fragments) and the k th grid point. Only non-hydrogen atoms A of the molecule are usually considered.

In contrast to other potentials, the lipophilicity potential is not obtained by calculating the interactions between a probe and the molecule.

Different MLP functions can be obtained according to the selection of the fragmental constant values and the distance function [Croizet, Langlois *et al.*, 1990; Gaillard, Carrupt *et al.*, 1994a, 1994b; Testa, Carrupt *et al.*, 1996; Carrupt, Testa *et al.*, 1997]. The MLP has been later adapted to a new atomic hydrophobic parameter called **Topological Lipophilicity Potential** (TLP) defined for each j th atom of the molecule as [Langlois, Audry *et al.*, 1993; Dubost, 1993]

$$\text{TLP}_j = \sum_{i=1}^A \frac{a_i}{1 + d_{ij}}$$

where d_{ij} is the $\rightarrow$ topological distance between atoms i and j of the molecule.

📖 [Gussio, Pattabiraman *et al.*, 1996; Masuda, Nakamura *et al.*, 1996; Testa, Raynaud *et al.*, 1999]

• hydrogen bonding fields

These are descriptors accounting for hydrogen-bonding interactions between ligand and receptor. Hydrogen bonding fields are obtained by calculating the energy E_{hb} due to the

formation of hydrogen-bonds between probe and target in each grid point [Leach, 1996; Oprea and Waller, 1997].

The hydrogen bonding potential energy is calculated as [Wade, 1993]

$$E_{hb} = \sum_s \sum_t E_r(r_{st}) \cdot E_s \cdot E_t$$

where the first sum runs over probe atoms and the second over atoms of the target molecule; E_r is an energy component dependent on the interatomic distance r_{st} between probe and target atoms involved in the hydrogen-bond; E_s and E_t are energy components dependent on the angle made by the hydrogen bond at the probe and target atoms, respectively. E_s and E_t values are between 0 and 1. The component E_r is usually defined by a Lennard-Jones function as

$$E_r = \frac{A}{r_{st}^m} - \frac{C}{r_{st}^n}$$

where A and C are constants dependent on the chemical type of the hydrogen-bonding atoms; m and n are parameters taking different values; for example, $m = 12$ and $n = 10$ are commonly used values, $m = 8$ and $n = 6$ were used in the GRID hydrogen bonding energy function [Boobbyer, Goodford *et al.*, 1989].

A more sophisticated hydrogen bonding potential energy based on the geometry of the hydrogen-bonding systems was proposed by Kim [Kim, 1993a, 1993f; Kim, Greco *et al.*, 1993] and implemented in the GRID program:

$$E_{hb} = \left(\frac{C}{r_{st}^6} - \frac{D}{r_{st}^4} \right) \cdot \cos(m \cdot \theta)$$

where the energy is evaluated in each grid point; r_{st} is the interatomic distance between probe and target atoms involved in the hydrogen-bond; C and D are parameters taken from tables; m is usually equal to one; and θ is the angle made by donor, hydrogen, and acceptor atoms. The probe used is a neutral H_2O molecule with an effective radius of 1.7 Å, free to rotate around the grid point.

• total interaction energy fields

These are potential energy descriptors accounting for the total noncovalent interaction potential energy, which determines the binding affinity of a molecule to the considered receptor. They are generally calculated as the pairwise sum of the interaction energies between each probe atom and each target atom as [Wade, 1993]

$$E = \sum_s \sum_t (E_{vdw} + E_{el} + E_{hb})_{st}$$

where the first sum runs over probe atoms and the second over atoms of the target molecule; E_{vdw} is the van der Waals interaction energy, E_{el} the electrostatic energy, and E_{hb} the hydrogen-bonding energy. Other noncovalent energy contributions can be included.

• desolvation energy fields

These are potential energy descriptors proposed as an indicator of hydrophobicity [Oprea and Waller, 1997]. Originally, they were calculated using the finite difference approximation method; the linearized Poisson-Boltzmann equation was solved numerically to compute the electrostatic

contribution to solvation at each grid point. Desolvation energy field values were calculated as the difference between solvated (grid dielectric = 80) and *in vacuum* (grid dielectric = 1).

📖 [Richard, 1991; Balogh and Naray-Szabo, 1993; Kim, 1993b; Naray-Szabo and Balogh, 1993; Nusser, Balogh *et al.*, 1993; van de Waterbeemd, Camenisch *et al.*, 1996; Liljefors, 1998] [Cruciani, Pastor *et al.*, 2001a; Cruciani, Benedetti *et al.*, 2004; Cianchetta, Li *et al.*, 2006; Goodford, 2006; Wade, 2006]

- **molecular lipophilicity potential** → molecular interaction fields (⊙ hydrophobic fields)
- **molecular lipophilicity potential model** → lipophilicity descriptors
- **molecular matrix** → molecular geometry
- **Molecular Modeling** ≡ *Computer-Aided Molecular Modeling* → drug design
- **molecular moment of energy** → self-returning walk counts
- **molecular negentropy** → information content
- **molecular orbital contour surface** → molecular surface
- **molecular orbital energies** → quantum-chemical descriptors
- **molecular orbital fields** → molecular interaction fields
- **molecular path code** → path counts
- **molecular path count** → path counts
- **molecular path number** ≡ *molecular path count* → path counts
- **molecular path/walk indices** → shape descriptors (⊙ path/walk shape indices)
- **molecular polarizability** → electric polarization descriptors
- **Molecular Polarizability Effect Index** → electric polarization descriptors (⊙ Polarizability Effect Index)

■ molecular profiles

These are molecular descriptors denoted by ${}^kD$ and derived from the → *distance distribution moments* of the → *geometry matrix G*, defined as the average row sum of its entries raised at the k th power, normalized by the factor $k!$:

$${}^kD = \frac{1}{k!} \cdot \frac{\sum_{i=1}^A \sum_{j=1}^A r_{ij}^k}{A}$$

where r_{ij}^k is the k th power of the i - j entry of the geometry matrix and A the number of atoms (Figure M8) [Randić, 1995a, 1995b; Randić and Razinger, 1995b].

Using several increasing k values, a sequence of molecular invariants called *molecular profile* is obtained as

$$\{{}^1D, {}^2D, {}^3D, {}^4D, {}^5D, {}^6D, \dots\}$$

As the exponent k increases, the contributions of the most distant pairs of atoms become the most important.

The maximum nonzero value of ${}^kD$ is for the power corresponding to the number of atoms of the molecule ($k = A$); to obtain → *uniform-length descriptors*, values for $k > A$ are set at zero.

For large k values, ${}^kD$ tends to zero, due to the effect of the factorial normalization factor.

Another set of theoretical invariants can be obtained by averaging the row sums as

$${}^k d = \frac{1}{k!} \cdot \frac{\sum_{i=1}^A \sum_{j=1}^A r_{ij}^k}{A^2}$$

obtaining the vector

$$\{ {}^1 d, {}^2 d, {}^3 d, {}^4 d, {}^5 d, {}^6 d, \dots \}$$

For characterization of 2D structures, molecular profiles are computed in the same way by the $\rightarrow$ *distance distribution moments* of the topological $\rightarrow$ *distance matrix D*.

If one is interested in the characterization of molecular local features, that is, **local profiles**, the calculation of the ${}^k D$ values can be restricted to the local environment of interest, that is, only the row sums corresponding to atoms of interest are considered, obtaining a vector of local theoretical invariants:

$$\{ {}^1 L, {}^2 L, {}^3 L, {}^4 L, {}^5 L, {}^6 L, \dots \}$$

By this way, different types of profiles can be derived, such as **shape profiles**, which are local profiles taking into account only atoms on molecular periphery:

$$\{ {}^1 S, {}^2 S, {}^3 S, {}^4 S, {}^5 S, {}^6 S, \dots \}$$

In this case, the row sums of the geometry matrix are obtained by summing only the geometric distance powers of the atoms belonging to the periphery and the average is made by the number of the contributing atoms only. Each atomic distance sum is considered as a local indicator of molecular shape and each molecular invariant ${}^k S$ is considered a global shape descriptor.

In the case of 3D space-filled molecular models, one can represent the molecule by **contour profiles**, which are shape profiles calculated for all individual contours used to map the molecule. Each contour profile is then defined by a sequence:

$$\{ {}^1 C, {}^2 C, {}^3 C, {}^4 C, {}^5 C, {}^6 C, \dots \}$$

where each element of the profile is the normalized average row sum of an augmented geometry matrix, where additional points defining the contour are also considered.

Particular contour profiles are obtained by randomly distributed points over the surface of the molecule.

Moreover, an arbitrary number of points can be considered along the molecule bonds, thus deriving **bond profiles**:

$$\{ {}^1 B, {}^2 B, {}^3 B, {}^4 B, {}^5 B, {}^6 B, \dots \}$$

Bond profiles constitute a generalization of atomic molecular profiles since they provide a characterization of molecular connectivity, which is not explicitly contained in the geometry matrix [Randić, 1996a; Randić and Krilov, 1996].

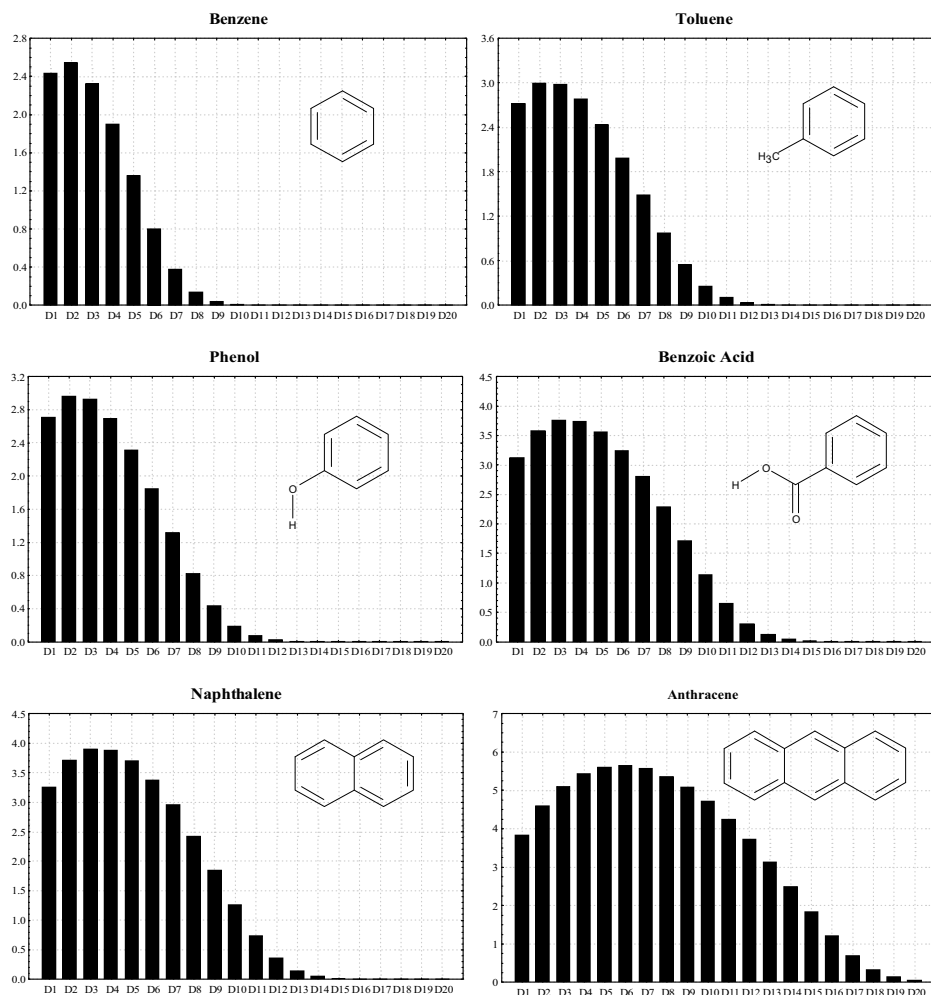


Figure M8 Molecular profiles for some compounds.

Volume profiles ${}^kV$ of a molecule can be calculated by distributing random points throughout the molecular interior defined by $\rightarrow$ *van der Waals molecular surface* and then constructing the corresponding augmented geometry matrices whose elements are raised at the k th power [Randić and Krilov, 1997b]. Moreover, to characterize the molecular surface, random points are restricted to the surface, thus obtaining **surface profiles** ${}^kSA$. Based on the same principles of the $\rightarrow$ *surface-volume ratio* $G' = SA/V$, the **volume-to-surface profiles** ${}^kV/{}^kSA$ have been proposed as $\rightarrow$ *shape descriptors* defined as the ordered sequence of the ratios of the volume over surface profile elements of corresponding order.

📖 [Randić, 1996c; Randić and Krilov, 1997a; Randić and Razinger, 1997; Zefirov and Tratch, 1997]

- **molecular pseudograph's adjacency matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **Molecular Quantum Self-Similarity Measures** → quantum similarity
- **molecular quantum similarity** ≡ *quantum similarity*
- **Molecular Quantum Similarity Indices** → quantum similarity
- **Molecular Quantum Similarity Measures** → quantum similarity
- **molecular representation** → molecular descriptors
- **molecular rigidity** → flexibility indices
- **molecular self-returning walk count** → self-returning walk counts
- **molecular sequence code** → sequence matrices
- **molecular sequence count** → sequence matrices

■ Molecular Shape Analysis (MSA)

A QSAR approach based on a set of methods, which combines molecular shape similarity and commonality measures with other → *molecular descriptors* to search both for similarities among molecules and build QSAR models [Hopfinger, 1980; Burke and Hopfinger, 1993]. The term *molecular shape similarity* refers to molecular similarity on the basis of a comparison of three-dimensional molecular shapes represented by some property of the atoms composing the molecule, such as the van der Waals spheres. The *molecular shape commonality* is the measure of molecular similarity when conformational energy and molecular shape are simultaneously considered [Hopfinger and Burke, 1990].

The main assumption of this approach is that the shape of the molecules is closely related to the shape of the → *binding site cavity* and, as a consequence, to the biological activity. Therefore, a shape reference compound is chosen, which represents the binding site cavity, and the similarity (or commonality) measured between the reference shape and the shape of other compounds is used to determine the biological activity of these compounds. Besides the shape similarity measures, other molecular descriptors such as those in → *Hansch analysis* can be used to evaluate the biological response. The MSA model is thus defined as

$$\hat{y}_i = b_0 + \sum_k b_k \cdot \Phi_{ik} + [f_0(M(i,j)) + \rho(n_j - n_i) - \beta \cdot \Delta E_i]$$

where i refers to any compound of the data set and j to the reference compound; $\hat{y}_i$ is the estimated biological response of the i th compound, usually expressed as a logarithm of the ligand inverse concentration; b_0 is a constant characteristic of the reference compound (j); Φ_k is any molecular descriptor representing → *physico-chemical properties* such as → *Hansch descriptors*, topological, geometrical, electronic, or thermodynamic features of the molecules. The last term (in squared brackets) is a 3D molecular structure term involving molecular shape and conformational thermodynamics; $f_0(M(i,j))$ is a molecular shape similarity function, that is, a function of the measure of the relative shape similarity between i and j , $\rho \cdot (n_j - n_i)$ is the difference in intramolecular conformational entropy (flexibility) between i and j , and $\beta \cdot \Delta E_i$ is a measure of the relative stability of the bioactive conformation of compound i with respect to its global intramolecular energy minimum. The quantity $I_c(i,j) = f_0(M(i,j)) - \beta \cdot \Delta E_i$ is the shape commonality index, which takes into account the balance between a gain in molecular shape similarity at the expense of loss in conformational stability.

There are seven operations involved in the MSA approach:

- (a) conformational analysis;
- (b) active conformation hypothesis;
- (c) shape reference compound selection;
- (d) pairwise molecular superimposition;
- (e) molecular shape similarity (or commonality) measure calculation;
- (f) other molecular descriptor calculation; and
- (g) trial QSAR model development.

For each MSA operation, there exists a set of choices that are experimented in the trial QSAR model; the final selection of the requirements for each MSA operation is based on optimizing the fitting ability of the QSAR model.

The most active compound is usually assumed as the reference structure, but also a set of overlapped structures can be assumed to define a reference shape.

Some **molecular shape similarity descriptors** (or **MSA descriptors**) are mentioned below. They represent a measure of the matching between the shapes of two molecules i and j , one of them being by definition the reference structure; the representation of molecular shape is given in different ways.

- **Common Overlap Steric Volume (COSV)**

Defined as the volume shared by two superimposed molecules, that is,

$$M_0(i,j) \equiv V_0(i,j) = V_i \cap V_j$$

where V_i and V_j are the $\rightarrow$ *van der Waals volume* of the i th and j th molecules, respectively.

Two arbitrary functions of the common overlap steric volume were also introduced as alternative molecular shape descriptors:

$$S_0 = V_0^{2/3} \quad \text{and} \quad L_0 = V_0^{1/3}$$

where S_0 is the **common overlap surface** (or **overlap surface**) and L_0 the **common overlap length**. Despite the terms, S_0 has the dimensions of area but is not a physical measure of the common surface area between two molecules, and the same holds for L_0 . Therefore, if the shape of the reference molecule is a good approximation for the acceptor site cavity, V_0 should measure the part of the cavity volume occupied by the considered ligand, whereas S_0 should be an approximation for the contact surface area of the ligand with receptor.

The **nonoverlap steric volume** V_{non} is another MSA descriptor defined as [Tokarski and Hopfinger, 1994]

$$V_{\text{non}}(i,j) = V_{ij} - V_j$$

where V_{ij} is the composite steric volume of the two aligned molecules i and j . In practice, the nonoverlap volume measures the regions of the i th molecule volume not shared by the reference compound, that is, it represents the $\rightarrow$ *steric misfit*.

- **atom-pair matching function**

Defined as

$$M_r(i, j) = \sum_{a=1}^{A_i} \sum_{a'=1}^{A_j} K_{aa'} \cdot r_{aa'}$$

where the sums run over all pairs of atoms of the two considered molecules i and j , $r_{aa'}$ is the interatomic distance between each pair of atoms from the i th and j th molecules, and $K_{aa'}$ is a user-defined constant providing the relative importance of the considered distance. For $M_r \rightarrow 0$, the superposition between i and j becomes better.

- **charge-matching function**

Defined as

$$M_c(i, j) = \sum_{a=1}^{A_i} \sum_{a'=1}^{A_j} \frac{q_a \cdot q_{a'}}{Q_T} \cdot r_{aa'}$$

where the sums run over all pairs of atoms of the two considered molecules i and j , q are atomic partial charges, r the $\rightarrow$ *interatomic distances* between atoms from molecules i and j , and Q_T is a normalization term calculated as

$$Q_T = \sum_{a=1}^{A_i} \sum_{a'=1}^{A_j} q_a \cdot q_{a'}$$

The partial charges q_a and $q_{a'}$ are assumed to always have the same sign, otherwise they would not be matched.

- **Integrated Spatial Difference in Field Potential (ISDFP)**

A field-based shape descriptor derived from the representation of molecular body by $\rightarrow$ *molecular interaction fields* and defined as

$$M_p(i, j) = \frac{1}{\Phi} \cdot \left[\int_{\Phi} [E_i(R, \Theta, \phi) - E_j(R, \Theta, \phi)]^2 \cdot d\Phi \right]^{1/2}$$

where E are the $\rightarrow$ *interaction energy values*, as measured by a probe, at the spherical coordinate position (R, Θ, ϕ) , and Φ is the considered integration volume. To calculate this descriptor, it is assumed that molecules i and j are superimposed.

- **weighted combination of COSV and ISDFP**

A combination of two complementary measures of shape similarity defined as

$$M_w(i, j) = w \cdot [(V_j \cap V_i) - M_0(i, j)] + (1 - w) \cdot M_p(i, j)$$

where $M_0(i, j)$ and $M_p(i, j)$ are the common overlap steric volume and the integrated spatial difference in field potential, and w is a weighting factor between zero and one. The two descriptors are considered complementary in the sense that the overlap volume measures the shape within the van der Waals surface formed by superimposition of i and j , whereas *ISDFP* measures the shape outside the van der Waals surface.

📖 [Battershell, Malhotra *et al.*, 1981; Hopfinger, 1981, 1983, 1984; Hopfinger and Potenzzone Jr, 1982; Mabilia, Pearlstein *et al.*, 1985; Walters and Hopfinger, 1986; Hopfinger, Compadre *et al.*, 1987; Rohrbaugh, Jurs *et al.*, 1988; Nagy, Tokarski *et al.*, 1994; Rowberg, Even *et al.*, 1994; Rhyu, Patel *et al.*, 1995; Holzgrabe and Hopfinger, 1996]

■ Molecular Shape Field (MSF)

The molecular shape field (MSF) is constituted by values of the → *molecular interaction potential* (MEP) of selected grid points that compose the molecular surface [Urbano-Cuadrado, Carbó *et al.*, 2007].

To obtain data suitable for later analysis, the local curvature values at each of the grid points of the molecular surface are computed using a cosine expression similar to that used by Pastor *et al.* [Pastor, Cruciani *et al.*, 2000]. The MSF values range between 0 and -1 for convex areas and 0 and $+1$ for concave ones.

From MSF and MEP values, → *autocorrelation descriptors* MSF–MSF, MEP–MEP, and MSF–MEP were proposed as molecular descriptors.

- **molecular shape similarity descriptors** → molecular shape analysis
- **molecular similarity matrices** → similarity/diversity

■ molecular structure

“... the term *molecular structure* represents a set of non-equivalent conceptual entities. There is no reason to believe that when we discuss different topics (e.g., organic synthesis, reaction rates theories, spectroscopic transitions, reaction mechanisms, *ab initio* calculations) using the concept of molecular structure, the different meaning we attach to the term *molecular structure* ultimately flows from the same concept” [Basak and Gute, 1997].

Together with the concepts of synthesis and chemical composition, the concept of molecular structure is one of the most fruitful of twentieth century scientific researches. This concept is conveniently studied by considering several levels of description, that is, the molecular structure is a part of a hierarchical system organized in different levels; to each level correspond characteristic language, properties, and relationships within its constitutional elements at that level as well as relationships between higher and lower levels. Thus, particles, atoms, molecules, compounds, cells, bodies, and so on are hierarchically organized levels of a complex system. At each level emergent properties arise from the organization of elements characterizing that level, that is, the presence of organizing relationships gives birth to new properties and constraints that influence the complexity of the system.

The above considerations also hold for different hierarchical descriptions of a system at a given level, that is, the same level is traversed by an inner hierarchical organization due to different descriptions of the same elements. The molecular representations are hierarchical descriptions of the molecular system; therefore, derived from the different representations of the molecular structure, several → *molecular descriptors* are calculated with different chemical information content.

Each molecular representation reflects hypotheses, ideas, a theory on unknown but supposed relevant relationships between molecules and their behavior. Much chemical research makes efforts to accurately predict properties, or to accurately classify chemical structures according to their properties, on the basis of chemical structure alone.

Each → *molecular representation* is a model that highlights only a part of the chemical reality, and, then, explaining only a part of the experimental evidence. Also a simple chemical formula

such as C_6H_5Cl already gives chemical information, at least about chemical composition and stoichiometric atom-type relationships.

Although chemical theories are the framework within which molecular structure has been developed, experimental properties define the reference framework in which the concept, or, better still, the concepts of molecular structure have been continuously verified, evaluated, and modified.

📖 [Woolley, 1978a, 1978b; Primas, 1981; Weininger, 1984; Turro, 1986; Wirth, 1986; Rouvray, 1989b; Weininger and Weininger, 1990; Ash, Warr *et al.*, 1991; Randić, 1992a; Wentang, Ying *et al.*, 1993; Dietz, 1995; Bauerschmidt and Gasteiger, 1997; Testa, Kier *et al.*, 1997; Ivanciuc, 2001a; Xu, 2003; Kuz'min, Artemenko *et al.*, 2005; Clark, Labute *et al.*, 2006]

- **molecular subgraph** → molecular graph
- **molecular supergraph** → hyperstructure-based QSAR techniques

■ molecular surface

The term molecular surface is usually referred to any surface surrounding some or all of the nuclei of the molecule. In the strict quantum mechanical sense, molecules do not have precisely defined surfaces; however, in analogy to macroscopic objects, the electron distribution may be regarded to as a 3D *molecular body* whose boundary is the molecular surface. In other words, the molecular surface can be viewed as the formal boundary that separates the 3D space into two parts: within the surface one is expected to find the whole molecule and beyond the rest of the universe [Meyer, 1986b, 1991c].

Different physical properties and molecular models have been used to define the molecular surface, the most common are reported below together with the descriptors proposed as measures of **surface areas** and molecular volume (→ *volume descriptors*). Molecular surface area and volume are parameters of molecules that are very important in understanding their structure and chemical behavior such as their ability to bind ligands and other molecules. An analysis of molecular surface shape is also an important tool in QSAR and → *drug design*, in particular both → *molecular shape analysis* and → *Mezey 3D shape analysis* were developed to search for similarities among molecules, based on their molecular shape.

• van der Waals molecular surface

The surface that envelops fused hard spheres centered at the atom coordinates (atomic nuclei) and having radii equal to some of the recommended values of the van der Waals radii. The spheres interpenetrate one another in such a way that the distance between the centers of two spheres equals the formal bond length.

In the hard-sphere model [Ciubotariu, Medeleanu *et al.*, 2004], the **van der Waals molecular surface** SA^{vdw} (also known as **Total molecular Surface Area, TSA**) is then defined as the exterior surface of the union of all such spheres in the molecule, that is, the area of the van der Waals molecular surface. It can be calculated by generating a uniform grid around each sphere of the molecule atoms, followed by the counting of the number of points generated on the surface n_s , consisting in the points that satisfy at least one of the following equalities:

$$(X_i - x)^2 + (Y_i - y)^2 + (Z_i - z)^2 \leq (R_i^{vdw})^2 \quad i = 1, \dots, A$$

where A is the number of atoms and R^{vdw} the van der Waals radius; X_i , Y_i , and Z_i are the coordinates of the i th atom and x , y , and z the coordinates of the generated points.

Then, the number of points n_e that are external to the surface have to be counted, that is, the number of points that do not satisfy the inequalities.

Therefore, the van der Waals surface SA_i^{vdw} of each i th atom is calculated as

$$SA_i^{vdw} = \frac{(n_e)_i}{n_s} \cdot 4 \cdot \pi \cdot (R_i^{vdw})^2$$

and the total van der Waals surface is calculated as the sum of the atomic van der Waals surfaces:

$$SA^{vdw} = \sum_{i=1}^A SA_i^{vdw}$$

• solvent-accessible molecular surface

In the case of large complex and folded molecular structures, a part of the van der Waals surface is buried in the interior and is thus inaccessible to solvent interactions, which mainly govern the chemical behavior of molecules in solution. Therefore, to obtain the best representation of the outer surface and overall shape of the molecule the solvent-accessible molecular surface was proposed. It was originally defined [Lee and Richards, 1971] as the surface across which the center of a spherical approximation of the solvent is passed when the solvent sphere is rolled over the van der Waals surface of the molecule (Figure M9). The radius (1.5 Å) of the solvent sphere is usually chosen to approximate the contact surface formed when a water molecule interacts with the considered molecule. If there are several grooves or minor cavities on the van der Waals surface where the rolling sphere cannot enter, then the solvent-accessible surface will be significantly different from the van der Waals surface (Figure M10).

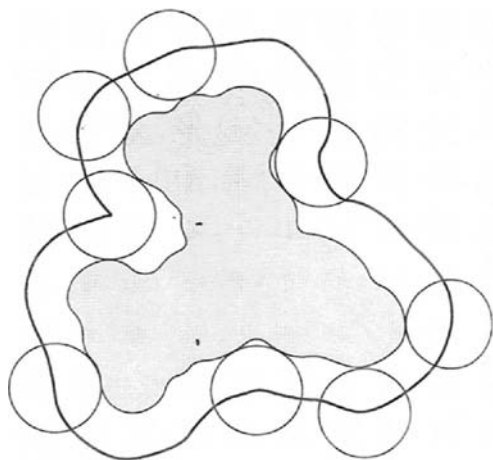


Figure M9 Solvent-accessible molecular surface defined by the centers of spheres rolled along the molecular contour surface. The radius of the sphere is chosen according to the size of the solvent molecule.

A few years later, Richards [Richards, 1977] gave a new definition of solvent-accessible surface, dividing it into two parts: the *contact surface* and the *reentrant surface*. The **contact surface** is that part of the van der Waals surface that is accessible to the probe sphere representing the solvent molecule. The **reentrant surface** comes from the inward-facing surface of the probe sphere when it is simultaneously in contact with more than one atom.

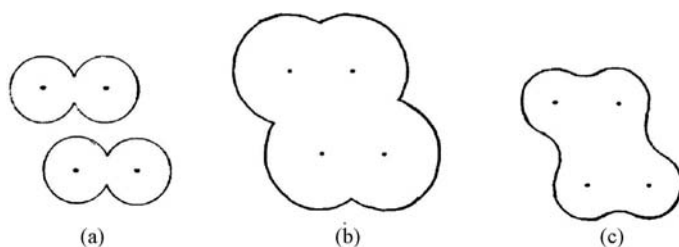


Figure M10 Comparison among (a) van der Waals surface, (b) solvent-accessible surface, and (c) contact surface.

The area of the solvent-accessible surface is called **Solvent-Accessible Surface Area**, *SASA* (or **Total Solvent-Accessible Surface Area**, *TSASA*). Several algorithms were proposed that implement both the first original definition of *SASA* and that of Richards. One of the most popular algorithms that implements Richards' solvent-accessible surface was proposed by Connolly [Connolly, 1983]. It is an analytical method for computing molecular surface, and is based on surface decomposition into a set of curved regions of spheres and tori that join at circular arcs; spheres, tori, and arcs are defined by analytical expressions in terms of atomic coordinates, van der Waals radii, and the probe radius. The molecular surface calculated in such a way is sometimes referred to as **Connolly surface area**. This algorithm also allows the calculation of solvent-accessible atomic areas.

An alternative to the hard sphere model is a recently proposed method for *SASA* calculations, based on atomic Gaussian functions describing the exposure of atoms and molecular fragments to solvent. A simple integral function of these atomic Gaussians is used to define a Gaussian neighborhood, which behaves in a complementary fashion to the conventional definition of solvent accessibility, that is, the smaller the Gaussian neighborhood, the more exposed the atom and hence the larger its accessibility [Grant and Pickup, 1995; Grant, Gallardo *et al.*, 1996].

Several $\rightarrow$ *charged partial surface area descriptors (CPSA)* and $\rightarrow$ *hydrogen-bond charged partial surface area descriptors (HB-CPSA)* are based on portions of the solvent-accessible surface area relative to polar or hydrophobic regions of the molecule, in some cases weighted by the corresponding local charges. Moreover, the **Hydrated Surface Area (HSA)** is the portion of the solvent-accessible surface area associated with hydration of polar functional groups.

The **Isotropic Surface Area (ISA)** is the surface of the molecule accessible to nonspecific interactions with the solvent, that is, the surface of the molecule involved in specific hydrogen-bonding with water is not considered [Collantes and Dunn III, 1995; Koehler, Grigoras *et al.*, 1988]. A hydration complex model needs to estimate the isotropic surface area. The **Polar Surface Area (PSA)** is defined as the part of the surface area of the molecule associated with

oxygen, nitrogen, sulfur, and the hydrogen bonded to any of these atoms. This surface descriptor is related to the hydrogen-bonding ability of compounds [Palm, Luthman *et al.*, 1998; Winiwarter, Bonham *et al.*, 1998] and is used to define some $\rightarrow$ *drug-like indices*.

The volume of space bounded by the solvent-accessible molecular surface is called **solvent-excluded volume** because it is the volume of space from which solvent is excluded by the presence of the molecule when the solvent molecule is also modeled as a hard sphere. Moreover, the **interstitial volume** is the volume consisting of packing defects between the atoms that are too small to admit a probe sphere of a given radius; in practice, it is calculated as the difference between the solvent-excluded volume and the van der Waals volume. An analytical method developed by Connolly was able to calculate the solvent-excluded volume [Connolly, 1983a]; several other numerical and analytical approaches have been proposed.

📖 [Silla, Tunon *et al.*, 1991; Hirono, Qian *et al.*, 1991]

- **electron isodensity contour surface**

The collection of all those points $\mathbf{r}$ of the space where the value of the $\rightarrow$ *electron density* $\rho(\mathbf{r})$ is equal to a threshold value m [Mezey, 1991b], that is,

$$G(m) = \{\mathbf{r} : \rho(\mathbf{r}) = m\}$$

Any positive value as threshold m can be chosen, even if a relatively small value is usually used to define a suitable molecular surface because the electron density converges rapidly to zero at short distances from the nuclei. The positive values of the threshold are due to the usual convention that a large negative charge means a large positive value of the electron density. For large values of m , the molecular surface is composed of several disconnected surfaces each surrounding one nucleus, whereas for too small values of m , the surface is an essentially spherical surface surrounding all of the nuclei and containing no information on the shape of the molecule.

- **molecular electrostatic potential contour surface**

The collection of all those points $\mathbf{r}$ of the space for which the value of the $\rightarrow$ *molecular electrostatic potential* (MEP) $V(\mathbf{r})$ is equal to a threshold value m [Mezey, 1991b], that is,

$$G(m) = \{\mathbf{r} : V(\mathbf{r}) = m\}$$

The contour parameter m as well as the electrostatic potential can take both positive and negative values. An analysis of the shape of MEP surfaces is of particular interest in $\rightarrow$ *drug design* as the electrostatic potential has a marked influence on the binding interactions between ligand and receptor. Moreover, the sum of all the surface minima values of the electrostatic potential, denoted as $\sum V_S^-$, was proposed as a molecular descriptor able to account for lipophilicity [Zou, Zhao *et al.*, 2002].

- **molecular orbital contour surface**

A molecular surface defined as the contour surface of individual molecular orbitals such as HOMO and LUMO, other frontier orbitals, or localized and delocalized orbitals [Mezey, 1991b]. In practice, it is the collection of all those points $\mathbf{r}$ of the space for which the value of the electronic wavefunction $\Psi(\mathbf{r})$ of the considered molecular orbital is equal to a threshold

value m , that is,

$$G(m) = \{\mathbf{r} : \psi(\mathbf{r}) = m\}$$

The contour parameter m can take both positive and negative values.

📖 [Hermann, 1972; Amidon, Yalkowsky *et al.*, 1975; Arteca, Jammal *et al.*, 1988b; Leicester, Finney *et al.*, 1988; Marsili, 1988; Lipkowitz, Baker *et al.*, 1989; Pascual-Ahuir and Silla, 1990; Valkó and Slegel, 1992; Brusseau, 1993; Leicester, Finney *et al.*, 1994a, 1994b; Schüürmann, 1995; Lee, Kwon *et al.*, 1996; Palm, Luthman *et al.*, 1996; Brickmann, 1997; Hermann, 1997; Randić and Krilov, 1997b; Zweerszeilmaker, Horbach *et al.*, 1997; Whitley, 1998; Jørgensen, Jensen *et al.*, 2001; Deanda and Pearlman, 2002; King, 2002]

■ molecular surface interaction terms (MSI)

These constitute a set of molecular descriptors including the molecular surface area and empirically derived descriptors accounting for dispersion, polar, and hydrogen-bonding interactions [Grigoras, 1990]. They were proposed to empirically express the molecular surface energy using atomic contributions to the total molecular surface. The molecular surface interaction terms are

$$A = \sum_i SA_i \quad A_- = \sum_{a-} SA_a \cdot b_a \cdot q_a^- \quad A_+ = \sum_{a+} SA_a \cdot b_a \cdot q_a^+ \quad A_{HB} = \sum_i SA_i \cdot b_i \cdot q_i^H$$

where A is the $\rightarrow$ *total molecular surface area* calculated as the sum of all the atomic surface areas SA_i ; this is a dispersion molecular surface interaction term. A_- is the electrostatic negative molecular surface interaction term calculated as the sum of surface areas of negatively charged atoms multiplied by their corresponding scaled $\rightarrow$ *net atomic charge* q_a^- . A_+ is the electrostatic positive molecular surface interaction term calculated as the sum of surface areas of positively charged atoms multiplied by their corresponding scaled net atomic charge q_a^+ . A_{HB} is the hydrogen-bonding molecular surface interaction term calculated as the sum of the surface areas of hydrogen-bonding hydrogen atoms multiplied by their corresponding scaled net atomic charge q_i^H . The coefficient b_i is an empirical charge scaling factor, which is the same for the atoms of the same chemical type (Table M12).

Table M12 Charge scaling factors b for different atom chemical types.

Atom/hybrid	Coefficient b	Atom/hybrid	Coefficient b
H (at C_{sp^3})	3.29	N_{sp^3}	0.155
H (at C_{sp^2})	7.77	N_{sp}	1.59
H_{HB}	1.00	N_{AROM}	2.79
C_{sp^3}	1.00	O_{sp^3}	1.32
C_{sp^2}	0.00	O_{sp^2}	1.51
C_{sp}	2.33	F	0.00
C_{AROM}	9.25	Cl	1.78

The hydrogen atom considered for the calculation of the hydrogen bonding term are not taken into account in the A_+ term.

- **molecular topological index** $\equiv$ *Schultz molecular topological index*
- **molecular topological indices** $\equiv$ *topological indices* $\rightarrow$ graph invariants

■ molecular transforms

These are descriptors based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves [Soltzberg and Wilkins, 1976, 1977]. A generalized scattering function can be used as the functional basis for deriving, from a known molecular structure, the specific analytic relationship of both X-ray and electron diffraction. The general molecular transform is

$$G(\mathbf{s}) = \sum_{i=1}^A f_i \cdot \exp(2\pi i \cdot \mathbf{r}_i \cdot \mathbf{s})$$

where $\mathbf{s}$ represents the scattering in various directions by a collection of A atoms located at points $\mathbf{r}_i$; f_i is a form factor taking into account the direction dependence of scattering from a spherical body of finite size. The scattering parameter s has the dimension of a reciprocal distance and depends on the scattering angle as

$$s = \frac{4\pi}{\lambda} \cdot \sin(\vartheta/2)$$

where ϑ is the scattering angle and λ the wavelength of the electron beam.

Usually, the above equation is used in a modified form as suggested in 1931 by Wierl [Wierl, 1931]. On substituting the form factors by an $\rightarrow$ *atomic property* w_i , considering the molecule to be rigid and setting the instrumental constant equal to one, the following function, usually called **radial distribution function**, is used to calculate molecular transforms:

$$I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}}$$

where $I(s)$ is the scattered electron intensity, w an atomic property, chosen as the atomic number Z by Soltzberg and Wilkins [Soltzberg and Wilkins, 1976], r_{ij} the $\rightarrow$ *geometric distance* between the i th and j th atom, and A the number of atoms in the molecule. The sum is performed over all the pairs of atoms in the molecule.

Soltzberg and Wilkins introduced a number of simplifications to obtain a binary code. Only the zero crossing of the $I(s)$ curve, that is, the s values at which $I(s) = 0$, in the range $1\text{--}31 \text{ \AA}^{-1}$ were considered. The s range was then divided into 100 equal-sized bins, each described by a binary variable equal to 1 if the bin contains a zero crossing, 0 otherwise. Thus, a vectorial descriptor consisting of 100 bins was finally calculated for each molecule.

Gabányi *et al.* proposed a modified molecular transform by replacing the geometric distance r_{ij} with the $\rightarrow$ *topological distance* d_{ij} [Gabanyi, Surjan *et al.*, 1982]. Moreover, two different functions of the scattering parameter s were evaluated to be used in place of the trigonometric term:

$$(a) \quad I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot e^{-1/2 \cdot (s \cdot d_{ij})^2} \quad (b) \quad I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot \left(1 - \frac{s \cdot d_{ij}}{\Delta_{ij}}\right)$$

where d_{ij} is the topological distance between atoms v_i and v_j , Δ_{ij} the corresponding $\rightarrow$ *detour distance* (i.e., the length of the largest path between two atoms), and w is a chosen atomic property [Csorvassy and Totser, 1991].

Raevsky and coworkers applied the molecular transform to study ligand–receptor interactions by using hydrogen-bond abilities, hydrophobicity, and charge of the atoms, instead of the atomic numbers Z [Novikov and Raevsky, 1982]. For each atomic property, a spectrum of interatomic distances, **interatomic interaction spectrum**, was derived to represent the 3D structure of molecules and the scattered intensities in selected regions of the spectrum were used as the molecular descriptors [Raevsky, Dolmatova *et al.*, 1995]. These $\rightarrow$ *vectorial descriptors* are based on local characteristics of different pairs of centers in the molecule. For a selected distance R , the following function is evaluated [Raevsky, Dolmatova *et al.*, 1995; Raevsky, 1997a, 1977b; Raevsky, Trepalin *et al.*, 2000]:

$$I(R) = \sum_{r^{\min}}^{r^{\max}} \sum_{i=1}^A \sum_{j=1}^A \frac{w_i \cdot w_j}{1 + \sqrt{\frac{(R-r_{ij})^2}{0.1}}} \quad i \neq j$$

where A is the number of atoms in the molecule, w_i and w_j are $\rightarrow$ *atomic properties* of the i th and j th atom, respectively, r_{ij} is the geometric interatomic distance; $r^{\min}$ and $r^{\max}$ define a distance range around R , which accounts for vibrations of atoms and allows to obtain a band instead of a line in the final spectrum for each pair of centers defined by R . Distances R are varied from 1.1 to 20  with step 0.1 , resulting in a total of 190 signals per spectrum.

Superimposition of all the bands for all the possible pairs of centers forms the final interatomic interaction spectrum. Seven types of spectrum are calculated for each molecule by using different atomic properties w . These include *atomic van der Waals radius* (steric interaction spectrum), *atomic charges* (spectrum of interactions between positively charged atoms, spectrum of interactions between negatively charged atoms, and spectrum of interactions of positively charged atoms with negatively charged atoms), *hydrogen-bond abilities* (spectrum of interactions between hydrogen-bond donors, spectrum of interactions between hydrogen-bond acceptors, and spectrum of interactions of hydrogen-bond donors with hydrogen-bond acceptors).

The **integrated molecular transform** (FT_m) is a molecular descriptor calculated from the square of the molecular transform, by integrating the squared molecular transform in a selected interval of the scattering parameter s to obtain the area under the curve and finally taking the square root of the area [King, Kassel *et al.*, 1990, 1991]. The square root of the integrated molecular transform, called **SQRT index**, was also proposed as molecular descriptor [Famini, Kassel *et al.*, 1991].

Applications of integrated molecular transforms found in literature are: [King and Kassel, 1992; King, 1993, 1994; Molnar and King, 1995, 1998; King and Molnar, 1996, 1997, 2000].

To calculate **3D-MoRSE descriptors** (*3D-MOLEcule Representation of Structures based on Electron diffraction*, or simply **MoRSE descriptors**), Gasteiger *et al.* [Schuur and Gasteiger, 1996, 1997] returned to the initial $I(s)$ curve and maintained the explicit form of the curve. As the atomic $\rightarrow$ *weighting scheme* w , various $\rightarrow$ *physico-chemical properties* such as atomic mass, partial atomic charges, and atomic polarizability were considered. To obtain $\rightarrow$ *uniform-length descriptors*, the intensity distribution $I(s)$ was made discrete, calculating its value at a sequence of evenly distributed values of, for example, 32 or 64 in the range of 1–31 ^{–1}. Clearly, the more the values are chosen, the finer the resolution in the representation of the molecule.

Applications of 3D-MoRSE descriptors found in literature are: [Gasteiger, Sadowski *et al.*, 1996; Schuur, Selzer *et al.*, 1996a, 1996b; Gasteiger, Schuur *et al.*, 1997; Baumann, 1999; Jelcic, 2004;

Pérez González and Moldes Teran, 2004; Pérez González, Helguera Morales *et al.*, 2004; Caballero and Fernández, 2006; Helguera Morales, Perez *et al.*, 2006; Saiz-Urra, Pérez González *et al.*, 2006, 2007; Yap, Li *et al.*, 2006].

RDF descriptors (or **Radial Distribution Function descriptors**) were proposed based on a radial distribution function different from that commonly used to calculate molecular transforms $I(s)$ [Hemmer, Steinhauer *et al.*, 1999; Selzer, Gasteiger *et al.*, 2000]. The radial distribution function selected here is that quite often used for interpretation of the diffraction patterns obtained in powder X-ray diffraction experiments.

Formally, the radial distribution function of an ensemble of A atoms can be interpreted as the probability distribution to find an atom in a spherical volume of radius R . The general form of the radial distribution function is represented by

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot e^{-\beta \cdot (R-r_{ij})^2}$$

where f is a scaling factor, w characteristic atomic properties of the atoms i and j , r_{ij} the interatomic distance between the i th and j th atom, and A the number of atoms. The exponential term contains the distance r_{ij} between the atoms i and j and the smoothing parameter β , which defines the probability distribution of the individual interatomic distances; β can be interpreted as a temperature factor that defines the movement of atoms. $g(R)$ is generally calculated at a number of discrete points with defined intervals. An RDF vector of 128 values was proposed, using a step size for R about 0.1–0.2 Å, whereas the β parameter is fixed in the range between 100 and 200 Å⁻². By including characteristic atomic properties w of the atoms i and j , RDF descriptors can be used in different tasks to fit the requirements of the information to be represented. These atomic properties enable the discrimination of the atoms of a molecule for almost any property that can be attributed to an atom.

The radial distribution function in this form meets all the requirements for a 3D structure descriptor. It is independent of the number of atoms, that is, the size of a molecule, it is unique regarding the three-dimensional arrangement of the atoms, and invariant against translation and rotation of the entire molecule. In addition, the RDF descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space, for example, to describe sterical hindrance or structure/activity properties of a molecule.

Moreover, the RDF vectorial descriptor is interpretable by using simple rules and, thus, it provides a possibility of $\rightarrow$ *reversible decoding*. Besides information about distribution of interatomic distances in the entire molecule, the RDF vector provides further valuable information; for example, about bond distances, ring types, planar and nonplanar systems, and atom types. This fact is a most valuable consideration for a computer-assisted code elucidation.

To account for stereochemistry of molecules, the $\rightarrow$ *Chirality Code* was proposed as a modification of the RDF code [Aires-de-Sousa and Gasteiger, 2001].

Applications of RDF descriptors reported in literature are: [Razdol'skii, Trepalin *et al.*, 2000; Yan and Gasteiger, 2003; Caballero and Fernández, 2006; Helguera Morales, Perez *et al.*, 2006; Podlipnik, Solmajer *et al.*, 2006; Saiz-Urra, Pérez González *et al.*, 2006, 2007; Schuffenhauer, Brown *et al.*, 2006; Yap, Li *et al.*, 2006; Hristozov, Da Costa *et al.*, 2007].

- **molecular volume** $\rightarrow$ volume descriptors (⊙ molar volume)
- **molecular volume index** $\rightarrow$ volume descriptors

- **molecular walk count** → walk counts
- **molecular weight** → physico-chemical properties
- **molecule center** $\equiv$ center of a molecule
- **MOLORD algorithm** → iterated line graph sequence

■ MOLMAP descriptors

MOLMAP (*MO*lecular *MA*p of *At*om-level *Pr*operties) descriptors are uniform-length → *vectorial descriptors* derived by mapping physico-chemical properties of all the bonds in a molecule into a 2D Kohonen → *self-organizing map* (SOM) [Zhang and Aires-de-Sousa, 2005; Gupta, Metthew *et al.*, 2006]. These descriptors encode local features of a chemical structure, being calculated on the basis of properties of single elements in a molecule, such as bonds.

A Kohonen map consists of a set of neurons (i.e., vectors of weights) organized into a square grid, each having as many weights as the number of input variables. In the MOLMAP approach, objects used for training the neural network are chemical bonds and the input variables are seven selected properties of bonds: resonance of stabilization, difference between the σ electronegativity of two bonded atoms, difference between the total charge of two bonded atoms, difference between the π charge of two bonded atoms, mean bond polarizability, bond dissociation energy, and bond polarity. These physico-chemical properties were chosen to encode information on the chemical reactivity of compounds, which is related to propensity for bond breaking and bond making.

All the bonds in all the molecules in the data set in analysis are used for network training. As some properties depend on the bond orientation, each bond is taken twice with different orientation. To focus on functional groups, only bonds involving a heteroatom, or an atom of a π system, can be considered. Once the training has been completed, the SOM provides similarities among chemical bonds. Indeed, similar bonds are mapped into the same or closely adjacent neurons.

By using a trained SOM, bonds in a molecule are mapped into the SOM and the pattern of activated neurons is interpreted as a fingerprint of the bonds of the molecule. For numerical processing, each neuron is assigned a value equal to the number of times it was activated by bonds of the molecule. The map, that is, a matrix, is then transformed into a vector by concatenation of columns. This vector is called MOLMAP descriptor. To account for proximity relationship, a value of 0.3 is added to each neuron multiplied by the number of times a neighbor was activated by a bond.

Unlike the common molecular descriptors, MOLMAP descriptors are data set dependent, which means that their values for a molecule change if another training set is used for SOM training or a different map size is chosen. However, their use in QSAR applications can lead to the identification of structural features responsible for the molecular property in analysis.

MOLMAP descriptors were originally proposed for automatic classification of chemical reactions with the name **reaction MOLMAPs** [Zhang and Aires-de-Sousa, 2005; Latino and Aires-de-Sousa, 2006]. These descriptors are calculated as the difference map between the MOLMAPs of the products of a reaction and the MOLMAPs of the reactants of the same reaction. This difference map can be interpreted as the reaction fingerprint. Zero values in the difference map are related to bonds far apart from the reaction center, remaining unchanged during the reaction; negative values concern bonds of the reactants that brake or change properties in the reaction; positive values concern new bonds appearing in the products. If more reactants (products) are involved in the reaction, the MOLMAPs of all reactants (products) are numerically summed.

Self-organizing map for describing chemical information of a molecule is also used in → *Comparative Molecular Surface Analysis* and → *topological feature maps*.

MOLMAP descriptors were used to predict mutagenicity (positive or negative Ames test) by CART classification tree and random forest [Zhang and Aires-de-Sousa, 2007]. Combined with other global molecular descriptors, error percentage of 15 and 16% were achieved for an external data set with 472 compounds and for the training set with 4038 compounds, respectively. They were also applied in modeling the radical scavenging activity of 47 naturally occurring phenolic antioxidants by counterpropagation neural networks obtaining a cross-validated Q^2 of 0.71 [Gupta, Matthew *et al.*, 2006].

- **MOLPRINT-2D fingerprints** → substructure descriptors (⊙ fingerprints)
- **MOLPRINT-3D fingerprints** → substructure descriptors (⊙ pharmacophore-based descriptors)

■ MolSurf descriptors

MolSurf descriptors comprise a set of physico-chemical properties estimated by quantum-chemical calculations [Norinder, Österberg *et al.*, 1997; Sjöberg, 1997; Norinder, Sjöberg *et al.*, 1998].

MolSurf descriptors include $\log P$, $\log D$, pK_a , polarizability, polarity, number and strength of H-bond acceptor nitrogen and oxygen atoms, number of H-bond donor atoms, and charge-transfer characteristics for all carbon atoms. The definition of specific substituents for which descriptors are calculated is also allowed. Moreover, MolSurf software includes a module allowing the construction of QSAR models by the Partial Least Squares (PLS) regression.

📖 [Norinder, Österberg *et al.*, 1999; Alifrangis, Christensen *et al.*, 2000; Egan, Merz Jr *et al.*, 2000; Stenberg, Norinder *et al.*, 2001; Norinder and Haeberlein, 2002; Nordqvist, Nilsson *et al.*, 2004]

- **moments about the mean** → statistical indices (⊙ moment statistical functions)

■ moments indices

These are molecular descriptors defined in terms of the weighted absolute central moment of first order, which is a statistical quantity used to measure variability of a distribution around a center. They are defined as

$$M(w) = \frac{\sum_{i=1}^A w_i \cdot r_i}{\sum_{i=1}^A w_i}$$

where w is an → *atomic property* and r the distance of an atom from the geometric center of the molecule; A is the number of atoms in the molecule.

The moment index based on the atomic weights m_i was called **normalized molecular moment**, M_n , and defined as [King and Molnar, 1997]

$$M_n = \frac{\sum_{i=1}^A m_i \cdot r_i}{MW}$$

where MW is the $\rightarrow$ *molecular weight* and r the distance of the atom from the geometric center of the molecule. This descriptor is a measure of absolute deviation of the distribution of the atomic masses and is similar to the $\rightarrow$ *radius of gyration*, which is defined in terms of the second-order central moments. Moreover, other moment indices encoding information on electronic features of the molecule were derived from quantum-chemical calculations, by replacing the atomic masses with atomic electron densities and charges [Molnar and King, 1998; King and Molnar, 2000].

- **moment of inertia** $\rightarrow$ principal moments of inertia
- **moment statistical functions** $\rightarrow$ statistical indices
- **Monge-Arrault-Marot-Morin-Allory scoring functions** $\rightarrow$ scoring functions
- **Monte Carlo version of MTD** $\rightarrow$ minimal topological difference
- **Moran coefficient** $\rightarrow$ autocorrelation descriptors
- **Moreau–Broto autocorrelation** $\rightarrow$ autocorrelation descriptors
- **Moreau chirality index** $\rightarrow$ chirality descriptors
- **morphological similarity** $\rightarrow$ Compass method
- **morphologic index** $\rightarrow$ functional coordination index
- **Morgan's extended connectivity algorithm** $\rightarrow$ canonical numbering
- **Moriguchi model based on structural parameters** $\equiv$ *MLOGP* $\rightarrow$ lipophilicity descriptors
- **Moriguchi model based on surface area** $\rightarrow$ lipophilicity descriptors
- **Moriguchi polar parameter** $\rightarrow$ lipophilicity descriptors

■ Morovitz information index (I_{MOR})

An information index accounting for the structural features of a molecule [Morovitz, 1955]. It is defined as

$$I_{MOR} = I_{AC} + I_{PB}$$

where the first term is the $\rightarrow$ *total information index on atomic composition* and the second term is the **information on the possible valence bonds** I_{PB} defined as

$$I_{PB} = \sum_{g=1}^G A_g \cdot \log_2 V_g$$

where g runs over all the different atom types, A_g is the number of atoms of g th type, and V_g the number of possible bonds that can be formed by an atom of g th type, calculated as

$$V_{g,\delta} = \binom{6 + \delta - 1}{\delta} = \frac{(6 + \delta - 1)!}{\delta! \cdot 5!}$$

where 6 is assumed as the maximum possible valence and δ is the actual valence of the g th-type atom. For example, $V_H = 6$, $V_O = 21$, $V_{N,3} = 56$, and $V_{C,4} = 126$ [Bonchev, 1983].

- **MoRSE descriptors** $\equiv$ *3D-MoRSE descriptors* $\rightarrow$ molecular transforms
- **motor octane number** $\rightarrow$ technological properties
- **Mozley similarity coefficient** $\equiv$ *Forbes–Mozley similarity coefficient* $\rightarrow$ similarity/diversity (Table S9)

- **MPEI** $\equiv$ *Molecular Polarizability Effect Index* $\rightarrow$ electric polarization descriptors ($\odot$ polarizability effect index)
- **M-PEOE** $\equiv$ *modified partial equalization of orbital electronegativities* $\rightarrow$ electronegativity
- **MP-MFP descriptors** $\rightarrow$ substructure descriptors ($\odot$ structural keys)

■ MPR approach

This is an approach designed to calculate $\rightarrow$ *local vertex invariants* (LOVIs) as the solutions of a linear equation system [Filip, Balaban *et al.*, 1987; Ivanciuc, Balaban *et al.*, 1992]:

$${}^a\mathbf{M} \cdot \mathbf{s} = \mathbf{r} \quad {}^a\mathbf{M} = \mathbf{M} + \mathbf{p} \cdot \mathbf{I}$$

where ${}^a\mathbf{M}$ is a square $A \times A$ matrix representing the $\rightarrow$ *molecular graph* and defined according to the scheme for the $\rightarrow$ *augmented matrices*, $\mathbf{p}$ an A -dimensional column vector containing weights for graph vertices that are used as diagonal elements of the matrix $\mathbf{M}$, $\mathbf{r}$ the A -dimensional column vector of atomic properties, $\mathbf{I}$ the identity matrix, and $\mathbf{s}$ the A -dimensional column vector that is the solution of the system.

MPR (*Matrix–Property–Response*) descriptors are thus the elements of the vector **MPR** calculated as

$$\mathbf{MPR} \equiv \mathbf{s} = ({}^a\mathbf{M}^T \cdot {}^a\mathbf{M})^{-1} \cdot {}^a\mathbf{M}^T \cdot \mathbf{r}$$

The vertex properties encoded in the column vectors $\mathbf{p}$ and $\mathbf{r}$ can be either topological, for example, $\rightarrow$ *vertex degree*, $\rightarrow$ *vertex distance degree*, or chemical, for example, atomic number, $\rightarrow$ *electronegativity*, and $\rightarrow$ *ionization potential*.

Among the LOVIs obtained by this general approach, the most known are **AZV descriptors** derived from the $\rightarrow$ *adjacency matrix* $\mathbf{A}$ whose diagonal elements are substituted by the atomic numbers Z_i and the A -dimensional vector $\mathbf{r}$ containing the vertex degrees δ_i .

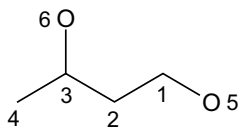
Different sets of LOVIs can be obtained by different choices of matrices and vectors defining the linear equation system; several combinations were studied on linear alkanes (Table M13).

Table M13 A, adjacency matrix; D, distance matrix; V, vector of vertex degrees δ_i ; S, vector of distance sums σ_i ; Z, vector of atomic numbers; N, vector of numbers of graph vertices; 1, unit vector. LOVI range values for linear alkanes.

ID	MPR	LOVIs range	ID	MPR	LOVIs range
1	AZV	0.1–1	11	DSN	0.05–0.7
2	ASV	0.01–0.2	12	DN ² N	0.06–0.2
3	DSV	–0.02–0.12	13	ANS	1–4
4	AZS	2–9	14	ANV	0.08–0.5
5	ASZ	0.1–1	15	AZN	0.3–1.5
6	DN ² S	0.1–0.3	16	ANZ	0.5–1.7
7	DN ² 1	0–0.09	17	AN1	0.1–0.3
8	AS1	0.02–0.1	18	DSZ	0.06–0.6
9	DS1	0–0.3	19	ANN	0.7–0.9
10	ASN	0.2–0.7	20	DN ² Z	0.03–0.5

Example M5

AZV descriptors for the H-depleted molecular graph of 1,3-butandiol.

	Atom	1	2	3	4	5	6	×	Atom	s_i	=	Atom	δ_i
	1	6	1	0	0	1	0		1	s_1		1	2
	2	1	6	1	0	0	0		2	s_2		2	2
	3	0	1	6	1	0	1		3	s_3		3	3
	4	0	0	1	6	0	0		4	s_4		4	1
	5	1	0	0	0	8	0		5	s_5		5	1
	6	0	0	1	0	0	8		6	s_6		6	1

$\begin{cases} 6 \cdot s_1 + s_2 + s_5 = 2 \\ s_1 + 6 \cdot s_2 + s_3 = 2 \\ s_2 + 6 \cdot s_3 + s_4 + s_6 = 3 \\ s_3 + 6 \cdot s_4 = 1 \\ s_1 + 8 \cdot s_5 = 1 \\ s_3 + 8 \cdot s_6 = 1 \end{cases} \Rightarrow$	$\begin{aligned} s_1 &\equiv \text{AZV}_1 = 0.28284 \\ s_2 &\equiv \text{AZV}_2 = 0.21334 \\ s_3 &\equiv \text{AZV}_3 = 0.43708 \\ s_4 &\equiv \text{AZV}_4 = 0.09382 \\ s_5 &\equiv \text{AZV}_5 = 0.08996 \\ s_6 &\equiv \text{AZV}_6 = 0.07036 \end{aligned}$

Triplet topological indices (*TTI*) or, simply, **triplet indices**, are derived from local vertex invariants calculated by the MPR approach by using the common functions defined for the calculation of $\rightarrow$ *graph invariants*. The most frequent functions used to generate triplet TIs are [Basak, Balaban *et al.*, 2000; Basak, Gute *et al.*, 2003]:

1. $TTI_1(\text{MPR}) = \sum_{i=1}^A \text{MPR}_i^\lambda$
2. $TTI_2(\text{MPR}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\text{MPR}_i \cdot \text{MPR}_j)^\lambda$
3. $TTI_3(\text{MPR}) = A \cdot \left(\prod_{i=1}^A \text{MPR}_i \right)^{1/A}$

where λ is a real exponent, usually taking values 1, 1/2, or 2 in function 1 and value $-1/2$ in function 2; A is the number of graph vertices and a_{ij} are the elements of the adjacency matrix.

Closely related to MPR descriptors are local vertex invariants called **graph potentials** and denoted by U_i [Golender, Drboglav *et al.*, 1981; Ivanciuc, Balaban *et al.*, 1992]. They are calculated as the solutions of a linear equation system defined as

$$\mathbf{W} \cdot \mathbf{s} = \mathbf{r}$$

where $\mathbf{W}$ is a weighted graph-theoretical matrix, $\mathbf{r}$ the A -dimensional column vector of atomic properties, and $\mathbf{s}$ the A -dimensional column vector of solutions of the system, which are local invariants U_i .

The weighted matrix $\mathbf{W}$ is defined as

$$[\mathbf{W}]_{ij} = \begin{cases} w_i + \sum_k w_{ik} & \text{if } i = j \\ -w_{ij} & \text{if } (i, j) \in E(G) \\ 0 & \text{if } (i, j) \notin E(G) \end{cases}$$

where w_i is any topological or chemical semipositive definite atomic property and the sum runs over the first neighbors of the i th atom; w_{ij} is any topological or chemical semipositive definite bond weight, and $E(G)$ the set of edges of the molecular graph G . If the weights w are all set equal to one, then the W matrix is

$$W = V - A + I = L + I$$

where V , A , I , and L are the diagonal $\rightarrow$ *vertex degree matrix*, the adjacency matrix, the identity matrix, and the $\rightarrow$ *Laplacian matrix*, respectively.

Similar to graph potentials, another set of LOVIs was proposed based on the $\rightarrow$ *geometry matrix* G , using as the diagonal terms the $\rightarrow$ *Balaban distance connectivity index* and as the response vector the adjacency matrix A multiplied by the column vector z collecting the atomic numbers of all the non-hydrogen atoms [Beteringhe, Filip *et al.*, 2005]:

$$MPR = {}^aG^{-1} \cdot A \cdot z \quad \text{and} \quad {}^aG = G + J \cdot I$$

where ${}^aG$ is the augmented geometry matrix; it must be noted that the diagonal terms of the geometry matrix are filled in by a constant term (the Balaban index J of the molecule and not by local vertex invariants). Moreover, the atomic properties r_i are obtained from the adjacency matrix and the atomic numbers Z_i as

$$r_i \equiv [A \cdot z]_i = \sum_{j=1}^A a_{ij} \cdot Z_j$$

where the summation accounts for the atomic numbers of vertices adjacent to the i th vertex.

From these local vertex invariants r_i , a molecular descriptor, called **Beteringhe–Filip–Tarko descriptor** and denoted as $GJ(AZ)$, was proposed as

$$GJ(AZ) = \frac{B}{A+B} \cdot \sum_{i=1}^A \log(r_i)^2$$

where A is the number of graph vertices and B the number of edges.

Note. The authors called this index as topological, although it depends on the molecular geometry.

📖 [Balaban, 1993a, 1994b]

- **MPS topological index** $\equiv$ *detour index* $\rightarrow$ detour matrix
- **MSA descriptors** $\equiv$ *molecular shape analysis descriptors* $\rightarrow$ molecular shape analysis
- **MS-WHIM descriptors** $\rightarrow$ grid-based QSAR techniques ($\odot$ G-WHIM descriptors)
- **MTD-ADJ method** $\rightarrow$ minimal topological difference
- **MTD descriptors** $\rightarrow$ minimal topological difference
- **MTD-MC method** $\rightarrow$ minimal topological difference
- **MTD model** $\rightarrow$ minimal topological difference
- **MTI' index** $\equiv$ *S index* $\rightarrow$ Schultz molecular topological index
- **MTD-PLS method** $\rightarrow$ minimal topological difference
- **m th order sparse matrix** $\rightarrow$ algebraic operators ($\odot$ sparse matrices)
- **Mulliken electronegativity** $\rightarrow$ atomic electronegativity

- **Mulliken population analysis** → quantum-chemical descriptors
- **MULTICASE** → lipophilicity descriptors (⊙ Klopman hydrophobic models)
- **multicriteria decision making** → chemometrics (⊙ ranking methods)
- **multigraph** → graph
- **multigraph distance degree** → weighted matrices (⊙ weighted distance matrices)
- **multigraph distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **multigraph factor** $\equiv$ *atomic multigraph factor* → bond order indices (⊙ conventional bond order)
- **multigraph information content indices** $\equiv$ *indices of neighborhood symmetry*
- **MultiLevel Chemical Compatibility** → scoring functions
- **Multilevel Neighborhoods of Atoms descriptors** → substructure descriptors (⊙ fingerprints)
- **multiple arc** → graph
- **multiple bond count** → multiple bond descriptors

■ multiple bond descriptors

The presence of multiple bonds in molecules is a fundamental chemical aspect, which characterizes molecular properties and reactivity.

The **bond multiplicity** m_{ij} represents the degree of bonding between two adjacent vertices v_i and v_j and the most common way to quantify it is by using → *bond orders* derived from quantum-chemical calculations or → *conventional bond orders*.

The most simple descriptors of the degree of unsaturation of a molecule are → *count descriptors* based on the presence of double bonds, triple bonds, and aromatic bonds; they are the **double-bond count (DB)**, the **triple-bond count (TB)**, and the **aromatic-bond count (AB)**.

The **MCB index** was proposed as the number of multiple CC bonds in the molecule accounting for double, triple, and aromatic bonds [Bakken and Jurs, 1999a] as

$$MCB = DB + TB + AB$$

→ *Partial Wiener indices* are other multiple bond descriptors derived by a splitting of the → *Wiener index* into different multiple bond contributions.

Among the first proposed simple multiple bond descriptors [Pellegrin, 1983], there are the **number of unsaturation sites (US)** defined as the number of double bonds plus the number of triple bonds in the molecule, that is, $US = DB + TB$, and the **unsaturation number (UN)** defined as the number of double bonds plus twice the number of triple bonds, that is, $UN = DB + 2 \times TB$. Moreover, the same author proposed the **degree of unsaturation (DU)** by also considering the number of rings C (the → *cyclomatic number*), that is, $DU = DB + 2 \times TB + C = UN + C$.

The unsaturation index UN can be expressed by a more general form, called **multiple bond count**, as

$$b^* = \sum_b \left(\pi_{ij}^* \right)_b - B$$

where π^* is the → *conventional bond order* and the summation runs over all B bonds. For saturated compounds, $b^* = UN = 0$.

The **formal oxidation number** of a carbon atom equals the sum of the → *conventional bond orders* with electronegative atoms; the C–N bond order in pyridine may be considered as 2 since

there is one such bond and 1.5 when there are two such bonds; the C–X bond order in pyrrole or furan may be considered as 1.

The **unsaturation index** *UI* was also defined as

$$UI = \log_2(1 + b)$$

where *b* is calculated as

$$b = 2N_C + 2 - N_H - N_X + N_N + N_P + 2(N_{O-S} - N_{SO_3})/2 - C$$

N_C , N_H , N_X , N_N , N_P , and C are the number of carbon atoms, hydrogen, halogen, nitrogen, phosphorous, and independent cycles, respectively. N_{O-S} and N_{SO_3} are the number of oxygen atoms bonded to sulfur and the number of SO_3 groups, respectively. When no sulfur atoms are present, this index can be easily calculated from the chemical formula; otherwise, it is coincident with the index calculated replacing *b* with b^* , the multiple bond count.

A general expression [Pellegrin, 1983], valid for any organic compound, was also given by defining the atom valencies as the following:

Symbol	Atom	Symbol	Atom
α	Monovalent	δ	Tetravalent
β	Divalent	ϵ	Pentavalent
γ	Trivalent	ξ	Hexavalent

The total number of atoms in a compound is

$$A = \alpha + \beta + \gamma + \delta + \epsilon + \xi$$

and the total number of bonds is

$$B = \frac{1}{2} \cdot \alpha + \frac{2}{2} \cdot \beta + \frac{3}{2} \cdot \gamma + \frac{4}{2} \cdot \delta + \frac{5}{2} \cdot \epsilon + \frac{6}{2} \cdot \xi$$

Then, starting from the $\rightarrow$ *Euler's formula* for a graph, corresponding to the usual expression for the calculation of the number of rings, that is, the $\rightarrow$ *cyclomatic number* *C*:

$$C = B - A + 1$$

the degree of unsaturation was defined as

$$DU = -\frac{\alpha}{2} + \frac{\gamma}{2} + \delta + \frac{3}{2} \cdot \epsilon + 2 \cdot \xi + 1$$

The sum of all the multiple bonds, that is, the *MCB* index, plus the number of rings is called **index of hydrogen deficiency** (*IHD*) or **double bond equivalents** (*DBE*). This last index can be derived from the following general equation [Pellegrin, 1983; Badertscher, Bsichofberger *et al.*, 2001]:

$$IHD = 1 + \frac{1}{2} \cdot \left[\sum_{i=1}^A (v_i - 2) \right]$$

where *A* is the number of atoms and v_i is the formal valence of the *i*th atom ($\alpha, \beta, \dots, \xi$).

One drawback of this index is that the formal valence of each element must be known. This is not a problem with most of the organic molecules containing only C, H, N, O, and halogens, but could become a problem for molecules containing sulfur and phosphorous. Moreover, it cannot be applied to radicals, ions, and disjoint parts. Finally, it is not invariant to different molecular representations.

Unlike the index of hydrogen deficiency, the **degree of unsaturation** has the same value for any structural representation corresponding to a molecular formula and can be calculated for much variety of structure representations. In this approach, the valence electrons of an element are partitioned into bond electrons and electrons localized on an atom, as shown in Table M14.

Table M14 Number of valence electrons, bond electrons, and localized valence electrons for some chemical elements.

Atom	No. valence electrons	Std. no. bond electrons	Std. no. localized valence electrons
H	1	1	0
Li	1	1	0
C	4	4	0
N	5	3	2
O	6	2	4
Halogens	7	1	6
Si	4	4	0
P	5	3	2
S	6	2	4

Then, the degree of unsaturation is defined for the molecular formula as [Badertscher, Bsichofberger *et al.*, 2001]

$$DU = 1 + \frac{1}{2} \cdot \left[-Q + \sum_{i=1}^A (b_i - 2) \right]$$

where Q is total charge of the molecule (signed) and b_i the standard number of bond electrons of the i th element (Table M14). For any structural molecule representation, DU is defined as

$$DU = DB + 2 \cdot TB + C + (1 - D) + \frac{1}{2} \cdot ELE$$

where DB and TB are the number of double and triple bonds, respectively, C the number of molecule rings, the so-called $\rightarrow$ *cyclomatic number*, D the number of disconnected parts. ELE is the number of excess localized electrons of a molecule, calculated as the difference between the actual number of electrons localized on all atoms and the sum of the standard numbers, as given in column 4 of Table M14.

Another unsaturation measure is the **Unsat index** proposed as [Zheng, Luo *et al.*, 2005]:

$$\text{Unsat} = NRG_{567} + DB + 2 \cdot TB + \frac{AB + 1}{2}$$

where NRG_{567} is the number of 5-, 6-, and 7-member rings, DB the number of double bonds, TB the number of triple bonds, and AB the number of aromatic bonds. A relative unsaturation

measure called **Unsat-p index** was also proposed as the ratio of the Unsat index to the number of atoms that do not have bonded hydrogens and halogens.

A multiple bond descriptor was proposed in terms of $\rightarrow$ *valence vertex degree* δ^v , called **DV index**, and defined as

$$DV = \sum_b \left[(\delta_i^v)^{-1/2} + (\delta_j^v)^{-1/2} \right]_b$$

where the sum runs over all the multiple bonds and i and j denote the atoms forming the considered bond [Millership and Woolfson, 1980].

The **induction parameter** was proposed to estimate the interaction ability of polar and nonpolar groups in the molecule [Thomas and Eckert, 1984]; it is based on the degree of unsaturation and defined as

$$q_{ind} = 1 - \frac{DB}{A}$$

where DB is the number of double bonds in the molecule. For saturated molecules, $q_{ind} = 1$. $\rightarrow$ *Multigraph information content indices* are $\rightarrow$ *information indices* encoding the bond multiplicity in the molecules.

To take into account the absolute contribution that a single double-bond makes to the whole size and shape of alkene molecules, **second-grade structural parameters** were derived from a $\rightarrow$ *molecular graph* [Zhang, Liu *et al.*, 1997]. The topological descriptors representing the size w and the shape $P_{\bar{w}}$ related to the presence of a double-bond are, respectively, as

$$w = \frac{\sum_{i=1}^A (d_{ik} + d_{il})}{2W} \quad P_{\bar{w}} = {}^3f_k + {}^3f_l$$

where W is the $\rightarrow$ *Wiener index*, that is, the total sum of distances in the molecular graph, and d represents the $\rightarrow$ *topological distance* between two vertices; k and l denote the vertices incident with the considered double-bond and the sum runs over all A vertices of the molecular graph. In the shape descriptor $P_{\bar{w}}$, ${}^3f_k$ and ${}^3f_l$ are the number of vertices at a distance 3 from vertices v_k and v_l , respectively, that is, their $\rightarrow$ *vertex distance count*. The size descriptor w is derived from the Wiener index, whereas $P_{\bar{w}}$ from the $\rightarrow$ *polarity number*. An extension giving information about the presence of several double bonds can be the sum of the w and $P_{\bar{w}}$ values defined above over all double bonds.

- **multiple correlation coefficient** $\rightarrow$ regression parameters
- **multiple edge** $\rightarrow$ graph
- **multiple graph** $\equiv$ *multigraph* $\rightarrow$ graph
- **multiple pharmacophore descriptors** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **multiplicative Wiener index** $\rightarrow$ Wiener index
- **multivariate K correlation index** $\rightarrow$ statistical indices ($\odot$ correlation measures)
- **multivariate entropy** $\rightarrow$ model complexity ($\odot$ information content ratio)
- **mutation and selection uncover models** $\rightarrow$ variable selection
- **mutation graph** $\equiv$ *Sachs graph* $\rightarrow$ graph
- **MVI** $\equiv$ *molecular volume index* $\rightarrow$ volume descriptors

N

- **Narumi–Katayama index** → vertex degree
- **NASAWIN descriptors** → substructure descriptors (⊙ fingerprints)
- **Natural Bond Orbital analysis** → quantum-chemical descriptors
- **NCD descriptors** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **ND indices** → spectral indices (⊙ A_{xi} eigenvalue indices)

■ Nearest Neighboring Code (NNC)

This is a → *local vertex invariant* derived both from the → *H-filled molecular graph* and the → *Graph of Atomic Orbitals* and defined to distinguish the different atom types in the framework of the → *OCWLI* approach [Toropov and Toropova, 2002a; Toropov, Nesterov *et al.*, 2003a]. From the H-filled molecular graph, it is calculated as

$$NNC_i = 100 \cdot \delta_i + 10 \cdot \delta_i(\text{C}) + \delta_i(\text{H})$$

where δ_i is the total number of neighbors of the i th atom, that is, its → *vertex degree*, and $\delta_i(\text{C})$ and $\delta_i(\text{H})$ are the number of neighbors of the i th atom, which are carbon and hydrogen atoms, respectively.

The Nearest Neighboring Code of the i th vertex in the Graph of Atomic Orbitals is calculated as [Toropov and Toropova, 2004]:

$$^{GAO}NNC_i = 100 \cdot \delta_i + 10 \cdot \delta_i(2p^2) + \delta_i(1s^1)$$

where δ_i is the total number of neighbors of the i th atom, and $\delta_i(2p^2)$ and $\delta_i(1s^1)$ are the number of neighbors that correspond to $2p^2$ atomic orbitals and the number of neighbors that correspond to $1s^1$ atomic orbitals in the graph, respectively.

- **negative predictive value** → classification parameters
- **negentropy** $\equiv$ *total information content* → information content
- **neighborhood-distance map matrix** → biodescriptors (⊙ proteomics maps)
- **neighborhood Euclidean matrix** $\equiv$ *neighborhood geometry matrix* → molecular geometry
- **neighborhood geometry matrix** → molecular geometry
- **neighborhood information content** → indices of neighborhood symmetry
- **neighborhood matrices** → matrices of molecules
- **neighborhood of a vertex** → indices of neighborhood symmetry

- **neighborhood total information content** → indices of neighborhood symmetry
- **net atomic charges** $\equiv$ *atomic charges* → charge descriptors
- **NICS index** → delocalization degree indices
- **Nikolić-Trinajstić-Randić index** → Wiener index
- **NOAEL** $\equiv$ *No-Observed-Adverse-Effect Level* → biological activity indices (⊙ toxicological indices)
- **NOEL** $\equiv$ *No-Observed-Effect Level* → biological activity indices (⊙ toxicological indices)
- **nonadjacent number** → Hosoya Z index
- **nonerror rate** → classification parameters
- **non-Omega polynomial** → Omega polynomial
- **nonoverlapping volume** → minimal topological difference
- **nonoverlap steric volume** → molecular shape analysis (⊙ common overlap steric volume)
- **No-Observed-Adverse-Effect Level** → biological activity indices (⊙ toxicological indices)
- **No-Observed-Effect Level** → biological activity indices (⊙ toxicological indices)
- **normal boiling point** → physico-chemical properties (⊙ boiling point)
- **normalized atomic walk count** → walk counts
- **normalized centric index** → centric indices (⊙ Balaban centric index)
- **normalized extended connectivity** → canonical numbering (⊙ Morgan's extended connectivity algorithm)
- **normalized fragment topological indices** → fragment topological indices
- **normalized Laplacian matrix** → Laplacian matrix
- **normalized molecular moment** → moment indices
- **normalized number of ring systems** → ring descriptors
- **normalized quadratic index** $\equiv$ *quadratic index* → Zagreb indices
- **normalized root mean square deviation** $\equiv$ *normalized root mean square error* → regression parameters
- **normalized root mean square error** → regression parameters
- **normalized Szeged property matrices** → Szeged matrices
- **normalized vertex distance complexity** → topological information indices
- **normalized Wiener index** → Wiener index
- **Norrington lipophilic constant** → lipophilicity descriptors (⊙ Hansch–Fujita hydrophobic substituent constants)
- **N_i index** → spectral indices
- **nuclear information content** → atomic information indices
- **nucleophilic atomic frontier electron density** → quantum-chemical descriptors
- **nucleophilic charge** → quantum-chemical descriptors (⊙ nucleophilic atomic frontier electron density)
- **nucleophilic frontier electron density index** → quantum-chemical descriptors (⊙ nucleophilic atomic frontier electron density)
- **nucleophilic indices** → reactivity indices
- **nucleophilicity–electrophilicity index** → quantum-chemical descriptors (⊙ Fukui functions)
- **nucleophilic substituent constant** → electronic substituent constants (⊙ resonance electronic constants)
- **nucleophilic superdelocalizability** → quantum-chemical descriptors
- **nucleus-independent chemical shift index** $\equiv$ *NICS index* → delocalization degree indices

- **number of atoms in substituent specific positions** → steric descriptors
- **number of ring systems** → ring descriptors
- **number of terms in the model** → model complexity
- **Nys-Rekker hydrophobic fragmental constants** → lipophilicity descriptors



■ **OASIS method** ($\equiv$ *Optimized Approach based on Structural Indices Set*)

The OASIS method is based on the same assumptions as $\rightarrow$ *Hansch analysis* and can be regarded to as an extended and optimized version of the Hansch approach [Mekenyan and Bonchev, 1986; Mekenyan, Karabunarliev *et al.*, 1990a].

Besides $\rightarrow$ *substituent constants*, the descriptors used in the OASIS approach are $\rightarrow$ *topological indices*, $\rightarrow$ *geometrical descriptors*, $\rightarrow$ *steric descriptors*, and $\rightarrow$ *electronic descriptors*, both for molecules and their fragments.

The OASIS model is defined as:

$$\text{biological activity} = f(\{\text{TI}\}, \{\Phi_1\}, \{\Phi_2\}, \{\Phi_3\})$$

where $\{\Phi_i\}$ are different sets of $\rightarrow$ *physico-chemical properties*, each set representing steric, electronic, and hydrophobic descriptors, respectively, and $\{\text{TI}\}$ is a set of topological descriptors.

This equation allows for the possibility that more than one descriptor of the different factors contribute to the overall biological activity. Moreover, these model components can be either local, that is, referring to atoms or fragments, or global, that is, describing the molecule as a whole.

📖 [Mekenyan, Bonchev *et al.*, 1986, 1988b; Mekenyan, Peitchev *et al.*, 1986a, 1986b; Mekenyan, Karabunarliev *et al.*, 1990b; Mercier, Mekenyan *et al.*, 1991; Bonchev, Mountain *et al.*, 1993; Mekenyan, Mercier *et al.*, 1993; Bonchev, Seitz *et al.*, 1994; Kamenska, Mekenyan *et al.*, 1996]

- **object** $\rightarrow$ data set
- **occupancy numbers** $\rightarrow$ cell-based methods
- **Ochiai similarity coefficient** $\equiv$ *cosine similarity coefficient* $\rightarrow$ similarity/diversity (☉ Table S9)
- **octanol-water distribution coefficient** $\rightarrow$ physico-chemical descriptors (☉ partition coefficients)
- **octanol-water partition coefficient** $\rightarrow$ physico-chemical descriptors (☉ partition coefficients)
- **octupole moment** $\rightarrow$ electric polarization descriptors
- **OCWLI** $\equiv$ *Optimization of Correlation Weights of Local Invariants* $\rightarrow$ variable descriptors
- **odd-even index** $\rightarrow$ Cao-Yuan indices
- **OEI** $\equiv$ *odd-even index* $\rightarrow$ Cao-Yuan indices
- **oligocenter** $\rightarrow$ center of a graph

■ Omega polynomial

The Omega polynomial is a $\rightarrow$ counting polynomial based on the counting of the so-called “quasiorthogonal cuts” (*qoc*) of a graph as

$$\Omega(G; x) = \sum_c m(G; c) \cdot x^c$$

where the coefficients $m(G; c)$ are the numbers of occurrences of quasiorthogonal cuts of length c (i.e., the number of edges cutoff) and the summation runs over the maximal length of *qoc* in the graph [Diudea, 2006; Diudea, Vizitiu *et al.*, 2007; Diudea, Cigher *et al.*, 2008].

The quasiorthogonal cuts are defined as the following. Two edges $e(u, v)$ and $e'(u', v')$ are called *codistant* (e co e') if for $k = 0, 1, 2, \dots$ there exist the relationships

$$d(u, u') = d(v, v') = k \quad \text{and} \quad d(u, v') = d(v, u') = k + 1$$

or *vice versa*. For some edges of a connected graph the following relationships may exist:

1. e co e'
2. e co $e' \Leftrightarrow e' \text{ co } e$
3. e co $e' \quad \wedge \quad e$ co $e'' \Leftrightarrow e' \text{ co } e''$

although the third relationship is not always valid.

Now, let $C(e)$ denote the set of all edges of the graph that are codistant to the edge e , that is, equidistant and “topologically” parallel. If all the elements of $C(e)$ satisfy the relationships 2 and 3, the $C(e)$ is called *orthogonal cut* (*oc*) of the graph and the graph is called *co-graph* if and only if the set $\mathcal{E}$ of all edges in the graph is the union of disjoint orthogonal cuts:

$$C_1 \cup C_2 \cup \dots \cup C_k = \mathcal{E} \quad \wedge \quad C_i \cap C_j = \emptyset \quad \text{for } i \neq j \text{ and } i, j = 1, 2, \dots, k$$

If any two consecutive edges of a cut edge sequence are codistant (i.e., obeying the relations 2 and 3) and belong to one and the same face of the covering, such a sequence is called a *quasiorthogonal cut* (*qoc*) strip. A *qoc* strip starts and ends either out of G (at an edge with end points of degree lower than 3, if G is an open lattice) or in the same starting polygon (if G is a closed lattice).

From the Omega polynomial, two molecular descriptors are derived. The first one, called **Cluj-Ilmenau index**, denoted as CI , is derived from the first Ω^I and second Ω^{II} derivatives of the Omega polynomial as

$$CI = (\Omega^I)^2 - (\Omega^I + \Omega^{II}) \quad \text{at } x = 1$$

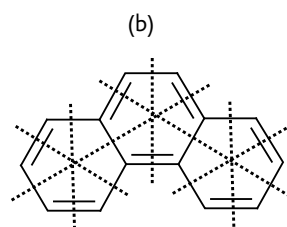
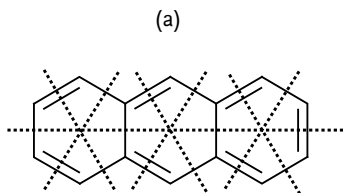
This index coincides with the $\rightarrow$ *PI index* for polycyclic graphs.

The second index, denoted as I_Ω , is calculated from the summation of all the possible Omega polynomial derivatives Ω^d at $x = 1$, and normalized to the first polynomial derivative (which is equal to the number of edges in the graph):

$$I_\Omega = \frac{1}{\Omega^I} \cdot \sum_d (\Omega^d)^{1/d} \quad \text{at } x = 1$$

Example O1

Omega polynomials, their derivatives, and CI and I_Ω indices for anthracene (a) and phenanthrene (b). PI is the PI index. Straight line segments indicate orthogonal edge cuts.



$$\Omega(a) = 6 \cdot x^2 + x^4$$

$$\Omega^I(a) = 12 \cdot x + 4 \cdot x^3 \quad \Omega^{II}(a) = 12 + 12 \cdot x^2$$

$$\Omega^{III}(a) = 24 \cdot x \quad \Omega^{IV}(a) = 24$$

$$CI(a) = 256 - (16 + 24) = 216 = PI(a)$$

$$I_\Omega(a) = \frac{1}{16} \cdot (16 + 24^{1/2} + 24^{1/3} + 24^{1/4}) = 1.624$$

$$\Omega(b) = 5 \cdot x^2 + 2 \cdot x^3$$

$$\Omega^I(b) = 10 \cdot x + 6 \cdot x^2 \quad \Omega^{II}(b) = 10 + 12 \cdot x$$

$$\Omega^{III}(b) = 12 \quad \Omega^{IV}(b) = 0$$

$$CI(b) = 256 - (16 + 22) = 218 = PI(b)$$

$$I_\Omega(b) = \frac{1}{16} \cdot (16 + 22^{1/2} + 12^{1/3}) = 1.436$$

The **non-Omega polynomial** is derived from the Omega polynomial so that to be a complementary quantity as

$$N\Omega(G; x) = \sum_c m(G; c) cx^{(B-c)}$$

where the coefficients $m(G; c)$ are the occurrences of quasiorthogonal cuts of length c and the summation runs over the maximum length of qoc in the graph; B is the total number of edges in the graph [Diudea, Vizitiu *et al.*, 2007]. The first derivative of the non-Omega polynomial at $x = 1$ coincides with the $\rightarrow PI$ index.

- **optimal descriptors** $\equiv$ *variable descriptors*
- **optimization** $\rightarrow$ chemometrics
- **Optimization of Correlation Weights of Local Invariants** $\rightarrow$ variable descriptors
- **Optimized Approach based on Structural Indices Set** $\equiv$ *OASIS method*
- **orbital electronegativity** $\rightarrow$ quantum-chemical descriptors ($\odot$ electronic chemical potential)

■ orbital information indices ($\bar{I}_{ORB}$)

The first information index proposed as a measure of complexity of a $\rightarrow$ *H-depleted molecular graph* is the **vertex orbital information content** $\bar{I}_{ORB}$, originally simply called **topological information content** $\bar{I}_{TOP}$ [Rashevsky, 1955; Trucco, 1956a, 1956b], defined as $\rightarrow$ *mean information content*:

$$\bar{I}_{ORB} \equiv \bar{I}_{TOP} = - \sum_{g=1}^G \frac{A_g}{A} \cdot \log_2 \frac{A_g}{A}$$

where A is the total number of vertices and A_g is the number of topologically equivalent vertices of g th type; the summation runs over all G different classes of topological equivalence. The atoms are distinguished not on the basis of their chemical nature but on their topological relationships to each other. Vertices are topologically equivalent if they belong to the same $\rightarrow$ *orbits* of the vertex $\rightarrow$ *automorphism group* of the graph. Note that in the original definition of the topological information content proposed by Rashevsky, the equivalence classes were defined by the $\rightarrow$ *vertex degree* of the neighboring vertices instead of the graph orbits.

The **total topological information content** was defined as

$$I_{TOP} = A \cdot \bar{I}_{TOP} = A \cdot \log_2 A - \sum_{g=1}^G A_g \cdot \log_2 A_g$$

where A is the number of vertices in the graph.

The **edge orbital information content** was defined by analogy as [Trucco, 1956a, 1956b]

$${}^E\bar{I}_{ORB} = - \sum_{g=1}^G \frac{B_g}{B} \cdot \log_2 \frac{B_g}{B}$$

where B is the total number of edges in the graph and B_g is the number of edges belonging to the g th edge orbit of the graph; the summation runs over all G different edge orbits. The corresponding **total edge orbital information content** was defined as:

$${}^E I_{ORB} = B \cdot {}^E\bar{I}_{ORB} = B \cdot \log_2 B - \sum_{g=1}^G B_g \cdot \log_2 B_g$$

Moreover, based on the connection orbits of the graph the **connection orbital information content** was also defined as [Bonchev, 1983]

$${}^{\text{CONN}}\bar{I}_{ORB} = - \sum_{g=1}^G \frac{(N_2)_g}{N_2} \cdot \log_2 \frac{(N_2)_g}{N_2}$$

where N_2 is the $\rightarrow$ *connection number* of the graph and $(N_2)_g$ is the number of connections belonging to the g th orbit of the graph; the summation runs over all G different connection orbits. The corresponding **total connection orbital information content**, defined as

$${}^{\text{CONN}}I_{ORB} = N_2 \cdot {}^{\text{CONN}}\bar{I}_{ORB} = N_2 \cdot \log_2 N_2 - \sum_{g=1}^G (N_2)_g \cdot \log_2 (N_2)_g$$

is a component of the $\rightarrow$ *Bertz complexity index* when calculated for $\rightarrow$ *multigraphs*.

The orbital information indices can also be calculated for multigraphs, giving a higher measure of graph complexity as the multiplicity of the edges provides more graph orbits than the simple graph. In the case of graph vertices of the same chemical element, the orbital information content for multigraph coincides with the $\rightarrow$ *neighborhood information content* of maximal order calculated for the H-depleted molecular graph:

$$\bar{I}_{ORB} = IC_{\max}$$

- **orbital interaction graph of linked atoms** → determinant-based descriptors ($\odot$ general a_N -index)
- **orbital interaction matrix of linked atoms** → determinant-based descriptors ($\odot$ general a_N -index)
- **orbits** → graph
- **ordered structural code** → self-returning walk counts
- **ordered walk count molecular code** → walk counts
- **order of a subgraph** → molecular graph
- **order of neighborhood** → indices of neighborhood symmetry
- **order parameter** → polymer descriptors
- **oriented graph** → graph ($\odot$ digraph)
- **orthogonal descriptors** $\equiv$ *orthogonalized descriptors*

■ **orthogonalized descriptors** ($\equiv$ *orthogonal descriptors*)

These are obtained by applying an orthogonalization procedure to a selected set of → *molecular descriptors*. A descriptor X_j is made orthogonal to another descriptor X_i simply by regressing it against X_i and using as the new orthogonal descriptor the residual $X_j - \hat{X}_j$, where $\hat{X}_j$ is the value of X_j calculated by the regression model. The residual, that is, the orthogonalized descriptor, represents the part of descriptor X_j not explained by the descriptor X_i ; the symbol ${}^i\Omega_j$ is usually used to indicate that descriptor X_j is made orthogonal to descriptor X_i . The orthogonalization process is an iterative procedure until the last considered descriptor is orthogonalized against the preceding orthogonalized descriptors [Randić, 1991b, 1991e, 1991f].

The orthogonalization procedure requires a prior ordering of the descriptors to which other descriptors are subsequently made orthogonal, that is, it requires → *basis descriptors*. Thus, it is evident that different orthogonalized descriptor bases derive from different ordering. However, in the case of path counts, connectivity indices, → *uniform length descriptors*, that are naturally ordered descriptors the orthogonalization procedure can be applied easily and the symbol Ω_j used simply for orthogonalized descriptors.

The most known orthogonalization technique is the Gram–Schmidt orthogonalization scheme [Golub and van Loan, 1983]. When a subset of descriptors has the same partial ordering they are simultaneously and mutually orthogonalized within the set itself after sequential orthogonalization to the preceding descriptors [Klein, Randić *et al.*, 1997]. Canonical orthogonalization, symmetrical orthogonalization, and optimal orthogonalization are techniques to perform this task.

To better explain this procedure a sequence of → *path counts* ${}^mP$ is considered. **Orthogonalized path counts** are calculated through the following steps: (a) The first path count ${}^1P$ in the sequence is chosen as the first orthogonalized descriptor Ω_1 ; it can also be decomposed into the mean contribution represented by Ω_0 and the deviation from the mean represented by Ω_1 . (b) The second orthogonalized descriptor Ω_2 is calculated as a residual of the regression between the second path count ${}^2P$ against the first one ${}^1P$; the residual contains information independent of ${}^2P$. (c) The orthogonalization continues by regressing the path number ${}^3P$ against the first path count ${}^1P$, and then the residual of this regression against the orthogonalized descriptor Ω_2 ; the residual of the last regression is the third orthogonalized descriptor Ω_3 . In general, the p th orthogonalized descriptor Ω_p is defined as the residual in the multiple regression of the p th descriptor of the sequence against the $p - 1$ previously orthogonalized descriptors [Šoškić, Plavšić *et al.*, 1996b].

Orthogonalized descriptors are used in $\rightarrow$ *similarity/diversity* analysis and quantitative $\rightarrow$ *structure/response correlations* with the aim of eliminating the bias provided by the interdependence of common molecular descriptors. Moreover, the interpretation of regression models should be facilitated as the information encoded in each descriptor is unique.

The similarity/diversity analysis based on previously orthogonalized descriptors is usually called **orthosimilarity** [Randić, 1996b].

📖 [Randić, 1991a, 1991g, 1993b, 1994a; Randić and Seybold, 1993; Pogliani, 1994a, 1994c, 1995b; Randić and Trinajstić, 1994; Amić, Davidović-Amić *et al.*, 1995b, 1997; Lučić, Nikolić *et al.*, 1995a, 1995b, 1995c; Araujo and Morales, 1996a, 1996b, 1998; Šoškić, Plavšić *et al.*, 1996a, 1996b; Mracec, Muresan *et al.*, 1997; Nikolić and Trinajstić, 1998; Ivanciuc, Taraviras *et al.*, 2000; Ivanciuc, Ivanciuc *et al.*, 2000c; González Díaz, Marrero *et al.*, 2003; Fernández *et al.*, 2004; Fernández, Duchowicz *et al.*, 2004; Du, Liang *et al.*, 2005]

- **orthogonalized path counts** $\rightarrow$ orthogonalized descriptors
- **orthogonal Wiener operator** $\rightarrow$ Wiener-type indices
- **orthosimilarity** $\rightarrow$ orthogonalized descriptors
- **Ostwald solubility coefficient** $\rightarrow$ physico-chemical properties ($\odot$ partition coefficients)
- **outdegree** $\rightarrow$ graph
- **ovality index** $\rightarrow$ shape descriptors
- **overall accuracy** $\equiv$ *nonerror rate* $\rightarrow$ classification parameters
- **overall connectivity index** $\rightarrow$ molecular complexity ($\odot$ Bonchev topological complexity indices)
- **overall degree of clustering of a graph** $\rightarrow$ adjacency matrix
- **overall electronic constants** σ_m and σ_p $\rightarrow$ electronic substituent constants
- **overall topological indices** $\equiv$ *Bonchev topological complexity indices* $\rightarrow$ molecular complexity
- **overlap surface** $\equiv$ *common overlap surface* $\rightarrow$ molecular shape analysis ($\odot$ common overlap steric volume)
- **overall Wiener indices** $\rightarrow$ molecular complexity ($\odot$ Bonchev topological complexity indices)
- **overall Zagreb indices** $\rightarrow$ molecular complexity ($\odot$ Bonchev topological complexity indices)

P

- **PAD descriptors** $\equiv$ *PEST Autocorrelation Descriptors* $\rightarrow$ TAE descriptor methodology
- **Padmakar–Ivan index** $\equiv$ *PI index* $\rightarrow$ Szeged matrix
- **pair correlation cutoff selection** $\rightarrow$ variable reduction
- **Palm steric constant** $\rightarrow$ steric descriptors ($\odot$ Taft steric constant)
- **parachor** $\rightarrow$ physico-chemical properties
- **Para-Delocalization Index** $\rightarrow$ delocalization degree indices ($\odot$ Delocalization Index)
- **partial atomic charge** $\rightarrow$ quantum-chemical descriptors
- **partial charge weighted topological electronic index** $\rightarrow$ charge descriptors
- **partial equalization of orbital electronegativities** $\rightarrow$ electronegativity
- **partial local invariant** $\rightarrow$ iterated line graph sequence
- **partial negative surface area** $\rightarrow$ charged partial surface area descriptors
- **partial-order ranking methods** $\rightarrow$ chemometrics ($\odot$ ranking methods)
- **partial positive surface area** $\rightarrow$ charged partial surface area descriptors
- **partial Wiener indices** $\rightarrow$ Wiener index
- **partition-based methods** $\equiv$ *cell-based methods*
- **partition coefficients** $\rightarrow$ physico-chemical properties
- **Pasaréti index** $\equiv$ *all-path Wiener index* $\rightarrow$ path counts

■ PASS ($\equiv$ *Prediction of Activity Spectra of Substances*)

The computer system PASS was built to predict several hundreds of biological activities (main and side pharmacological activities, $\rightarrow$ *mode of action*, mutagenicity, carcinogenicity, teratogenicity, and embryotoxicity) [Filimonov and Poroikov, 1996, 2001; Poroikov, Filimonov *et al.*, 2000, 2003; Anzali, Barnickel *et al.*, 2001]. Most of the biological active compounds reveal a wide spectrum of different effects. Some of them are useful in the treatment of defined diseases, while others cause various side and toxic effects. The whole complex of activities caused by the compounds is called “biological activity spectrum of the substance.” This spectrum is defined as the intrinsic property of a compound depending only on its molecular structure and physico-chemical characteristics.

PASS was trained on more than 30 000 compounds that reveal more than 500 kinds of different biological activities. The molecular descriptors used by PASS are $\rightarrow$ *MNA descriptors*.

- **path** $\rightarrow$ graph
- **path-Cluj matrices** $\rightarrow$ Cluj matrices
- **path-cluster subgraph** $\rightarrow$ molecular graph

- **path connectivity** → weighted matrices ($\odot$ weighted distance matrices)
- **path count** $\equiv$ *molecular path count* → path counts

■ **path counts** ($\equiv$ *path numbers*)

Path counts are atomic and molecular descriptors obtained from a → *H-depleted molecular graph* G , based on the counting of graph → *paths*. Analogous to the → *atomic walk count*, the **atomic path count** (or **atomic path number**) ${}^mP_i$ is a → *local vertex invariant* encoding the atomic environment, defined as the number of paths of length m starting from the i th vertex to any other vertex in the graph. The length m of the path, that is, the number of edges along the path, is called **path order** [Randić, Brissey *et al.*, 1979; Randić and Wilkins, 1979b; Randić, 1979].

The **vertex path code** (or **Randić atomic path code**) of the i th vertex is the ordered sequence of atomic path counts, with respect to the path length:

$$\{{}^1P_i, {}^2P_i, \dots, {}^LP_i\}$$

where $L = {}^A\eta_i$ is the → *atom detour eccentricity* of the i th vertex, that is, the length of the longest path starting from the vertex v_i ; it can be derived from the → *detour matrix* as the maximum value entry in the i th row. The atomic path count of first order ${}^1P_i$ is the → *vertex degree* δ_i , while the atomic path count of zero order ${}^0P_i$ is always equal to 1. Vertex path codes for all nonhydrogen atoms in the molecule can be collected into a rectangular matrix that has been called → *path-sequence matrix* **SP**. The sum of all the elements in the vertex path code is the total number of paths of any length starting from the considered vertex and is called **atomic path count sum** P_i :

$$P_i = \sum_{m=1}^L {}^mP_i$$

The **molecular path count**, also called **path count**, **molecular path number** or **topological bond index** with the symbol K_m , is the total number of paths of length m in the graph and is denoted by ${}^mP$ ($m = 0, 1, \dots, L$), where L is the length of the longest path in the graph. ${}^0P$ coincides with the number A of graph vertices, ${}^1P$ with the number B of graph edges, ${}^2P$ with the → *connection number* N_2 , that is, the number of two contiguous edges.

The molecular path count of order m is calculated by adding the corresponding atomic path counts of all vertices, then dividing by 2 since each path has been counted twice:

$${}^mP = \frac{1}{2} \cdot \sum_{i=1}^A {}^mP_i \quad m \neq 0$$

The path count ${}^0P$ is simply equal to A .

The **molecular path code** of the graph is the ordered sequence of molecular path counts:

$$\{{}^0P, {}^1P, {}^2P, \dots, {}^LP\}$$

Molecular path codes are → *vectorial descriptors*, used, for example, to search for similarities among molecules, by choosing a suitable value for the maximum length L with respect to the set of studied molecules to obtain → *uniform-length descriptors*.

It is noteworthy that, for acyclic graphs, the molecular path code coincides with the → *graph distance code*.

Summing up all the elements of the molecular path code gives the **total path count** P (also called **total path number**):

$$P = \sum_{m=0}^L {}^m P = A + \frac{1}{2} \cdot \sum_{m=1}^L \sum_{i=1}^A {}^m P_i = A + \frac{1}{2} \cdot \sum_{i=1}^A P_i$$

This descriptor is considered a quantitative measure of $\rightarrow$ *molecular complexity*.

For acyclic graphs, the total path count is simply calculated from the number A of graph vertices as

$$P = \frac{A^2 + A}{2}$$

For simple structures, the path counts can be derived directly from the molecular graphs; otherwise specific algorithms are needed. For example, Randić's algorithm results [Randić, Brissey *et al.*, 1979] in path counts for nonequivalent vertices from the $\rightarrow$ *adjacency matrix*.

Table P1 Outline of a generic path-sequence matrix with row and column sums.

Atom ID	Path length, m					Atomic path count sums
	0	1	2	...	L	
1	1	${}^1P_1$	${}^2P_1$	...	${}^LP_1$	P_1
2	1	${}^1P_2$	${}^2P_2$	...	${}^LP_2$	P_2
...	...	...	...	...	...	...
...	...	...	...	...	...	...
A	1	${}^1P_A$	${}^2P_A$	...	${}^LP_A$	P_A
Molecular path counts	A	${}^1P$	${}^2P$	...	${}^LP$	P

The $\rightarrow P$ matrix is a graph representation of molecules based on the total path count.

Five **path count-based indices** were proposed by Balaban, defined as [Balaban, Beteringhe *et al.*, 2007]

$$Q = \sum_{m=1}^L \frac{{}^m P^2}{(C+1)} \quad S = \sum_{m=1}^L \frac{{}^m P^{1/2}}{(C+1)} \quad D = \sum_{m=1}^L \frac{{}^m P^{1/2}}{m \cdot (C+1)}$$

$$A = \sum_{m=1}^L \frac{{}^m P}{m \cdot (C+1)} \quad P = \sum_{m=1}^L \frac{{}^m P^{1/2}}{m^{1/2} \cdot (C+1)}$$

where C is the $\rightarrow$ *cyclomatic number* and the summations run over the increasing path lengths.

The index Q increases with the molecule size and branching, whereas index S increases with size but decreases with branching; index D increases with cyclicity and decreases with branching. For acyclic graphs, index A is the $\rightarrow$ *Harary index* (denoted as H by Trinajstić and $RDSUM$ by Balaban). Finally, index P increases with size and decreases with branching; for hydrocarbons, this index shows the minimum number of degeneracies with respect to the other path count-based indices.

Atom-type path counts ${}^m P_X$ are defined as the number of paths originating from all the atoms of a given type. For example, the number of paths of length 3 originating from oxygen atoms in a molecule was used to predict boiling points of alcohols [Randić and Basak, 2001a].

To take into account multiple bonds and heteroatoms, **weighted path counts** can be calculated, either by introducing the weighting factors after the paths have been enumerated or by computing the weighted paths directly [Randić and Basak, 1999]. The sums of path weights obtained by applying different $\rightarrow$ *weighting schemes* to the graph edges are known as the $\rightarrow$ *ID numbers*; the most common weighting schemes are based on $\rightarrow$ *bond order indices*. Moreover, the **WTPT index** was proposed as the sum of the weights of all the paths starting from heteroatoms in the molecule [Bakken and Jurs, 1999b]; it is closely related to path counts of heteroatoms used in $\rightarrow$ *start-end vectors*.

$\rightarrow$ *Variable path counts* are obtained by weighting the graph edges involving heteroatoms with one or more variable parameters [Amić, Basak *et al.*, 2002].

Valence shell counts or **graph valence shells**, denoted by ${}^m S_i$, are weighted path counts calculated by adding valence shells at the same separation m for all atoms in a molecule [Randić, 2001b]. The concept of valence shell is similar to the concept of atomic path count; the difference is that instead of counting for each atom the number of neighbors at increasing length, one adds the $\rightarrow$ *vertex degree* of neighbors at increasing separation. The **valence shell** of order m for the i th vertex is then defined as

$${}^m S_i = \sum_{j=1}^A \sum_{p_{ij}} \delta_j \cdot \delta(|p_{ij}|; m)$$

where δ_j is the vertex degree of the j th atom, $|p_{ij}|$ is the length of a path connecting vertices v_i and v_j , and $\delta(|p_{ij}|; m)$ the Dirac delta function equal to 1 when the length of the path p_{ij} is equal to m , and zero otherwise. The first summation goes over all vertices in the graph, while the second one over all the paths connecting two vertices v_i and v_j . A shell of order zero represents the vertex degree of a vertex, while a shell of order one represents the $\rightarrow$ *extended connectivity* of the vertex. The valence shell of a vertex can be viewed as the count of weighted paths starting from the vertex, where the weights are determined by the vertex degree of the other terminal vertex of the path. For acyclic graphs, the valence shells for the i th vertex reduce to the elements lb_{im} in the i th row of the $\rightarrow$ *branching layer matrix* and are defined as the following [Lukovits, 2001a]:

$${}^m S_i \equiv lb_{im} = \sum_{j=1}^A \delta_j \cdot \delta(d_{ij}; m)$$

where d_{ij} is the topological distance between vertices v_i and v_j .

Then, the molecular valence shell count of m th order is calculated as

$${}^m S = \frac{1}{2} \cdot \sum_{i=1}^A {}^m S_i \quad m \neq 0$$

Based on the length of the paths in the molecular graph, other local vertex invariants and molecular descriptors have been proposed.

The **path degree** or **vertex path sum**, is a local invariant, denoted by ξ_i and defined as the sum of the lengths m of all paths starting from vertex v_i :

$$\xi_i = \sum_{m=1}^L {}^m P_i \cdot m$$

where $L = {}^A\eta_i$ is the atom detour eccentricity of the i th vertex, that is, the length of the longest path starting from v_i and ${}^mP_i$ is the number of paths of length m from v_i . For acyclic graphs, the path degree ξ_i coincides with the $\rightarrow$ vertex distance degree σ_i . Moreover, the path degrees are used as the weighting scheme for vertices to generate the $\rightarrow$ path degree layer matrix **LPD**.

By summing up path degrees over all vertices in the graph, the **all-path Wiener index** W^{AP} (or **Pasaréti index**) is derived. This is a molecular descriptor proposed as a variant of the $\rightarrow$ Wiener index but with more discriminating power among cycle-containing structures, defined as [Lukovits, 1998a; Lukovits and Linert, 1998]

$$W^{AP} = \frac{1}{2} \cdot \sum_{i=1}^A \xi_i = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \sum_{p_{ij}} |p_{ij}|$$

where the two outer summations on the right side run over all pairs of vertices in the graph and the inner summation runs over all paths p_{ij} between the vertices v_i and v_j ; $|p_{ij}|$ denotes the length of the considered path. Its maximum value is equal to $A^2 \times (A-1) \times 2^{(A-4)}$ for a $\rightarrow$ complete graph with A vertices.

It has to be noted that the all-path Wiener index coincides with a previously proposed global index obtained as the half-sum of any row of the path degree layer matrix **LPD**.

The all-path Wiener index can be calculated more easily from the **all-path matrix** Ω^{AP} that is a square symmetric $A \times A$ matrix, A being the number of graph vertices, whose i - j entry is the sum of the lengths of all the paths p_{ij} connecting vertices v_i and v_j :

$$[\Omega^{AP}]_{ij} = \begin{cases} \sum |p_{ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $|p_{ij}|$ denotes the length of a path between v_i and v_j . Diagonal elements are equal to zero by definition. $\rightarrow$ Distance matrix **D**, $\rightarrow$ detour matrix **Δ** and $\rightarrow$ detour distance – topological distance combined matrix $\mathbf{\Delta} \wedge \mathbf{D}$ are closely related to the all-path matrix as they are based on the length of the shortest, longest, and longest plus shortest paths between any two vertices in the graph, respectively. It must be noted that for acyclic graphs all these matrices coincide, there being a unique path between two vertices.

The row sums of the all-path matrix are the path degrees ξ_i :

$$\xi_i \equiv VS_i(\Omega^{AP}) = \sum_{j=1}^A [\Omega^{AP}]_{ij}$$

where VS_i indicates the $\rightarrow$ vertex sum operator.

The all-path Wiener index is then derived from the all-path matrix as the following:

$$W^{AP} \equiv Wi(\Omega^{AP}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Omega^{AP}]_{ij}$$

where Wi is the $\rightarrow$ Wiener operator.

Because the all-path Wiener index increases exponentially with the number A of graph vertices, it was proposed [Lukovits, 1998a] to divide it by the average number k of paths between vertices and the resulting quantity was called **Vérhalom index**:

$$\bar{W}^{AP} = W^{AP}/k$$

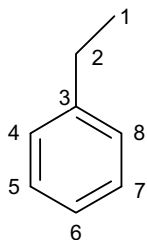
where k is obtained by the ratio of the total number of paths (of order greater than zero) over the number of vertex pairs $A \times (A - 1)/2$, where A is the number of vertices in the graph:

$$k = \frac{2 \cdot P}{A \cdot (A-1)}$$

For simple cycles, $k = 2$.

Example P1

Path-sequence matrix **SP**, all-path matrix Ω^{AP} , path degrees ξ_i , and related indices for ethylbenzene.



SP =

Atom	0	1	2	3	4	5	6	7	P_i
1	1	1	1	2	2	2	2	2	12
2	1	2	2	2	2	2	2	0	12
3	1	3	3	2	2	2	0	0	12
4	1	2	3	3	2	2	1	1	14
5	1	2	2	3	3	3	1	0	14
6	1	2	2	2	4	4	0	0	14
7	1	2	2	3	3	3	1	0	14
8	1	2	3	3	2	2	1	1	14
mP	8	8	9	10	10	10	4	2	106

$$P = \sum_{m=0}^7 mP = 8 + \frac{1}{2} \cdot \sum_{i=1}^8 P_i = 53$$

$$Q = \sum_{m=1}^7 \frac{mP^2}{(1+1)} = 232.5$$

$$S = \sum_{m=1}^7 \frac{mP^{1/2}}{(1+1)} = 9.365$$

$$D = \sum_{m=1}^7 \frac{mP^{1/2}}{m \cdot (1+1)} = 3.670$$

$$A = \sum_{m=1}^7 \frac{mP}{m \cdot (1+1)} = 10.643$$

$$P = \sum_{m=1}^7 \frac{mP^{1/2}}{m^{1/2} \cdot (1+1)} = 5.561$$

$$W^{AP} = \frac{1}{2} \cdot \sum_{i=1}^8 \xi_i = 184$$

$$W^{AP} = \frac{1}{2} \cdot \sum_{i=1}^8 \sum_{j=1}^8 [\Omega^{AP}]_{ij} = 184$$

$$\bar{W}^{AP} = \frac{W^{AP}}{k} = \frac{184}{53/28} = 97.2$$

$$\xi_1 = 1 \times 0 + 1 \times 1 + 1 \times 2 + 2 \times 3 + 2 \times 4 + 2 \times 5 + 2 \times 6 + 2 \times 7 = 53$$

$$\xi_2 = 1 \times 0 + 2 \times 1 + 2 \times 2 + 2 \times 3 + 2 \times 4 + 2 \times 5 + 2 \times 6 + 0 \times 7 = 42$$

$$\xi_3 = 1 \times 0 + 3 \times 1 + 3 \times 2 + 2 \times 3 + 2 \times 4 + 2 \times 5 + 0 \times 6 + 0 \times 7 = 33$$

.....

$$\xi_8 = 1 \times 0 + 2 \times 1 + 3 \times 2 + 3 \times 3 + 2 \times 4 + 2 \times 5 + 1 \times 6 + 1 \times 7 = 48$$

Ω^{AP} =

Atom	1	2	3	4	5	6	7	8	ξ_i
1	0	1	2	10	10	10	10	10	53
2	1	0	1	8	8	8	8	8	42
3	2	1	0	6	6	6	6	6	33
4	10	8	6	0	6	6	6	6	48
5	10	8	6	6	0	6	6	6	48
6	10	8	6	6	6	0	6	6	48
7	10	8	6	6	6	6	0	6	48
8	10	8	6	6	6	6	6	0	48

Table P2 Molecular path counts for C8 data set (Appendix C – Set 1).

C8	¹ P	² P	³ P	⁴ P	⁵ P	⁶ P	⁷ P	C8	¹ P	² P	³ P	⁴ P	⁵ P	⁶ P	⁷ P
n-Octane	7	6	5	4	3	2	1	33MM	7	9	7	4	1	0	0
2M	7	7	5	4	3	2	0	34MM	7	8	8	4	1	0	0
3M	7	7	6	4	3	1	0	2M3E	7	8	8	5	0	0	0
4M	7	7	6	5	2	1	0	3M3E	7	9	9	3	0	0	0
3E	7	7	7	5	2	0	0	223MMM	7	10	8	3	0	0	0
22MM	7	9	5	4	3	0	0	224MMM	7	10	5	6	0	0	0
23MM	7	8	7	4	2	0	0	233MMM	7	10	9	2	0	0	0
24MM	7	8	6	5	2	0	0	234MMM	7	9	8	4	0	0	0
25MM	7	8	5	4	4	0	0	223MMMM	7	12	9	0	0	0	0

📖 [Randić and Wilkins, 1979a, 1979c; Randić, 1980a, 1990a, 1991c, 1992c, 1996b, 1997a; Randić, Brissey *et al.*, 1980; Quintas and Slater, 1981; Wilkins, Randić *et al.*, 1981; Randić, Kraus *et al.*, 1983; Randić, 1984a; Randić, Jerman-Blazic *et al.*, 1987; Kunz, 1989; Randić and Jurs, 1989; Clerc and Terkovich, 1990; Hall, Kier *et al.*, 1993; Hall, Dailey *et al.*, 1993; Kier, Hall *et al.*, 1993; Pisanski and Žerovnik, 1994; Plavšić, Šoškić *et al.*, 1996b; Amić, Lučić *et al.*, 2001; Lukovits, Nikolić *et al.*, 2002]

- **path count-based indices** → path counts
- **path degree** → path counts
- **path degree layer matrix** → layer matrices
- **path-distance map matrix** → biodescriptors (⊙ proteomics maps)
- **path-distance-sum-connectivity matrix** → weighted matrices (⊙ weighted distance matrices)
- **path eccentricity** ≡ *atom detour eccentricity* → detour matrix
- **PathFinder fingerprints** → shape descriptors
- **path graph** ≡ *linear graph* → graph
- **path graphical bond order** → bond order indices (⊙ graphical bond order)
- **path-layer matrix** ≡ *path-sequence matrix* → sequence matrices
- **path length** → graph
- **path matrix** → double invariants
- **path matrix** ≡ *P-matrix* → bond order indices (⊙ graphical bond order)
- **path numbers** ≡ *path counts*
- **path order** → path counts
- **path-sequence matrix** → sequence matrices
- **path subgraph** → molecular graph
- **path-Szeged matrices** → Szeged matrices
- **path/walk shape indices** → shape descriptors
- **path-Wiener matrix** → Wiener matrix
- **path-χ matrix** → weighted matrices (⊙ weighted distance matrices)
- **pendent matrix** → superpendent index
- **Pauling bond number** → delocalization degree indices (⊙ Krygowski bond energy)
- **PCA** ≡ *Principal Component Analysis*

- **PC-based drug-like index** → scoring functions
- **PDQ descriptors** $\equiv$ *Pharmacophore-Derived Query descriptors* → substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **PDR-FP fingerprints** → cell-based methods
- **PDT fingerprints** → substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **Pearson's correlation coefficient** → statistical indices ($\odot$ correlation measures)
- **Pearson's first index** → statistical indices ($\odot$ moment statistical functions)
- **Pearson coefficient** → classification parameters
- **Pearson coefficient** → similarity/diversity
- **PEI** $\equiv$ *Polarizability Effect Index* → electric polarization descriptors
- **PEOE** $\equiv$ *Partial Equalization of Orbital Electronegativities* → electronegativity
- **per(D) index** → algebraic operators ($\odot$ determinant)

■ periphery codes

These are binary molecular codes proposed to characterize the periphery shape of molecules embedded on a 2D hexagonal lattice [Balaban and Harary, 1968; Balaban, 1976b]. They are suitable for the shape characterization of planar benzenoids and annulenes. "Inside" and "Outside" regions of closed curves are indicated by binary labels 1 and 0, respectively, associated with the graph vertices [Randić and Razinger, 1995a, 1995b, 1997]. In other words, digit 1 is associated with movement toward Inside and digit 0 with movement Outside of each ring; a clockwise direction is adopted and the starting point on the periphery is the vertex satisfying the convention of lexicographic minimum. Other different canonical rules can be chosen to define periphery codes [Jerman-Blazic Dzonova and Trinajstić, 1982; Müller, Szymanski *et al.*, 1990a].

Periphery codes can be used to evaluate → *similarity/diversity* based on molecular shape among several compounds [Randić and Razinger, 1995b]. Moreover, periphery codes can also be used to distinguish between *cis*- and *trans*-isomers [Oth and Gilles, 1968; Balaban, 1969, 1997a] and recognize whether a atom molecule is chiral or not [Randić, 1998a]. In particular, for 2D-embedded molecules, the → *Randić chirality index* was proposed by calculating a particular periphery code from left to right and from right to left: if different results are obtained, then the molecule is chiral.

📖 [Balaban, 1971, 1988a; Randić and Mezey, 1996]

- **permanent** → algebraic operators ($\odot$ determinant)
- **permittivity** $\equiv$ *dielectric constant* → physico-chemical properties
- **persistence** → environmental indices
- **persistence length** → size descriptors
- **perturbation connectivity indices** → connectivity indices
- **perturbation delta value** → vertex degree
- **perturbation geodesic matrices** → weighted matrices ($\odot$ weighted distance matrices)
- **perturbation graph matrices** → weighted matrices ($\odot$ weighted distance matrices)
- **Perturbation of an Environment Limited Concentric and Ordered** → DARC/PELCO analysis
- **PEST Autocorrelation Descriptors** → TAE descriptor methodology
- **PEST descriptors** → TAE descriptor methodology

- **Petitjean shape indices** → shape descriptors
- **pfaffian** → algebraic operators (⊙ determinant)
- **pH** → physico-chemical properties
- **pharmacological indices** → biological activity indices
- **pharmacophore** → drug design
- **pharmacophore-based descriptors** → substructure descriptors
- **Pharmacophore Definition Triplets fingerprints** $\equiv$ *PDT fingerprints* → substructure descriptors (⊙ pharmacophore-based descriptors)
- **Pharmacophore-Derived Query descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **Pharmacophore Point Filter** → scoring functions
- **pharmacophore signature** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **PharmPrint descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **phase capacity ratio** $\equiv$ *capacity factor* → chromatographic descriptors

■ physico-chemical properties

They constitute the most important class of experimental measurements and play a fundamental role as → *molecular descriptors* both for their availability as well as for their interpretability [Exner, 1966; Lyman, Reehl *et al.*, 1982; Reid, Prausnitz *et al.*, 1988; Horvath, 1992; Abraham, 1993c; Baum, 1997; Lide, 1999; Reinhard and Drefahl, 1999]. Physico-chemical properties are used both as the molecular properties to be correlated with molecular structure in QSPR modeling and as the molecular descriptors when searching for relationships with biological activities. Physico-chemical properties are constitutive parts of → *volume descriptors*, → *electric polarization descriptors*, → *spectra descriptors*, → *chromatographic descriptors*, and so on. Combinations of physico-chemical properties are largely used in the definition of → *environmental indices*. Other important physico-chemical properties are the so-called → *technological properties* useful to characterize materials.

Definitions of some important physico-chemical properties are given below.

• boiling point (BP)

Boiling point is the temperature at which the liquid and gas phases of a pure substance are in equilibrium at a specified pressure, that is, the temperature at which the substance changes its state from a liquid to a gas at a given pressure. The **normal boiling point** is the boiling point at normal atmospheric pressure (101.325 kPa). The SI units are Kelvin degrees K, nevertheless the Celsius degrees °C are still very much in use ($^{\circ}\text{C} = \text{K} - 273.15$).

In terms of intermolecular interactions, the boiling point represents the temperature at which molecules possess enough thermal energy to overcome the various intermolecular attractions binding the molecules into the liquid (e.g. hydrogen bonds, dipole–dipole attraction, instantaneous-dipole induced-dipole attractions). Therefore the boiling point is also an index of the strength of intermolecular attractive forces.

The boiling point of a pure compound increases with the increase in the molecule size and molecular branching, with the presence of hydrogen-bonds and dipole–dipole interactions.

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **critical constants**

The critical pressure P_c , critical volume V_c , and critical temperature T_c are the values of the pressure P , volume V_m , and thermodynamic temperature T at which the densities of coexisting liquid and gaseous phases become identical.

The **critical temperature**, T_c , of a substance is the temperature above which distinct liquid and gas phases do not exist, that is, the temperature above which a gas cannot be liquefied by an increase of pressure. As the critical temperature is approached, the properties of the gas and liquid phases become the same resulting in only one phase: the supercritical fluid.

The **critical pressure**, P_c , is the vapor pressure at the critical temperature and critical volume.

The **critical volume**, V_c , is the volume of a fixed mass of a fluid at critical temperature and pressure.

📖 [Needham, Wei *et al.*, 1988; Grigoros, 1990; Katritzky, Mu *et al.*, 1998; Turner, Costello *et al.*, 1998; Espinosa, Yaffe *et al.*, 2001; Wakeham, Cholakov *et al.*, 2002; Yao, Wang *et al.*, 2002]

- **density (ρ)**

The density of a substance is the mass m per unit volume V . For the common case of a homogeneous substance, it is expressed as

$$\rho = \frac{m}{V}$$

where m is the mass of the substance and V its volume. The SI units are kg m^{-3} .

In general, density can be changed by changing either the pressure or the temperature. Increasing the pressure will always increase the density of a material. Increasing the temperature generally decreases the density, but there are notable exceptions to this generalization (e.g., water).

- **dielectric constant (ϵ)**

The dielectric constant ϵ , also called **permittivity** and sometimes denoted by κ , is a measure of the ability of a substance to attenuate the transmission of an electrostatic force from one charged body to another [Karelson, 2001]. The lower the value, the greater the attenuation.

Based on the dielectric constant, the **Kirkwood function** is defined as [Kirkwood and Westheimer, 1938; Reichardt, 1990]

$$K_f = \frac{\epsilon - 1}{2 \cdot \epsilon + 1}$$

This function is used to study solvent effects and for classification of solvents. Moreover, the dielectric constant enters the definition of the $\rightarrow$ *molar refractivity*.

📖 [Schweitzer and Morris, 1999; Sulea and Purisima, 1999; Sild and Karelson, 2002]

- **dielectric susceptibility (χ^e)**

The dielectric susceptibility χ^e of a dielectric material is a measure of how easily it polarizes in response to an electric field. This, in turn, determines the electric permittivity of the material

and thus influences many other phenomena in that medium, from the capacitance of capacitors to the speed of light.

It is defined as the constant of proportionality relating the electric field $\mathbf{E}$ to the induced dielectric polarization density $\mathbf{P}$ such that

$$\mathbf{P} = \epsilon_0 \cdot \chi^e \cdot \mathbf{E}$$

where ϵ_0 is the electric permittivity in vacuum.

The electric displacement $\mathbf{D}$ is related to the polarization density $\mathbf{P}$ by

$$\mathbf{D} = \epsilon_0 \cdot \mathbf{E} + \mathbf{P} = \epsilon_0 \cdot (1 + \chi^e) \cdot \mathbf{E} = \epsilon_0 \cdot \epsilon \cdot \mathbf{E}$$

where ϵ is the $\rightarrow$ *dielectric constant* of the medium; the dielectric constant is related to the electric susceptibility as follows:

$$\epsilon = 1 + \chi^e$$

• enthalpies (H)

The **enthalpy** or *heat content* (denoted as H) is a thermodynamic quantity describing the thermodynamic potential of a system, which can be used to calculate the “useful” work obtainable from a closed thermodynamic system under constant pressure.

The **standard reaction enthalpy** (ΔH^0) is the variation of the enthalpy of a chemical reaction relatively to one mole of a specified reagent when both reagents and products are in their standard state (the most stable form of the element at 100 kPa of pressure and the specified temperature, usually 298 K or 25 °C).

Different enthalpies can be defined, depending on the involved thermodynamic process (Table P3) and their values are usually quoted in kJ/mol or kcal/mol or cal/g.

Table P3 Usual symbols for standard reaction enthalpies.

Symbol	Reaction	Symbol	Reaction
ΔH_f^0	Formation	ΔH_{fus}^0	Fusion
ΔH_c^0	Combustion	ΔH_{trans}^0	Transition phases
ΔH_{vap}^0	Vaporization	ΔH_{mix}^0	Mixing of fluids
ΔH_{sub}^0	Sublimation	ΔH_{ads}^0	Adsorption

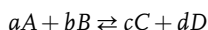
Negative values of standard reaction enthalpies indicate exothermic reactions, whereas positive values indicate endothermic reactions. Together with the $\rightarrow$ *molar volume*, vaporization enthalpy is used in determining the $\rightarrow$ *Hildebrand solubility parameter*.

📖 [Exner, 1973; Randić, 1991a; Li and You, 1993a; Pogliani, 1997b; Estrada, Torres *et al.*, 1998; Mercader, Castro *et al.*, 2000; Mercader, Castro *et al.*, 2001; Yao, Zhang *et al.*, 2001; Chickos, Nichols *et al.*, 2002; Puri, Chickos *et al.*, 2002a, 2002b, 2003; Toropov, Toropova *et al.*, 2004; Cao and Gao, 2005; Zhokova, Palyulin *et al.*, 2007]

- **equilibrium constants (K)**

The equilibrium constant is a dimensionless quantity characterizing a chemical equilibrium in a chemical reaction. It is a useful tool in determining the concentration of various reactants and products in a system where chemical equilibrium occurs.

For example, for the reaction in solution,



where A and B are reactant chemical species, C and D are product species, and a , b , c , and d are the stoichiometric coefficients of the respective reactants and products, the equilibrium constant is given by

$$K = \frac{[C]^c \cdot [D]^d}{[A]^a \cdot [B]^b}$$

where $[A]$, $[B]$, $[C]$, and $[D]$ are the concentrations of the species involved in the reaction. A more precise definition is in terms of activity rather than concentration.

Equilibrium constants are often represented by the quantity pK that is the negative logarithm (base 10) of an equilibrium constant K : $pK = -\log_{10}K$.

A **dissociation constant** is a constant whose numerical value depends on the equilibrium between the dissociated and undissociated forms of a molecule. Higher the dissociation constant, greater the dissociation. Examples of dissociation constants are *substrate-enzyme dissociation constant* and the **acid dissociation constant** pK_a . This latter is defined as

$$pK_a = pH + \log_{10} \left(\frac{AH}{A^-} \right)$$

where pH is the concentration of H^+ species, AH is the conjugated acid and A^- the conjugated base ($pK_a < 2$ means strong acid; $pK_a > 2$ and $pK_a < 7$ mean weak acid; $pK_a > 7$ and $pK_a < 10$ mean weak base; $pK_a > 10$ means strong base).

In chemistry and biochemistry, a dissociation constant is a specific type of equilibrium constant that measures the propensity of a larger object to separate (dissociate) reversibly into smaller components, as when a complex falls apart into its component molecules, or when a salt splits up into its component ions. The dissociation constant is often also denoted as K_d and is the inverse of the *affinity constant*. In the special case of salts, the dissociation constant can also be called *ionization constant*.

The dissociation constant is commonly used in QSAR studies to describe the affinity between a ligand (such as a drug) and a protein, that is, how tightly a ligand binds to a particular protein. Ligand-protein affinities are influenced by noncovalent intermolecular interactions between the two molecules such as hydrogen-bonding, electrostatic interactions, hydrophobic, and Van der Waals forces.

Fundamental thermodynamic equations relate the equilibrium constant to Gibbs (G) free energy, enthalpy (H), and entropy (S):

$$\Delta G^0 = \Delta H^0 - T \cdot \Delta S^0 = -RT \cdot \ln K$$

 Additional references are collected in the thematic bibliography (see Introduction).

- **flash point (FP)**

The flash point is the temperature at which the vapor above a volatile liquid forms a combustible mixture with air. At the flash point, the application of a naked flame gives a momentary flash rather than continuous combustion, for which the temperature is too low.

At this temperature, the vapor may cease to burn when the source of ignition is removed. However, as the temperature rises still further, the combustible substance reacts with oxygen in the air in an exothermic oxidation process.

Closely related to the flash point, the **autoignition temperature** is defined as the lowest temperature at which a substance in air will ignite in the absence of a spark or flame. Autoignition occurs when the rate of heat evolved is greater than the rate at which heat is lost to the surroundings.

Flash point and autoignition temperature are → *technological properties* of compounds and important safety parameters [Katritzky, Maran *et al.*, 2000], often used as one descriptive characteristic of liquid fuel, but also used to describe liquids that are not used intentionally as fuels.

QSPR studies on flash points and autoignition temperatures are [Egolf and Jurs, 1992; Murugan, Grendze *et al.*, 1994; Katritzky, Lobanov *et al.*, 1996; Tetteh, Metcalfe *et al.*, 1996; Mitchell and Jurs, 1997; Tetteh, Suzuki *et al.*, 1999; Katritzky, Petrukhin *et al.*, 2001a; Stefanis, Constantinou *et al.*, 2004].

A → *group contribution method* was also proposed for the calculation of the flash point of chemicals [Albahri and George, 2003].

- **fugacity**

Fugacity is the tendency of a substance to move from one environmental compartment to another, that is, to prefer one phase (liquid, solid, gas) over another. At a fixed temperature and pressure, a chemical will have a different fugacity for each phase: the phase with the lowest fugacity will be the most favorable.

Originally, the term was applied to the tendency of a gas to expand or escape and related to its pressure in the system being studied.

- **Henry's law constant (H)**

The Henry's law gives the relationship between the partial pressure P of a solute above the solution and its concentration c in the solution; it is defined as

$$e^P = e^{H \cdot c}$$

or, using the natural logarithm, as

$$P = H \cdot c$$

where H is the Henry's law constant; its units are $\text{L} \cdot \text{atm}/\text{mol}$, $\text{atm}/(\text{mol fraction})$, or $\text{Pa} \cdot \text{m}^3/\text{mol}$.

The Henry's law constant varies with the solvent and the temperature.

📖 [Nirmalakhandan and Speece, 1989b; Dunnivant, Elzerman *et al.*, 1992; ; Russell, Dixon *et al.*, 1992; Suzuki, Ohtaguchi *et al.*, 1992a; English and Carroll, 2001; Mariussen, Andersson *et al.*, 2001; Delgado and Alderete, 2002; Zhong, Yang *et al.*, 2002; Dearden and Schüürmann, 2003; Taskinen and Yliruusi, 2003; Wang, Tang *et al.*, 2003; Yaffe, Cohen *et al.*, 2003]

- **magnetic susceptibility (χ^m)**

It is the degree of magnetization of a material in response to an applied magnetic field. To distinguish magnetic susceptibility from $\rightarrow$ *dielectric susceptibility*, it is often denoted by χ^m and it relates the magnetization $\mathbf{M}$ of a material with the intensity of the applied magnetic field $\mathbf{H}$:

$$\mathbf{M} = \chi^m \cdot \mathbf{H}$$

The magnetic induction $\mathbf{B}$ is related to $\mathbf{H}$ by the relationship

$$\mathbf{B} = \mu_0 \cdot (\mathbf{H} + \mathbf{M}) = \mu_0 \cdot (1 + \chi^m) \cdot \mathbf{H} = \mu \cdot \mathbf{H}$$

where μ_0 is the magnetic permeability in the vacuum and μ the **magnetic permittivity** of the material.

If χ^m is positive, that is, $(1 + \chi) > 1$, the material is called paramagnetic and the magnetic field is strengthened by the presence of the material. Alternatively, if χ^m is negative, that is, $(1 + \chi) < 1$, the material is diamagnetic and the magnetic field is weakened by the presence of the material.

📖 [Dauben, Wilson *et al.*, 1968; Schmalz, Klein *et al.*, 1992; Estrada, 1998a]

- **melting point (MP)**

It is the temperature at which the solid and liquid states of a pure substance can exist in equilibrium; the melting point of a crystalline solid is the temperature at which it changes state from solid to liquid.

As heat is applied to a solid, its temperature increases until it reaches the melting point. At this temperature, additional heat converts the solid into a liquid without a change in temperature.

When considered as the temperature of the reverse change from liquid to solid, it is referred to as the **freezing point**. For most substances, melting and freezing points are equal.

Molecular size and symmetry usually increase the melting point; however, unlike the boiling point, the melting point is relatively insensitive to pressure. Melting points are often used to characterize organic compounds and to ascertain the purity. The melting point of a pure substance is always higher than the melting point of that substance when a small amount of an impurity is present. Moreover, together with the $\rightarrow$ *octanol–water partition coefficient*, melting point is used in the $\rightarrow$ *general solubility equation* to predict solubility of compounds.

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **molar refractivity (MR)**

The molar refractivity is the volume of the substance taken up by each mole of that substance. In SI units, MR is expressed as m^3/mol . MR is a molecular descriptor of a liquid, which contains both information about molecular volume and polarizability, usually defined by the Lorenz–Lorentz equation [Lorentz, 1880a, 1880b] (also known as the Clausius–Mosotti equation):

$$\text{MR} = \frac{n_D^2 - 1}{n_D^2 + 2} \cdot \frac{\text{MW}}{\rho} = \frac{\epsilon - 1}{\epsilon + 2} \cdot \bar{V}$$

where MW is the $\rightarrow$ *molecular weight*, ρ the liquid $\rightarrow$ *density*, and $\bar{V}$ the $\rightarrow$ *molar volume*, and n_D the $\rightarrow$ *refractive index* of the liquid referred to the sodium D line, and its square coincides with the $\rightarrow$ *dielectric constant* ϵ .

Molar refractivity is also proportional to $\rightarrow$ polarizability α , by the following [Hansch and Leo, 1995]:

$$MR = \frac{4}{3} \cdot \pi \cdot N_A \cdot \alpha$$

where N_A the Avogadro number (or Loschmidt constant), equal to $6.022\,141\,79 \times 10^{23} \text{ mol}^{-1}$, that is, the number of molecules in a mole of substance.

Molar refractivity can be used to design a set of bioactive molecules so that covariance between MR and hydrophobicity is minimized; MR can serve as a measure of binding force between the polar portions of an enzyme and its substrate.

Alternative definitions of molar refractivity were proposed by Gladstone and Dale (MR_{GD}) [Gladstone and Dale, 1858] and Vogel (MR_V) [Vogel, 1948] as

$$MR_{GD} = (n-1) \cdot \frac{MW}{\rho} \quad MR_V = n \cdot MW$$

where n is the refractive index.

Molar refractivity estimates by substituting the molar volume by $\rightarrow$ Mc Gowan's characteristic volume V_X were proposed by Abraham *et al.* [Abraham, Whiting *et al.*, 1990b] as

$$MR_A = 10 \cdot f(n) \cdot V_X$$

where $f(n)$ is the $\rightarrow$ refractive index function. Moreover, to remove cohesive dispersion interactions, it was proposed to subtract the molar refractivity of the n -alkane with the same characteristic volume V_X :

$$R_2 = MR_A - MR_A^* = MR_A - (2.83195 \cdot V_X - 0.52553)$$

where MR_A is the molar refractivity of the considered compound and MR_A^* the molar refractivity of the n -alkane with the same characteristic volume V_X . The parameter R_2 can be considered a polarizability descriptor and is called **excess molar refractivity**. By definition, $R_2 = 0$ for all n -alkanes, and the same holds for branched alkanes.

When molar refractivity is determined using the sodium D-line, it coincides with the $\rightarrow$ electron polarization. Therefore, it can be considered contemporarily as being an $\rightarrow$ electronic descriptor as well as a $\rightarrow$ steric descriptor of compounds.

As molar refractivity is essentially an additive property, **group molar refractivity** is calculated as the difference between the molar refractivity of an X-substituted compound and the reference compound:

$$MR_X = MR_{X+REF} - MR_{REF}$$

This parameter is often used as a substituent steric constant in $\rightarrow$ Hansch analysis. To put the molar refractivities of the substituents on approximately the same scale as the $\rightarrow$ hydrophobic substituent constants π , the substituent MR values are often scaled down by a factor 0.1.

The difference between the molar refractivity of a substituent MR_X and hydrogen MR_H was used to estimate the difference in the interaction energy of a hydrogen-substituted parent

compound and an X-substituted analogue compound:

$$\Delta E_{INT} = \frac{-1673.6}{r_{XB}^6} \cdot (MR_X - MR_H) \text{ kJ/mol}$$

where r_{XB} is the distance in angstroms between the group and the binding site [Pauling and Pressman, 1945].

Values for the atomic molar refractivity were also estimated by $\rightarrow$ *group contribution methods* [Ghose and Crippen, 1987].

The **molar refractivity partition index**, denoted as ${}^P MR_X$, is a $\rightarrow$ *Randić-like index* derived from the $\rightarrow$ *H-depleted molecular graph* of a compound as [Padrón, Carrasco *et al.*, 2002]

$${}^P MR_X = \sum_b \left[\gamma_i^{MR} \cdot \gamma_j^{MR} \right]_b^{-1/2} \quad i \neq j$$

where the summation goes over all the bonds and γ_i is the **atomic refractivity** of the i th atom plus the atomic refractivity of the hydrogens bonded to the i th atom; i and j indicate the two atoms forming the bond b .

Table P4 Molar refractivity values for different atom types.

Atom-type	Atomic refractivity	Atom type	Atomic refractivity
Csp ³	2.8128	N(Ar)	2.7662
Csp ²	3.8278	NO ₂	3.5054
Csp	3.8974	Ar–N=X	3.8095
C(Ar)	3.5090	F	1.0632
C=X	3.0887	Cl	5.6105
H	0.9155	Br	8.6782
–O–	1.6351	I	13.8741
=O	1.7956	Ssp ³	7.3190
O=N	2.1407	Ssp ²	9.1680
Nsp ³	3.0100	R–SO–R	6.0762
Nsp ² , Nsp	3.2009	R–SO ₂ –R	5.3321

 Additional references are collected in the thematic bibliography (see Introduction).

• molecular weight (MW)

Among the $\rightarrow$ *size descriptors*, molecular weight is the most simple and used molecular $\rightarrow$ *0D-descriptor*, calculated as the sum of the atomic weights of all the atoms in a molecule. It is related to molecular size and is atom-type sensitive. It is defined as

$$MW = \sum_{i=1}^A m_i$$

where m is the atomic mass and i runs over the A atoms of the molecule. The **average molecular weight** defined as

$$\overline{MW} = \frac{1}{A} \cdot \sum_{i=1}^A m_i = \frac{MW}{A}$$

is also used as molecular descriptor and is related to $\rightarrow$ atomic composition indices.

Square root molecular weight (MW2), defined as $MW2 = MW^{1/2}$, and **cubic root molecular weight (MW3)**, defined as $MW3 = MW^{1/3}$ and corresponding to a linear dimension of size, are also used as descriptors of molecule size.

- **parachor (PA)**

The parachor is defined by the Sudgen equation as [Sudgen, 1924]

$$PA = \gamma^{1/4} \cdot \frac{MW}{\rho_L - \rho_V} \approx \gamma^{1/4} \cdot \frac{MW}{\rho_L} = \gamma^{1/4} \cdot \bar{V}$$

where MW is the $\rightarrow$ molecular weight, γ the liquid $\rightarrow$ surface tension, and ρ_L and ρ_V the $\rightarrow$ density at a given temperature of liquid and vapor, respectively. The second relationship holds when the vapor density is negligible with respect to the liquid density, $\bar{V}$ being the $\rightarrow$ molar volume. This expression is considered to be an additive quantity, that is, can be approximately expressed as a sum of empirical increments PA_i corresponding to the single atoms or groups in the molecule. As an additive quantity, the parachor has been used in solving various structural problems.

The parachor is related to physico-chemical properties depending on the molecule volume, that is, $\rightarrow$ boiling point. It is essentially constant over wide ranges of temperature.

📖 [Vogel, 1948; Quayle, 1953; Ahmad, Fyfe *et al.*, 1975; Briggs, 1981; Zhao, Abraham *et al.*, 2003a; Tiwari and Pande, 2006]

- **partition coefficients**

A partition coefficient or distribution coefficient is a measure of the equilibrium between two different means, such as two different phases or two different immiscible liquids [Dearden, 1985]. It is usually denoted by K or P and defined as the ratio of the concentrations of a compound in a two-compartment system under equilibrium conditions:

$$K \equiv P = \frac{[C]_1}{[C]_2}$$

where $[C]_1$ and $[C]_2$ are the concentrations of the solute in the two systems. The partition coefficients are usually transformed in a logarithmic form as

$$\log P = \log \frac{[C]_1}{[C]_2} = \log[C]_1 - \log[C]_2$$

Partition coefficients are dimensionless measures of the relative affinity of a molecule with respect to the two phases and depend on absorption, transport, and partitioning phenomena.

In most of the cases, the two phases are an organic phase and an aqueous phase, that is, the partitioning of a compound between a lipidic and an aqueous phase.

The best known of these partition coefficients is the one based on the solvents 1-octanol and water. The **octanol–water partition coefficient** K_{ow} , very often expressed in its logarithmic form

$\log K_{ow}$ (also denoted as $\log P_{ow}$ or, often, simply as $\log P$) is a measure of the hydrophobicity and hydrophilicity of a substance measured as partition between 1-octanol (the lipidic phase) and water (the polar phase):

$$K_{ow} \equiv P = \frac{[C]_{1-octanol}}{[C]_{water}}$$

To avoid possible associations of the solute in the organic phase, partition coefficients should be measured at low concentrations or extrapolated to infinite dilution of the solute.

In the context of drug-like substances, hydrophobicity is related to absorption, bioavailability, hydrophobic drug–receptor interactions, metabolism and toxicity. Closely related to $\log P$ is the **octanol–water distribution coefficient** ($\log D_{pH}$), accounting for partition of pH-dependent mixture of ionizable species. Ionization of any compound makes it more water soluble and then less lipophilic. The $\log D$ can be calculated from $\log P$ and $\rightarrow$ acid dissociation constant pK_a by the following expression [Cronin, Aptula *et al.*, 2002b; Livingstone, 2003]:

$$\log D_{pH} = \log P - \log(1 + 10^{(pH - pK_a) \cdot I_{ab}})$$

where I_{ab} is equal to 1 for acids and to -1 for bases.

Due to its importance in QSAR studies, several approaches were proposed for modeling $\rightarrow$ lipophilicity of chemical compounds.

Other common partition coefficients are soil sorption partition coefficient, gas–solvent partition coefficient and micelle–water partition coefficient, together with $\rightarrow$ leaching indices, which are partition indices thought of for environmental studies.

Soil sorption partition coefficient (or **soil–water partition coefficient**), denoted as K_{oc} or $\log K_{oc}$, accounts for sorption from water into soil. Because this often depends primarily on the soil's organic carbon content, measured values are usually normalized for the organic carbon (OC) content of soil, in which case the soil sorption equilibrium constant is expressed as

$$K_{oc} = \frac{[C_{soil}]/[C_{soil}^0]}{[C_w]/[C_w^0]}$$

where $[C_{soil}]$ is the concentration of solute per gram of carbon in a standard soil and $[C_w]$ is the concentration of solute per volume of aqueous solution. The standard state concentrations $[C_{soil}^0]$ and $[C_w^0]$ are typically chosen as $1 \mu\text{g}$ of solute/g of organic carbon for soil and $1 \mu\text{g}$ of solute/ml for aqueous solution.

Several models for estimating soil sorption coefficients take advantage of the correlation between K_{oc} and other experimental partition coefficients, specially K_{ow} . For example, K_{oc} values have been estimated from experimental octanol–water partition coefficients by

$$\log K_{oc} = m \cdot \log K_{ow} + b$$

where m and b are slope and intercept, respectively, of the developed linear regression models. Published values of m and b range from 0.5 to 1.1 and from -0.2 to 1.3, respectively, depending on the range of data employed in the individual regression [Winget, Cramer *et al.*, 2000]. Moreover, the $\rightarrow$ adsorbability index was proposed as a K_{oc} descriptor.

Gas–solvent partition coefficient is known as the **Ostwald solubility coefficient** L and is usually written in the logarithmic form as [Katritzky, Mu *et al.*, 1996a; Katritzky,

Oliferenko *et al.*, 2003a]

$$\log L = \log \left(\frac{[C_l]}{[C_g]} \right)$$

where $[C_l]$ and $[C_g]$ are the concentrations of the substance in the liquid solvent and in the gas, respectively. It is used in the $\rightarrow$ *Linear Solvation Energy Relationships*.

Micelle–water partition coefficient, denoted as K_{mw} or in its logarithmic form as $\log K_{mw}$ or $\log P_{mw}$, is the partition of a solute between micellar and aqueous phases [Tanaka, Nakamura *et al.*, 1994; Abraham, Chadha *et al.*, 1995a].

Micellization is typical of surfactants that are organic molecules having a chemical structure combining both a polar (amphiphobic) and a nonpolar (amphiphilic) group into a single molecule. When dissolved in a solvent at low concentration, they have the ability to adsorb at interfaces, thereby alter significantly physical properties of the interfaces. In particular, micellization is observed in surfactant solutions when the concentration exceeds the *critical micelle concentration* (cmc), whereas the physico-chemical properties of the aqueous solution change abruptly [Li, Zhang *et al.*, 2004; Jalali-Heravi and Konouz, 2005].

Micelle–water partition coefficients are extracted by micelle chromatography (high performance liquid chromatography, HPLC) using micelle aqueous solution as mobile phase. For determination of K_{mw} , $\rightarrow$ *retention times* are measured using a usual HPLC system at various concentrations of micelle in the aqueous mobile phase and then estimated from the following equation:

$$K_{mw} = k' \cdot \phi = k' \cdot \frac{V_{mc}}{V_{aq}}$$

where k' is the $\rightarrow$ *retention factor* and ϕ is the phase ratio, defined as the volume of the micellar pseudostationary phase over that of the bulk aqueous phase (V_{mc}/V_{aq}), which is related to two intrinsic properties of the surfactant [Liu, Yao *et al.*, 2006; Katritzky, Pacureanu *et al.*, 2007].

As it was for the soil sorption partition coefficient, models for estimating micelle–water partition coefficients take advantage of the correlation with K_{ow} . For example, K_{mw} values have been estimated from experimental octanol–water partition coefficients by

$$\log K_{mw} = m \cdot \log K_{ow} + b$$

where m and b are slope and intercept, respectively, of the developed linear regression model [Ishihama, Oda *et al.*, 1996; Trone, Leonard *et al.*, 2000].

📖 [Tanaka and Fujiwara, 1996; Winget, Cramer *et al.*, 2000; Chen, Harner *et al.*, 2003; Fichert, Yazdani *et al.*, 2003; Basak, Mills *et al.*, 2004; Kahn, Fara *et al.*, 2005]

• pH

It is the common measure of the acid-base character of a solution, defined as

$$\text{pH} = -\log_{10}[\text{H}^+]$$

where $[\text{H}^+]$ is the concentration of hydrogen ions in moles per liter. The most precise definition is in terms of activity rather than concentration.

A solution of pH below 7 is acid, pH of 7 is neutral, pH over 7 is alkaline.

- **refractive index (n)**

The refractive index (or **index of refraction**) of a medium is defined as a ratio of the velocity of light in vacuum over the velocity of light in the substance of interest (a medium), or, in other words, is a measure for how much the speed of light (or other waves such as sound waves) is reduced inside the medium. For example, typical glass has a refractive index of 1.5, which means that light travels at $1/1.5 = 0.67$ times the speed in air or vacuum.

Used as an indicator of the purity of organic compounds, it is related to several electric and magnetic properties such as polarizability as well as to molar refractivity, critical temperature, surface tension, density, and boiling point. Usually, the refractive index is measured at the sodium D-line and indicated as n_D^2 . Moreover, the **refractive index function** $f(n)$ defined as

$$f(n) = \frac{n^2 - 1}{n^2 + 2}$$

was proposed as a molecular descriptor, accounting for composite solute interactions [Fuchs, Abraham *et al.*, 1982]. The refractive index of polymers is also among the important $\rightarrow$ *technological properties* of polymers.

📖 [Vogel, Cresswell *et al.*, 1951; Huggins, 1956; Katritzky, Sild *et al.*, 1998a, 1998b; Holder, Ye *et al.*, 2006a, 2006b; Cao and Gao, 2007]

- **solubility (S)**

Solubility is the maximum amount of solute that dissolves in a given quantity of solvent at a specific temperature, that is solubility refers to the ability for a given substance, the solute, to dissolve in a solvent. The resulting solution is called a saturated solution. Certain substances are soluble in all proportions with a given solvent, such as, for example, ethanol in water. This property is more correctly described as miscible.

Generally, for a solid in a liquid, solubility increases with temperature; for a gas, solubility decreases. Common measures of solubility include the mass of solute per unit mass of solution (mass fraction), mole fraction of solute, molality, molarity, and others.

Aqueous solubility is among the most important characteristics in ADME studies and plays a relevant role as physico-chemical descriptor in QSAR studies.

Solute-solvent interactions were largely studied and modeled by $\rightarrow$ *Linear Solvation Energy Relationships* and the $\rightarrow$ *Hildebrand solubility parameter*.

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **surface tension (γ)**

It is the attraction of molecules to each other on a liquid's surface, or, more specifically, the attractive intermolecular forces that liquid molecules below the surface exert on molecules at the surface. It is defined as

$$\gamma = [PA \cdot (\rho_L - \rho_V)]^4$$

where PA is the $\rightarrow$ *parachor*, and ρ_L and ρ_V are the liquid and vapor densities, respectively.

Surface tension creates a strong boundary between the air and liquid and is among the important → *technological properties* of substances.

📖 [Sudgen, 1924; Wiener, 1948a; Stanton and Jurs, 1992; Gutman, Popovic *et al.*, 1997; Kauffman and Jurs, 2001a; Knotts, Wilding *et al.*, 2001]

• **vapor pressure (V_P)**

The vapor pressure of a liquid is the pressure exerted by its vapor when the liquid and vapor are in dynamic equilibrium.

Vapor pressure is an indication of a liquid's evaporation rate. It relates to the tendency of molecules and atoms to escape from a liquid or a solid. A substance with a high vapor pressure at normal temperature is often referred to as volatile. The higher the vapor pressure of a material at a given temperature, the lower the boiling point.

The vapor pressure of any substance increases nonlinearly with temperature according to the Clausius–Clapeyron relation.

📖 [Wiener, 1948b; Pitzer, Lippmann *et al.*, 1955; Balaban and Feroiu, 1990; Basak, Gute *et al.*, 1997; Myrdal and Yalkowsky, 1997; Katritzky, Wang *et al.*, 1998; Liang and Gallagher, 1998; Goll and Jurs, 1999b; Simmons, 1999; Beck, Breindl *et al.*, 2000; Engelhardt McClelland and Jurs, 2000; Basak and Mills, 2001b; Chalk, Beck *et al.*, 2001; Olsen and Nielsen, 2001; Dearden, 2003b; Raevsky, Raevskaja *et al.*, 2007]

- **PI index** → Szeged matrices
- **Pisanski–Zerovnik index** → Wiener index
- **pK_a** ≡ *acid dissociation constant* → physico-chemical properties (⊙ equilibrium constants)
- **planted tree** → graph (⊙ tree)
- **Platt number** ≡ *total edge adjacency index* → edge adjacency matrix
- **PLS-based variable selection** → variable selection
- **P-matrix** → bond order indices (⊙ graphical bond order)
- **Pogliani cis/trans connectivity index** → *cis/trans* descriptors
- **Pogliani index** → Zagreb indices
- **point-by-point alignment** → alignment rules
- **polar effect** → electronic substituent constants
- **polar hydrogen factor** → electric polarization descriptors
- **polarity/polarizability descriptors** → electric polarization descriptors
- **polarity number** → distance matrix
- **polarizability** → electric polarization descriptors
- **polarizability effect index** → electric polarization descriptors
- **polarizability tensor** → electric polarization descriptors
- **polarizability volume** → electric polarization descriptors (⊙ mean polarizability)
- **polarization** → electric polarization descriptors
- **polar surface area** → molecular surface (⊙ solvent-accessible molecular surface)
- **Politzer hydrophobic model** → lipophilicity descriptors
- **polycenter** → center of a graph

■ polymer descriptors

Polymers are large molecules constituted of repeating structural units connected by covalent chemical bonds. Polymers are characterized by some specific → *physico-chemical properties*, → *technological properties* and conformational characteristics such as steric hindrance, characteristic ratio, persistence length, statistical chain segment (or Kuhn segment) length, molar stiffness function (also called molar limiting viscosity number function), intrinsic viscosity, and glass transition temperature [Katritzky, Maran *et al.*, 2000].

Molecular descriptors for polymers with an infinite number of repeating units are often calculated for small sequences (dimers, trimers) or for the single repeating unit.

Polymer descriptors ranges from → *quantum chemicals descriptors* [Holder, Ye *et al.*, 2006a; Yu, Yi *et al.*, 2007] to → *graph invariants* [Gutman, Kolaković *et al.*, 1989b, 1989a; Hosoya, 1991; Bonchev, Mekenyan *et al.*, 1992; Patil, Bora *et al.*, 1995; Gutman and Rosenfeld, 1996; Liu and Zhong, 2005], from the typical descriptors used in → *Linear Solvation Energy Relationships* [Kamlet, Abraham *et al.*, 1984; Taft, Abraham *et al.*, 1985; Kamlet, Doherty *et al.*, 1987a; Moody, Willauer *et al.*, 2005] to quantities computed by → *group contribution methods* [Elbro, Fredeslund *et al.*, 1991].

Polymer descriptors are the → *characteristic ratio*, and, among the → *size descriptors*, the → *Kuhn length*, the → *end-to-end distance*, the → *persistent length*.

Other examples of polymer descriptors are given below.

The **Wiener polymer index** is a normalized Wiener index for infinite polymers defined as [Balaban, Balaban *et al.*, 2001]

$$W_{\infty} = \frac{d}{3 \cdot (A_{pc} + C_{pc})}$$

where d is the shortest topological distance between two equivalent atoms in two neighboring polymer units, A_{pc} and C_{pc} are the number of atoms and cycles in the polymer unit.

The **mean overcrossing number** $\bar{N}$ is a descriptor for polymer chains accounting for the occurrence of entanglements caused by polymer chains interpenetrating each other (Figure P1). The mean overcrossing number is a → *geometrical descriptor* defined as the number of bond–bond crossings in a regular 2D projection of the chain, averaged over all possible projections and calculated on the → *molecular geometry* [Arteca, 1999]. It is a suitable descriptor of DNA chains, polymer geometrical shape, rheological and dynamic properties of polymer melts and concentrated solutions being explained by the occurrence of entanglements that cause geometrically constrained chain motion.

Moreover, the **average writhe** $\bar{W}_r$ was also defined as the observed overcrossing sum for each given 2D-projection, distinguishing right-handed crossing (+1) from left-handed crossing (−1). By definition, $\bar{N} \geq \bar{W}_r$. Both $\bar{N}$ and $\bar{W}_r$ provide useful information: in a compact random configuration a large value of $\bar{N}$ and a vanishing value of $\bar{W}_r$ are expected, whereas in a configuration with regular dihedral angles (e.g., a compact helix) both $\bar{N}$ and $\bar{W}_r$ are expected to be large.

These two descriptors can be combined to produce an effective polymer shape parameter, called **order parameter** ζ , such as

$$\zeta = \bar{N} - \frac{\bar{W}_r}{\bar{N}}$$

which exhibits two regular trends: $\zeta \rightarrow 0$ in a nonentangled regular configuration and $\zeta \rightarrow 1$ in an entangled random configuration.

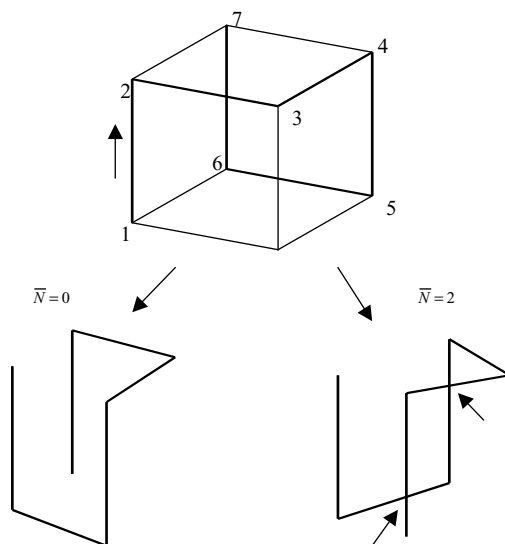


Figure P1 Example of calculation of the mean overcrossing number.

Another descriptor of the macromolecular topology is the **linking number** L that characterizes the entanglements of molecules having at least two molecular loops [White, 1969]. For two disjoint curves C_1 and C_2 , viewed along a direction in space, the linking number is computed as the sum of the handedness indices of only overcrossings for which curve C_1 is underneath C_2 , ignoring the overcrossings of each curve with itself. Two separate, nonentangled curves yield $L = 0$; the simplest nontrivial link of two loops, $L = 1$.

📖 [Small, 1953; Mekenyan, Dimitrov *et al.*, 1963; Volkenstein, 1963; Fuller, 1971; Bonchev and Mekenyan, 1980; Bonchev, Mekenyan *et al.*, 1981a, 1981b; Kamlet, Doherty *et al.*, 1986a, 1987c; Artemi and Balaban, 1987; Balaban and Artemi, 1987; Arteca and Mezey, 1990; Maranas, 1996; Balaban and Artemi, 1998; Katritzky, Sild *et al.*, 1998a; Katritzky, Sild *et al.*, 1998c; Sundaram and Venkatasubramanian, 1998; Camarda and Maranas, 1999; Zhong, Yang *et al.*, 2002; Arteca, 2003a, 2003b; Camacho-Zuñiga and Ruiz-Treviño, 2003; Edvinsson, Arteca *et al.*, 2003; Adams and Schubert, 2004; Afantitis, Melagraki *et al.*, 2005; Bonchev, Markel *et al.*, 2005; Funar-Timofei, Kurunczi *et al.*, 2005; Liu and Zhong, 2005; Shevade, Homer *et al.*, 2006; Yu, Wang *et al.*, 2006; Xu, Liu *et al.*, 2007]

- **polynomial** → algebraic operators
- **population analysis** → quantum-chemical descriptors
- **population trace** → DARC/PELCO analysis
- **positive predictive value** → classification parameters
- **potential of a charge distribution** → charge descriptors
- **Potential Pharmacophore Point pairs** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **power matrices** → matrices of molecules

- **power matrix** → algebraic operators ($\odot$ product of matrices)
- **PPFS** $\equiv$ *Property and Pharmacophore Features Score* → scoring functions
- **P'/P index** → bond order indices ($\odot$ graphical bond order)
- **PPP eigenvalues** → spectral indices
- **PPP pairs** $\equiv$ *Potential Pharmacophore Point pairs* → substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **PPP-triangle descriptors** → substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **Pratt measure** → statistical indices ($\odot$ concentration indices)
- **precision** → classification parameters
- **prediction error sum of squares** → regression parameters
- **predictive residual sum of squares** → regression parameters
- **predictive square error** → regression parameters
- **predictor variables** $\equiv$ *independent variables* → data set
- **prime ID number** → ID numbers
- **principal axes of a molecule** → principal moments of inertia
- **principal components** → Principal Component Analysis

■ Principal Component Analysis (PCA)

A fundamental chemometric technique for → *exploratory data analysis*, transforming the p variables in the data matrix $\mathbf{X}$ ($n \times p$), where n is the number of objects, into linear combinations of the common factors $\mathbf{T}$ ($n \times M$), called **principal components** and denoted by $\mathbf{t}_m$:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{L}^T$$

where $\mathbf{T}$ is the **score matrix**, $\mathbf{L}$ ($p \times M$) the **loading matrix**, and M the number of significant principal components ($M \leq p$). The columns of the loading matrix represent the eigenvectors $\mathbf{l}_m$; the eigenvector coefficients ℓ_{jm} ($-1 \leq \ell_{jm} \leq +1$), called *loadings*, represent the importance of each original variable (the rows of the loading matrix) in the considered eigenvector [Jolliffe, 1986; Jackson, 1991; Basilevsky, 1994].

The principal components are calculated according to the maximum variance criterion, that is, each successive component is an orthogonal linear combination of the original variables such that it covers the maximum of the variance not accounted for by the previous components. The eigenvalue λ_m associated with each m th component represents the variance explained by that component. Moreover, the sum of the variances of all the components equals the variance of the original variables.

The principal components are as linear combinations of the p original variables:

$$\mathbf{T} = \mathbf{X} \cdot \mathbf{L}, \quad \text{that is,} \quad t_{im} = x_{i1} \cdot \ell_{1m} + x_{i2} \cdot \ell_{2m} + \cdots + x_{ip} \cdot \ell_{pm} = \sum_{j=1}^p x_{ij} \cdot \ell_{jm}$$

where ℓ_{jm} are the coefficients of the linear combinations (i.e., the *loadings*) and t_{im} is the PCA score of the i th object (e.g., molecule, amino acid, etc.) in the m th principal component.

Mathematically, PCA consists in the diagonalization of the → *correlation matrix* (or covariance matrix) of the data matrix $\mathbf{X}$ with size $p \times p$ (the number of variables).

The main advantages of principal components are that

- (1) each component is orthogonal to all the remaining components, that is, the information carried by this component is unique;
- (2) each component represents a *macrovariable* of the data;
- (3) components associated with the lowest eigenvalues do not usually contain useful information (noise, spurious information, etc.).

Once M significant components have been chosen, each i th object is represented by the M -dimensional score vector:

$$\{t_{i1}; t_{i2}; \dots; t_{iM}\}$$

often called **z-scores** (or **z-scales**) and denoted as $\{z_{i1}; z_{i2}; \dots; z_{iM}\}$ or **principal properties** PP and denoted as $\{PP_{i1}; \dots; PP_{iM}\}$.

Principal properties PPs (or z-scores) are $\rightarrow$ *vectorial descriptors* of compounds, which summarize the main information of the original molecular descriptors or are empirical scales describing the physico-chemical properties of the training set objects [Alunni, Clementi *et al.*, 1983; Carlson, 1992; Clementi, Cruciani *et al.*, 1993a]. The number of significant PPs and their meaning depend closely on the original variables used to perform PCA. Since the PPs derived from PCA are orthogonal to each other and their number is usually small, they are suitable for design problems [Skagerberg, Bonelli *et al.*, 1989; Eriksson, Johansson *et al.*, 1997].

Principal properties can be calculated for both whole molecules and substituent groups, fragments, amino acids, and so on. For example, the i th substituent can be represented by four PPs, each having a different meaning such as PP_1 = steric, PP_2 = lipophilic, PP_3 = electrostatic, PP_4 = H-bonding properties of the substituent, respectively.

$\rightarrow$ *BC(DEF) parameters* are principal properties of a data matrix given by six physico-chemical properties describing 114 diverse liquid-state compounds.

Principal properties calculated on molecular $\rightarrow$ *interaction energy values* obtained by $\rightarrow$ *grid-based QSAR techniques* are usually referred to as **3D principal properties (3D-PP)** [van de Waterbeemd, Clementi *et al.*, 1993]. They were originally proposed for a theoretical description of the amino acids [Norinder, 1991; Cocchi and Johansson, 1993]. **3D-PP** were also calculated from $\rightarrow$ *ACC transforms*. Several other principal properties were proposed as the $\rightarrow$ *amino acid descriptors*.

When different data sets of descriptors are used separately to derive the principal properties of the same compounds, **disjoint principal properties (DPP)** are obtained as the whole set of significant PPs derived from each block of descriptors:

$$\{PP_1^A, PP_2^A, \dots, PP_{M_A}^A; PP_1^B, PP_2^B, \dots, PP_{M_B}^B; PP_1^C, PP_2^C, \dots, PP_{M_C}^C\}$$

where A, B, C represent three different blocks of variables on which PCA was performed and M_A , M_B , and M_C the corresponding numbers of significant principal components [van de Waterbeemd, Costantino *et al.*, 1995].

📖 [Weiner and Weiner, 1973; Dunn III and Wold, 1978, 1980; Dunn III, Wold *et al.*, 1978; Wold, 1978; Lukovits and Lopata, 1980; Streich, Dove *et al.*, 1980; Lukovits, 1983; McCabe, 1984; Maria, Gal *et al.*, 1987; Eriksson, Jonsson *et al.*, 1988, 1989, 1990; van de Waterbeemd,

El Tayar *et al.*, 1989; Hemken and Lehmann, 1992; Ridings, Manallack *et al.*, 1992; Suzuki, Ohtaguchi *et al.*, 1992a; Tysklind, Lundgren *et al.*, 1992; Caruso, Musumarra *et al.*, 1993; Cristante, Selves *et al.*, 1993; Ordorica, Velazquez *et al.*, 1993; Rodríguez Delgado *et al.*, 1993; Rodríguez Delgado, Sánchez *et al.*, 1993; Bazylak, 1994; Franke, Gruska *et al.*, 1994; Norinder, 1994; Azzaoui and Morinallory, 1995; Bjorsvik and Priebe, 1995; Cocchi, Menziani *et al.*, 1995; Clementi, Cruciani *et al.*, 1996; Gibson, McGuire *et al.*, 1996; Kimura, Miyashita *et al.*, 1996; Bjorsvik, Hansen *et al.*, 1997; Young, Profeta *et al.*, 1997; Balasubramanian and Basak, 1998; Langer and Hoffmann, 1998a; Kuanar, Kuanar *et al.*, 1999a; Vendrame, Braga *et al.*, 1999; Xue, Godden *et al.*, 1999b]

- **principal component analysis feature selection** → variable reduction
- **principal inertia axes** $\equiv$ *principal axes of a molecule* → principal moments of inertia

■ **principal moments of inertia** (I_A , I_B , I_C) ($\equiv$ *inertia principal moments*)

They are physical quantities related to the rotational dynamics of a molecule. The **moment of inertia** about any axis is defined as

$$I = \sum_{i=1}^A m_i \cdot r_i^2$$

where A is the atom number, m_i and r_i are the atomic mass and the perpendicular distance from the chosen axis of the i th atom of the molecule, respectively. For any rectangular coordinate system, with axes X , Y , Z , three moments of inertia are defined as

$$I_{XX} = \sum_{i=1}^A m_i \cdot (y_i^2 + z_i^2) \quad I_{YY} = \sum_{i=1}^A m_i \cdot (x_i^2 + z_i^2) \quad I_{ZZ} = \sum_{i=1}^A m_i \cdot (x_i^2 + y_i^2)$$

where (x, y, z) are the coordinates of the atoms.

The corresponding cross-terms are called **products of inertia** and are defined as

$$I_{XY} = Y_{YX} = \sum_{i=1}^A m_i \cdot x_i \cdot y_i \quad I_{XZ} = Y_{ZX} = \sum_{i=1}^A m_i \cdot x_i \cdot z_i \quad I_{YZ} = Y_{ZY} = \sum_{i=1}^A m_i \cdot y_i \cdot z_i$$

Therefore the **inertia matrix**, denoted by $\mathbf{I}$, is a square symmetric matrix 3×3 , collecting the three moment of inertia and six products of inertia.

Principal moments of inertia are the moments of inertia corresponding to that particular and unique orientation of the axes for which one of the three moments has a maximum value, another a minimum value, and the third is either equal to one of the others or intermediate in value to the other two. The corresponding axes are called **principal axes of a molecule** (or **principal inertia axes**). Moreover, the products of inertia all reduce to zero and the corresponding inertia matrix is diagonal. Conventionally, principal moments of inertia are labeled as

$$I_A \leq I_B \leq I_C$$

In general, the three principal moments of inertia have different values, but, depending on the molecular symmetry, they show characteristic equalities such as those shown in Table P5.

A number of $\rightarrow$ *shape descriptors* is defined in terms of principal moments of inertia. Moreover, principal moments of inertia are used to provide a unique reference framework for the calculation of the $\rightarrow$ *shadow indices*, and, in general, are used to define $\rightarrow$ *alignment rules* of the molecules. They constitute the basic starting point for the calculation of $\rightarrow$ *WHIM descriptors* and $\rightarrow$ *CoMMA method*.

Table P5 Principal moments for some selected symmetries.

Symmetry	Principal moments	Example
Spherical top	$I_A = I_B = I_C$	CCl_4
Symmetric top	$I_A = I_B \neq I_C$	NH_3
Asymmetric top	$I_A \neq I_B \neq I_C$	CH_2FCl
Linear symmetry	$0 = I_A \neq I_B = I_C$	$\text{HC}\equiv\text{CH}$
Planar symmetry	$I_A + I_B = I_C$	C_6H_6

- **principal properties** $\rightarrow$ Principal Component Analysis
- **privileged pharmacophore keys** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)

■ Probabilistic Receptor Potential (PRP)

This is a 3D-QSAR method designed to predict, in a qualitative manner, the types of receptor atoms to which a compound would prefer to bind [Labute, 2001].

To this end, molecules with different binding activities are aligned and common hydrogen-bond and hydrophobic regions are determined. Then, the type of interactions that most likely occur at different regions around the compounds are evaluated.

- **probability matrices** $\equiv$ *stochastic matrices* $\rightarrow$ algebraic operators
- **probe** $\rightarrow$ grid-based QSAR techniques
- **products of inertia** $\rightarrow$ principal moments of inertia
- **product of matrices** $\rightarrow$ algebraic operators
- **product of row sums index** $\equiv$ *PRS index* $\rightarrow$ distance matrix
- **preference** $\rightarrow$ DARC/PELCO analysis
- **Property and Pharmacophore Features Score** $\rightarrow$ scoring functions
- **Property and Pharmacophore Features fingerprints** $\rightarrow$ scoring functions ($\odot$ Property and Pharmacophore Features Score)
- **Property-Encoded Surface Translator descriptors** $\equiv$ *PEST descriptors* $\rightarrow$ TAE descriptor methodology

■ property filters

A property filter is a set of general and objective rules based on limits on structural features and physico-chemical properties that are shared by drugs or lead compounds. These rules are extracted from large collections of chemicals, containing both generic chemicals and drugs. By comparing a collection of known drugs with a collection of nondrugs, distribution of structural features and properties of compounds are analyzed by different methods to identify those features and value ranges of properties qualifying a compound to be a drug.

To focus drug discovery toward effective and orally adsorbable compounds, properties considered are usually related to *Absorption, Distribution, Metabolism, Excretion* (→ *ADME properties*).

Property filters are largely used in screening of virtual libraries and design of combinatorial libraries, allowing selection from large chemical database of compounds with desired properties to be potential drugs or, alternatively, removal of existing compounds with undesired properties [Clark and Pickett, 2000; Oprea, 2003; Leach, Hann *et al.*, 2006]. When filters are used to extract good drug candidates, they are usually referred to as **drug-like indices**. When they are applied to identify those chemicals that is likely to fail the development process, the term **alert indices** is more appropriate. When filters are only based on limits on functional groups they are properly called **functional group filters** [Muegge, 2003; Walters and Murcko, 2002] or **chemical filters** [Oprea, Gottfries *et al.*, 2000].

Different authors are using the term “drug-like” with slightly different meaning. Muegge says that “drug-likeness is a general descriptor of the potential of a small molecule to become a drug. It is not a unified descriptor but a global characteristic of a compound possessing many specific characteristics such as good solubility, membrane permeability, half-life, and having a pharmacophore pattern to interact specifically with a target protein. In reality, highly potent compounds against a drug target may not be efficacious because of pharmacokinetic problems; they may be toxic or unfavorably interact with other drugs” [Muegge, 2003].

Lipinski defines drug-like “*those compounds that have sufficiently acceptable ADME properties and sufficiently acceptable toxicity properties to survive through the completion of human Phase I clinical trials*” [Lipinski, 2000]. Walters and Murcko define drug-like compounds as those “*molecules which contain functional groups and/or have physical properties consistent with the majority of known drugs*” [Walters and Murcko, 2002].

There is a large variety of molecular descriptors used to address drug-likeness: they range from constitutional and counting descriptors to topological descriptors, from physico-chemical properties to pharmacophore description, from thermodynamic considerations to the synthetic accessibility, from presence of functional groups to ADME/Tox properties.

Several chemoinformatic approaches were proposed to evaluate drug-likeness of compounds; these include simple counting rules, such as property filters, and more complex regression and classification models, obtained by machine learning algorithms based on → *artificial neural networks* and recursive partitioning. These models have been used to derive descriptor weights and → *scoring functions* that classify compounds as drug or nondrug.

Moreover, to improve the chances of finding a drug candidate, it has been suggested to select small rational libraries from large libraries, or, in other words, to select a set of compounds with properties representative of the large library [Ashton, Jaye *et al.*, 1996]. This set of representative compounds can be selected by means of clustering or cell-based methods. → *Cell-based methods* require one or more quantitative molecular properties accounting for ligand–receptor binding interactions and properties involved in the transport of the drug to its target. Unlike common clustering methods, cell-based methods are more suitable to identify missing diversity in a chemical library and to highlight underrepresented or unrepresented regions of the overall chemical space.

Property filters usually are binary variables assuming a value equal to 1, if the molecule shows a specific property (drug-likeness, ADME properties, and toxicities) and equal to zero otherwise. These filters are not comprised of many molecular descriptors and a threshold or a range of values is associated to each descriptor together with a condition on the descriptor value: if the

conditions are fulfilled for all the descriptors, the studied property is considered as potentially present in the molecule. Usually, a few violations are allowed.

In the following, some property filters are reported. They are divided into drug-like indices and lead-like indices depending on whether they address drug-likeness or lead-likeness of compounds. In the section functional group filters, a survey of the most common functional groups for database filtering is given. All the property filters that allow a drug-likeness ranking of compounds instead of a simple yes/no response are reported elsewhere under $\rightarrow$ *scoring functions*.

• drug-like indices

The **Lipinski drug-like index** (or **rule-of-five**, RO5) is the first drug-like filter proposed to predict oral bioavailability of compounds that have achieved phase II clinical status [Lipinski, Lombardo *et al.*, 1997, 2005]. This filter predicts that poor absorption or permeation is more likely when more than one violation is registered for the four following rules: molecular weight (MW) ≤ 500 , $\log P \leq 5$, number of hydrogen-bond acceptors (HBA) ≤ 10 ; number of hydrogen-bond donors (HBD) ≤ 5 .

The Lipinski rules were derived from an analysis of 2245 drugs from the WDI database; they identify compounds lying in a region of property space where the probability of useful oral activity is very high. A compound that fails the filter, that is, two or more properties are out of range, will likely be poorly bioavailable because of poor absorption or permeation.

As the rule-of-five was designed to predict compound bioavailability, it is not really able to distinguish between drugs and nondrugs [Frimurer, Bywater *et al.*, 2000; Oprea, 2000]. Moreover, there are some limitations of the rule-of-five [Keller, Pichota *et al.*, 2006]: (1) RO5 applies only to compounds that are delivered by the oral route (not applicable for substrates of transporter and natural products); (2) RO5 applies only to compounds that are absorbed by passive mechanisms [Lipinski, Lombardo *et al.*, 2001]; (3) important RO5 violations come from antibiotics, antifungals, vitamins, and cardiac glycosides [Walters and Murcko, 2002]; (4) compliant compounds are not necessarily good drugs; (5) RO5 says nothing about specific chemical structural features found in drugs or nondrugs [Lipinski, 2004].

Bhal *et al.* proposed a revised rule-of-five by using the logarithm of the $\rightarrow$ *octanol–water distribution coefficient* ($\log D_{\text{pH}}$), at pH 5.5 ($\log D_{5.5}$), instead of $\log P$ because $\log D$ is a better descriptor for lipophilicity accounting for the ionization of compounds under physiological conditions [Bhal, Kassam *et al.*, 2007]. The idea underpinning this replacement is that since ionization of molecules results in decreased lipophilicity with respect to the neutral state, it is necessary to take into account the ionic state of the compound when describing the lipophilicity of potential drugs.

To achieve a better distinguishing between drugs and nondrugs, other property filters are defined which are extensions of the rule-of-five. Some of them are collected in Table P6 and briefly commented in the text below.

The drug-like filter proposed by Chen *et al.* is applied after a first structural screening aimed at excluding compounds containing atoms different from C, H, O, N, S, P, F, Cl, Br, or I [Chen, Zheng *et al.*, 2005]. Moreover, the three descriptors added to those of the RO5 are $\rightarrow$ *combined descriptors* defined as ratio of $\rightarrow$ *count descriptors*: C3p is the ratio of the number of C(sp³) atoms over the total number of nonhalogen heavy atoms; h-p is the ratio of the number of hydrogen atoms over the total number of nonhalogen heavy atoms; Unsat-p is the ratio of molecular unsaturation, as defined in $\rightarrow$ *multiple bond descriptors*, over the number of nonhalogen heavy atoms. The same authors also proposed a simple filter based only on two molecular descriptors

Table P6 Lipinski rule-of-five (RO5) and related drug-like filters.

Descriptor	RO5	Oprea <i>et al.</i>	Chen <i>et al.</i>	Monge <i>et al.</i>	Walters <i>et al.</i>	Rishton
MW	≤500	[200; 450]	[78; 500]	[100; 800]	[200; 500]	≤500
log <i>P</i>	≤5	[−2; 4.5]	[−0.5; 5]	≤7	[−5; 5]	≤5
<i>HBA</i>	≤10	[1; 8]	[2; 10]	≤10	≤10	≤10
<i>HBD</i>	≤5	≤5	≤5	≤5	≤5	≤5
<i>RBN</i>		[1; 9]		≤15	≤8	≤10
<i>NRG</i>		≤5		≤6		
<i>PSA</i> (Å ²)						≤140
C3p			[0.15; 0.8]			
h-p			[0.6; 1.6]			
Unsat-p			[0.10; 0.45]			
Charge					[−2; +2]	
Halogens				≤7		
O + N				≥1		

MW, molecular weight; *HBA*, number of hydrogen-bond acceptors; *HBD*, number of hydrogen-bond donors; *RBN*, number of rotatable bonds; *NRG*, number of rings (cyclomatic number); *PSA*, partial surface area. Noncited descriptors are defined in the text. Data from [Oprea, Gottfries *et al.*, 2000; Oprea, 2000; Chen, Zheng *et al.*, 2005; Monge, Arrault *et al.*, 2006; Walters and Murcko, 2002; Rishton, 2003].

that are independent of the molecular size [Zheng, Luo *et al.*, 2005]: one is the unsaturation-related descriptor Unsat-p and the other is a descriptor related to the proportion of heteroatoms NO_C3, defined as the ratio of the total number of oxygen and nitrogen atoms over the number of carbon atoms with sp³ hybridization. The filter for drug-like compounds is then,

$$\text{Unsat-p} \leq 0.43 \quad \text{and} \quad 0.10 \leq \text{NO\_C3} \leq 1.8$$

The filter of Monge *et al.* includes some additional rules based on molecular structural features. In particular: (a) compounds with atoms other than C, H, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, or Li are not allowed to pass the filter; (b) no reactive functions; (c) no perfluorinated chains (e.g., −CF₂CF₂CF₃); (d) no rings with more than seven members; (e) alkyl chains ≤ −(CH₂)₆CH₃. This filter was derived from the analysis of 2.6 million compounds collected from 32 diverse chemical databases.

The property filter of Walters *et al.* is implemented in the program REOS where a set of more than 200 functional group filters is also available to enable one to remove compounds with toxic, reactive, and otherwise undesirable moieties.

The filter proposed by Rishton is based on data taken from the literature.

Another property filter designed to predict oral bioavailability was proposed by [Veber, Johnson *et al.*, 2002] by substituting the four Lipinski rules with the following two rules: (a) number of rotatable bonds ≤10, and (b) polar surface area (*PSA*) ≤ 140 Å² or the sum of H-bond acceptors and H-bond donors ≤ 12.

Eight **GVW drug-like indices** have been proposed by Ghose–Viswanadhan–Wendoloski [Ghose, Viswanadhan *et al.*, 1999] to help streamline the design of combinatorial chemistry libraries for drug design and develop guidelines for prioritizing large sets of compounds for biological testing. They are based on a consensus definition and have been derived from analysis

of the distribution of some physico-chemical properties ($\log P$, molar refractivity, molecular weight, number of atoms) and chemical constitutions of known drug molecules available in the Comprehensive Medicinal Chemistry (CMC) database and seven drug classes defined by disease state.

Among the eight proposed indices is a general drug-like index that has been derived from the analysis of the whole CMC database and seven specific drug-like indices derived from the property distributions within the single drug classes (Table P7).

$\log P$ ($\rightarrow$ ALOGP) and $\rightarrow$ molar refractivity (AMR) are calculated by using the atomic contribution method of Ghose, Crippen, and Viswanadhan. The drug-like indices are dummy variables taking value equal to 1 when all the criteria of the consensus definition of a drug-like molecule are satisfied, 0 otherwise. Specifically, a drug-like index equals 1 when $\log P$, molar refractivity, molecular weight (MW), and number of atoms (A) of a compound are in the property range reported in Table P7; moreover, the compound must be a combination of some of the following functional groups: a benzene ring, a heterocyclic ring (both aliphatic and aromatic), an aliphatic amine, a carboxamide group, an alcoholic hydroxyl group, a carboxy ester, and a keto group. For example, according to the CMC-80 index, an organic compound is a drug-like molecule if: the calculated ALOGP is between -0.4 and 5.6 , the molar refractivity AMR between 40 and 130 , the molecular weight MW between 160 and 480 , the total number of atoms A between 20 and 70 , and it includes at least one of the above mentioned functional groups.

Two property ranges have been proposed: the *qualifying range* that covers approximately 80% of the drugs studied and the *preferred range* that is the smallest range within the qualifying range occupied by approximately 50% of the drugs. If large compound databases are screened by means of the indices based on the qualifying range (80%), the chance of missing drug-like

Table P7 Value ranges of the descriptors used in defining GVW drug-like indices.

Drug class	P%	ALOGP	AMR	MW	A
CMC	80	[-0.4 ; 5.6]	[40 ; 130]	[160 ; 480]	[20 ; 70]
CMC	50	[1.3 ; 4.1]	[70 ; 110]	[230 ; 390]	[30 ; 55]
Antiinflammatory	80	[1.4 ; 4.5]	[59 ; 119]	[212 ; 447]	[24 ; 59]
Antiinflammatory	50	[2.6 ; 4.2]	[67 ; 97]	[260 ; 380]	[28 ; 40]
Antidepressant	80	[1.4 ; 4.9]	[62 ; 114]	[210 ; 380]	[32 ; 56]
Antidepressant	50	[2.1 ; 4.0]	[75 ; 95]	[260 ; 330]	[37 ; 48]
Antipsychotic	80	[2.3 ; 5.2]	[85 ; 131]	[274 ; 464]	[40 ; 63]
Antipsychotic	50	[3.3 ; 5.0]	[94 ; 120]	[322 ; 422]	[49 ; 61]
Antihypertensive	80	[-0.5 ; 4.5]	[54 ; 128]	[206 ; 506]	[28 ; 66]
Antihypertensive	50	[1.0 ; 3.4]	[68 ; 116]	[281 ; 433]	[36 ; 58]
Hypnotic	80	[0.5 ; 3.9]	[43 ; 97]	[162 ; 360]	[20 ; 45]
Hypnotic	50	[1.3 ; 3.5]	[43 ; 73]	[212 ; 306]	[29 ; 38]
Antineoplastic	80	[-1.5 ; 4.7]	[43 ; 128]	[180 ; 475]	[21 ; 63]
Antineoplastic	50	[0.0 ; 3.7]	[60 ; 107]	[258 ; 388]	[30 ; 55]
Antimicrobial	80	[-0.3 ; 5.1]	[44 ; 144]	[145 ; 455]	[12 ; 64]
Antimicrobial	50	[0.8 ; 3.8]	[68 ; 138]	[192 ; 392]	[12 ; 42]

ALOGP, Ghose–Crippen–Viswanadhan $\log P$; AMR, Ghose–Crippen–Viswanadhan molar refractivity; MW, molecular weight; A, number of atoms; P%, the percentage of covering.

compounds is less than 20%. To make the search/design for new drugs more efficient the indices based on the preferred range (50%) may be used, even if the chance of missing good compounds increases in this case.

Note that as these indices depend on ALOGP, their values are provided only for compounds having C, H, O, N, S, Se, P, B, Si, and halogens.

The **rule-of-unity**, proposed by Yalkowski *et al.* [Sanghvi, Ni *et al.*, 2003; Yalkowsky, Johnson *et al.*, 2006], is a drug-like filter based on a single **absorption parameter** Π calculated by the ratio of the $\rightarrow$ octanol–water partition coefficient, K_{ow} , over the *luminal oversaturation number* O_{Lumen} , that is,

$$\Pi = \frac{K_{ow}}{O_{Lumen}} = \frac{K_{ow}}{\max\left(1, \frac{4 \cdot Dose}{S_w}\right)}$$

The absorption parameter was defined to predict whether or not at least half of the administered drug will be absorbed.

The **luminal oversaturation number** is defined as the maximum of either unity or four times the dose in grams per 0.250 l of water divided by the aqueous solubility, S_w , of the drug in grams per liter [Sanghvi, Ni *et al.*, 2003].

The luminal oversaturation number is a dimensionless number that cannot be less than unity and distinguishes between drugs that are soluble in the gastrointestinal contents from drugs that are not. The former will dissolve readily, whereas the latter will exist as suspensions that will maintain a saturated solution in the gut until sufficient absorption has taken place so that no suspended particles remain. For the calculation of solubility, the **general solubility equation** of Jain–Yalkowsky is used [Yalkowsky, 1999; Jain and Yalkowsky, 2001]:

$$\log S_w = 0.5 - \log K_{ow} - 0.01 \times (MP - 25)$$

where MP is the $\rightarrow$ *melting point*.

Drugs with a Π absorption parameter greater than unity tend to be well absorbed (i.e., absorbed fraction > 0.5), while drugs with Π values of less than or equal to 1 are poorly absorbed (absorbed fraction < 0.5). Thus, absorption is most efficient and hence drug-likeness more likely when the absorption parameter Π is greater than unity. This most often occurs when the partition coefficient is greater than unity and/or the oversaturation number is equal to unity.

• lead-like indices

The term “drug-like” is used for compounds resembling existing drugs, while the term “lead-like” for compounds possessing the structural and physico-chemical profile of a quality lead [Verheij, 2006]. The concept of lead-like is more restrictive for some terms with respect to the concept of drug-like [Monge, Arrault *et al.*, 2006], depending on the fact that optimization of a lead compound often results in an increase of molecular weight, $\log P$ and complexity and in a decrease of solubility [Teague, Davis *et al.*, 1999; Hann, Leach *et al.*, 2001; Oprea, 2002a].

Leads should display the following properties to be considered for further development [Oprea, Davis *et al.*, 2001]: (1) relative simple chemical features; (2) membership to a well-established structure–activity relationship series, wherein compounds with similar (sub)-structure exhibit similar target binding affinity; (3) favorable patent situations; and (4) good $\rightarrow$ *ADME properties*. Moreover, in a strict sense, the definition of leads requires the presence of at least one marketed drug, derived from that particular lead structure.

On an average, compared to drugs, leads have lower molecular complexity (lower molecular weight, less rings and rotatable bonds), lower polarizability, are less hydrophobic (their $\log P$ is 0.5–1.0 units less than that of drugs), and have lower drug-like scores [Hann, Leach *et al.*, 2001; Oprea, Davis *et al.*, 2001]. Therefore, in general, physico-chemical property values used as a measure of lead-likeness should be lower than those traditionally used for drug-likeness. Moreover, structural features need to be accounted for in defining lead-likeness since there are various different types of structures that yield false positive hits, such as reactive structures or those that irreversibly bind to the target [Rishton, 2003; Lipinski, 2004].

A collection of lead-like indices is reported in Table P8.

Table P8 Lead-like filters.

	Congreve <i>et al.</i>	Oprea <i>et al.</i>	Hann–Oprea	Verheij	Monge <i>et al.</i>	Wenlock <i>et al.</i>
MW	<300	≤450	≤460	≤450	≤460	≤473
$\log P$	≤3	[−3.5; 4.5]	[−4; 4.2]	[−2.0; 4.5]	[−4.0; 4.2]	[−2.0; 5.5]
HBA	≤3	≤8	≤9	≤10	≤9	≤7
HBD	≤3	≤5	≤5	≤5	≤5	≤4
RBN	≤3	≤10	≤10	≤10	≤10	≤10
NRG	—	≤4	≤4	—	≤4	≤4
$\log D_{7.4}$	—	[−4; 4]	—	—	—	≤4.3
PSA (Å ²)	≤60	—	—	≤150	—	—
$\log S_w$	—	—	≥−5	≥−6	—	—
Halogens	—	—	—	*	≤7	—
N + O	—	—	—	—	≥1	—

MW, molecular weight; HBA, number of hydrogen-bond acceptors; HBD, number of hydrogen-bond donors; RBN, number of rotatable bonds; NRG, number of rings (cyclomatic number); $\log D_{7.4}$, log of the distribution coefficient at pH 7.4; PSA, partial surface area; $\log S_w$, water solubility; Halogens, number of halogen atoms; N + O, total number of nitrogen and oxygen atoms. Data from [Congreve, Carr *et al.*, 2003; Oprea, Davis *et al.*, 2001; Hann and Oprea, 2004; Verheij, 2006; Monge, Arrault *et al.*, 2006; Wenlock, Austin *et al.*, 2003].

The filter proposed by Congreve *et al.* was called the **rule-of-three** (RO3) because, by analogy with the Lipinski → *rule-of-five*, the limits on molecular properties are all multiples of three instead of five.

The filter proposed by Verheij for lead-like compound selection is based on seven molecular descriptors representing molecular properties involved in early discovery, such as oral availability and permeability [Verheij, 2006]. Cutoff values of the descriptors were derived from [Lipinski, Lombardo *et al.*, 1997; Lipinski, 2000; Hann, Leach *et al.*, 2001; Oprea, 2002a; Veber, Johnson *et al.*, 2002]. Moreover, the polar surface area is estimated by the model of the → *topological polar surface area* (TPSA). The filter of Monge *et al.* is an extension of the filter of Hann and Oprea, which includes some additional structural rules. To the limits on the number of halogen atoms and the total number of oxygen and nitrogen atoms, the filter also includes the following rules: (a) no atoms other than C, H, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, or Li; (b) no reactive functions; (c) no perfluorinated chains (e.g., $-\text{CF}_2\text{CF}_2\text{CF}_3$); (d) no rings with more than seven members; (e) alkyl chains $\leq -(\text{CH}_2)_6\text{CH}_3$.

The filter of Wenlock *et al.* is derived from a statistical analysis of a set of marketed oral drugs that are compounds with acceptable physico-chemical properties that have successfully enabled them to overcome the obstacles of development for their desired therapeutic indication.

• **functional group filters** ($\equiv$ *chemical filters*)

Functional group filters are applied to exclude from a chemical database those structures that possess undesired functionalities. These can be structures having more than one aldehyde group, structures containing metals, reactive alkyl halides, peroxides, carbazides.

In general, these filters are designed to recognize those functional groups that tend to be toxic or unstable under physiological conditions. A survey of reactive structures that should be avoided in selection of drug or lead candidates is reported by [Rishton, 2003] (Table P9).

Table P9 List of functional groups responsible for electrophilic protein-reactive false positive from Rishton [Rishton, 2003].

Sulfonyl halides	Acyl halides	Alkyl halides
Anhydrides	Halopyrimidines	α -Halocarbonyl compounds
1,2-Dicarbonyl compounds	Aldehydes	Aliphatic ketones
Perhalo ketones	Aliphatic esters	Imines
Epoxides	Aziridines	Thioesters
Sulfonate esters	Phosphonate esters	Heteroatom–heteroatom
Michael acceptors	β -Heterosubstituted carbonyl compounds	single bonds

Examples of structural filters implemented in the program REOS are listed in Table P10 [Walters and Murcko, 2002].

Table P10 List of REOS functional group filters from [Walters and Murcko, 2002].

Sulfonyl halides	Nitro groups	Aldehydes
Primary alkyl halides	Epoxides	Aziridines
Sulfonate esters	Phosphonate esters	Long aliphatic chains
Peroxides	1,2-Dicarbonyl compounds	Acyl halides

To remove potentially toxic compounds, functional group filters primarily draw from mutagenicity, carcinogenicity, and acute toxicity database [Muegge, 2003].

The **structural alerts** (SA) are chemical filters highlighting molecular substructures or reactive groups that are mainly related to the carcinogenic and mutagenic properties of the chemicals, and represent a sort of “codification” of a long series of studies aimed at highlighting the mechanisms of action of the mutagenic and carcinogenic chemicals [Benigni and Bosa, 2006]. A review about carcinogenic and mutagenic effects and related QSAR models was published by [Frierson, Klopman *et al.*, 2006].

A very effective representation of the structural alerts has been provided by Ashby [Ashby, 1985; Ashby and Tennant, 1988] in the form of a hypothetical poly-carcinogen chemical comprised of most of the known SAs (Figure P2). In Table P11, the structural alert groups are collected.

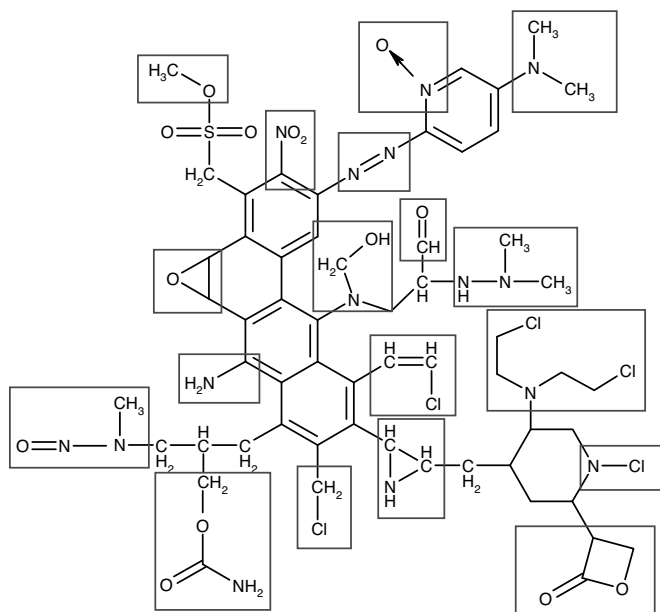


Figure P2 The hypothetical poly-carcinogen chemical proposed by Ashby [Ashby, 1985; Ashby and Tennant, 1988]

Table P11 Structural alerts proposed by Ashby [Ashby, 1985; Ashby and Tennant, 1988].

Structural alert	Structural alert
Aromatic nitro groups	Alkyl esters of either phosphoric or sulfonic acids
Aromatic rings <i>N</i> -oxides	Aromatic mono- and dialkylamino groups
Alkyl hydrazines	Aromatic azo groups (because of possible reduction to aromatic amines)
Alkyl aldehydes	Aromatic and aliphatic aziridinyl derivatives
<i>N</i> -methyl derivatives	Aromatic and aliphatic substituted primary alkyl halides
Monoalkenes	Aromatic amines (including their <i>N</i> -hydroxy derivatives and the derived esters)
β -Haloethyl mustards	Propiolactones and propiosultones
<i>N</i> -Chloroamines	Derivatives of urethane (carbamates)
Alkyl <i>N</i> -nitrosoamines	Aliphatic and aromatic epoxides

Each of the SAs is a “code” for a well-characterized chemical class, with its own specific mechanism of action. However, there are also general physico-chemical factors that may influence the potential reactivity of a chemical, that is, one could expect to observe compounds with structurally alerting features but that are biologically inactive because of a number of reasons, such as molecular weight, solubility, reactivity, and so on.

Starting from eight general toxicophores from the Ashby compilation, a list of 29 toxicophores containing new substructures was proposed to classify compounds according to their mutagenicity (Table P12) [Kazius, McGuire *et al.*, 2005].

Table P12 Extended list of structural alerts according to [Kazius, McGuire *et al.*, 2005].

Structural alerts	Structural alerts
Specific aromatic nitro	Unsubstituted heteroatom-bonded heteroatom
Specific aromatic amine	Nitrogen and sulfur mustard
Aromatic nitroso	Polycyclic aromatic systems (PAH)
Alkyl nitrite	Bay-region in PAH
Nitrosoamine	K-region in PAH
Epoxide	Aliphatic <i>N</i> -nitro
Aziridine	α,β -Unsaturated aldehydes (including R-carbonyl aldehydes)
Azide	Diazonium
Diazo	β -Propiolactone
Triazene	α,β -Unsaturated alkoxy group
Aromatic azo	1-Aryl-2-monoalkyl hydrazine
Carboxylic acid halide	Aromatic methylamine
Aromatic hydroxylamine	Ester derivative of aromatic Hydroxylamine
Aliphatic halide	Polycyclic planar systems
Sulfonate-bonded carbon (alkyl alkane sulfonate or dialkyl sulfate)	

Structural alerts were also searched for within the framework of the **Threshold Toxicological Concern (TTC)**, aimed at reducing extensive toxicity evaluations [Benigni and Bosa, 2006]. This approach refers to the establishment of a generic human exposure threshold value for groups of chemicals below which there would be no appreciable risk to human health. The underlying principle is that such a value can be identified for many chemicals, including those of unknown toxicity, when considering their chemical structures and the known toxicity of chemicals that share similar structural characteristics. Moreover, the concept that there are levels of exposure that do not cause adverse effects is strictly related to the possibility of setting $\rightarrow$ *acceptable daily intakes* for chemicals with known toxicological profiles. A general *TTC* approach, mainly based on carcinogenicity data, was adopted by the US Food and Drug Administration Threshold of Regulation for indirect food additives. An extension of this approach to a range a dietary concentrations was proposed by using QSARs, genotoxicity and short term toxicity data [Cheeseman, Machuga *et al.*, 1999]. This resulted in the identification of eight more complex, less generalized structural alerts, that include a majority of the most potent of the 709 carcinogens (Table P13). This study shows that the inclusion of structural alerts as criteria for substances proposed for approval under a threshold of regulation process, can significantly increase the safety

Table P13 Structural alerts in the *TTC* approach according to [Cheeseman, Machuga *et al.*, 1999].

Structural alerts	Structural alerts
<i>N</i> -nitroso compounds	α -Nitro furyl compounds
Endocrine disruptors	Hydrazines/triazine/azides/azoxy compounds
Strained heteronuclear rings	Polycyclic amines
Heavy metal compounds	Organophosphorous compounds

assurance margin. Substances that do not belong to any of the structural alert classes are likely to have much lower carcinogenic potencies, and therefore may qualify for a higher threshold level.

📖 [Gayoso and Kimri, 1990b, 1990a; Bemis and Murcko, 1996; Stahl and Böhm, 1998; Clark, 1999b, 1999a; Kelder, Grootenhuys *et al.*, 1999; Walters, Ajay *et al.*, 1999; Egan, Merz Jr. *et al.*, 2000; Sakaeda, Okamura *et al.*, 2001; Borodina, Filimonov *et al.*, 2002; Egan and Lauri, 2002; Norinder and Haeberlein, 2002; Engkvist, Wrede *et al.*, 2003; Fichert, Yazdanian *et al.*, 2003; Olah, Bologa *et al.*, 2004b; Vieth, Siegel *et al.*, 2004; te Heesen *et al.*, 2007; te Heesen, Schlitter *et al.*, 2007]

- **properties matrix** → topoelectric matrices
- **protein folding degree index** → biodescriptors (⊙ peptide sequences)
- **protein sequences** → biodescriptors (⊙ peptide sequences)
- **protein TOMOCOMD descriptors** → TOMOCOMD descriptors
- **proteo-chemometrics approach** → Structure/Response Correlations
- **PRP** ≡ *Probabilistic Receptor Potential*
- **proteomics maps** → biodescriptors
- **PRS index** → distance matrix
- **pruning of the graph** → centric indices (⊙ Balaban centric index)
- **pruning partition** → centric indices (⊙ Balaban centric index)
- **pseudocenter** → center of a graph
- **pseudoconnectivity indices** → electrotopological state indices
- **pseudograph** → graph

■ P_VSA descriptors

These are molecular descriptors defined as the amount of van der Waals surface area (VSA) having a property P in a certain range [Labute, 2000]. These descriptors correspond to a partition of the molecular surface area conditioned by the atomic values of the property P.

To generate P_VSA descriptors, first, the van der Waals surface area VSA_i of each atom is estimated according to the following:

$$VSA_i = 4 \cdot \pi \cdot R_i^2 - \pi \cdot R_i \cdot \sum_{j=1}^A a_{ij} \cdot \left(\frac{R_j^2 - (R_i - g_{ij})^2}{g_{ij}} \right)$$

where R is the atomic van der Waals radius, the summation goes over all the atoms, but accounts only for contributions from atoms bonded to the i th atom, a_{ij} being the elements of the → *adjacency matrix*. The quantity g_{ij} is calculated as

$$g_{ij} = \min \{ \max \{ |R_i - R_j|, b_{ij} \}, (R_i + R_j) \}$$

where the term b_{ij} is the ideal length of the bond formed by atoms i and j , calculated according to the formula:

$$b_{ij} = r_{ij}^* - c_{ij}$$

where r_{ij}^* is a reference bond length and c_{ij} a correction term depending on the $\rightarrow$ bond multiplicity: 0 for single bond, 0.1 for aromatic, 0.2 for double, and 0.3 for triple bonds.

In Tables P14 and P15, the van der Waals radii and the reference bond lengths used for P_VSA calculations are collected.

Table P14 van der Waals radii (in Angstrom) used for P_VSA calculations.

Atom-type	R	Atom-type	R
H (-O)	0.8	O (other)	1.779
H (-N, -P)	0.7	F	1.496
H (other)	1.485	P	2.287
C	1.950	S	2.185
N	1.950	Cl	2.044
O (oxide)	1.810	Br	2.166
O (acid)	2.152	I	2.358

Table P15 Reference bond lengths (in Angstrom) used for P_VSA calculations. The symbol $\sim$ indicates any kind of bond.

Bond-type	r^*	Bond-type	r^*	Bond-type	r^*
C $\sim$ C	1.540	H-N	1.010	N-N	1.450
C-H	1.060	H-O	0.970	N-O	1.460
C-N	1.470	H-P	1.410	N-P	1.600
C-O	1.430	H-S	1.310	N-S	1.760
C-P	1.850	H-F	0.870	O-O	1.470
C-S	1.810	H-Cl	1.220	O-P	1.570
C-F	1.350	H-Br	1.440	O-S	1.570
C-Cl	1.800	H-I	1.630	P-P	2.260
C-Br	1.970			P-S	2.070
C-I	2.120			S-S	2.050

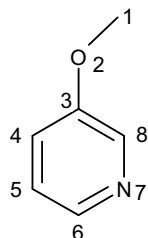
Let P_i be a property of the i th atom. Then, the P_VSA descriptors are defined as

$$P\_VSA_k = \sum_{i=1}^A VSA_i \cdot \delta(P_i \in [a_{k-1}, a_k]) \quad k = 1, 2, \dots, n$$

where the summation goes over all the atoms, VSA_i is the van der Waals surface area of the i th atom, $\delta(P_i \in [a_{k-1}, a_k])$ is the Dirac delta function that is equal to 1 for atoms with property value in the specified range, and zero otherwise; $a_0 \leq a_k < a_n$ are interval boundaries such that $[a_0, a_n]$ bounds all values of the property P in any molecule of the data set.

Example P2

A hypothetical atomic property is used to partition the van der Waals surface area into six different regions so that a six-dimensional P_VSA vector results.



Atom	1	2	3	4	5	6	7	8
P_i	2.4	1.2	4.5	5.9	5.7	3.1	0.2	3.9
VSA_i	9.2	6.3	2.2	4.5	4.5	4.4	4.6	4.4

$P\_VSA(0, 2)$	$= VSA_2 + VSA_7 = 6.3 + 4.6$	$= 10.9$
$P\_VSA(2, 3)$	$= VSA_1$	$= 9.2$
$P\_VSA(3, 4)$	$= VSA_6 + VSA_8 = 4.4 + 4.4$	$= 8.8$
$P\_VSA(4, 5)$	$= VSA_3$	$= 2.2$
$P\_VSA(5, 6)$	$= VSA_4 + VSA_5$	$= 9.0$
$P\_VSA(6, 7)$	$= -$	$= 0$

P_VSA descriptors were calculated from several properties, such as atomic weight (m_VSA), atom polarizabilities (p_VSA), atom-type counts (a-nX_VSA), $\log P$ ($S \log P\_VSA$), molar refractivity (SMR_VSA), connectivity (δ_VSA), van der Waals volume (vdw_VSA), van der Waals surface (vsa_VSA), and van der Waals density (molecular weight divided by van der Waals volume, den_VSA), hydrogen-bond donor (HBD_VSA) and acceptor (HBA_VSA), polar atom (hydrogen-bond donors plus hydrogen-bond acceptors (POL_VSA), hydrophobic atom (hyd_VSA), and partial charges (PEOE_VSA).

Table P16 P_VSA descriptors in terms of $\log P$ [Labute, 2000], molar refractivity [Wildman and Crippen, 1999], and partial charges [Gasteiger and Marsili, 1980]. Property interval boundaries were optimized using data from a database of 44 795 small organic compounds.

Property	No.	Interval boundaries for the calculation of P_VSA descriptors
$\log P$	10	$(-\infty, -0.4, -0.2, 0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, +\infty)$
Molar refractivity	8	$(0, 0.11, 0.26, 0.35, 0.39, 0.44, 0.485, 0.56, +\infty)$
Partial charges	14	$(-\infty, -0.3, -0.25, -0.20, -0.15, -0.10, -0.05, 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, +\infty)$

This methodology can be easily extended replacing atomic properties with any $\rightarrow$ local vertex invariant $\mathcal{L}_i$ and, in particular, with local vertex invariants obtained by atomic

properties using the Randić-like formula, that is, taking into account all the bonds incident to the atom.

📖 [Baurin, Mozziconacci *et al.*, 2004; Burton, Ijjaali *et al.*, 2006; Dubus, Ijjaali *et al.*, 2006; Ijjaali, Petitet *et al.*, 2007; Klon and Diller, 2007; Moorthy, Karthikeyan *et al.*, 2007]

➤ **P weighting scheme** → weighting schemes



■ QikProp descriptors

QikProp is a software [QikProp – Schrödinger, 2003] that calculates a set of 31 physico-chemical descriptors generated by using the program MacroModel [MacroModels – Schrödinger, 1990] for $\rightarrow$ ADME property prediction of drug candidates. The list of QikProp descriptors is given in Table Q1.

Table Q1 List of the descriptors calculated by QikProp software from Jensen, Sørensen *et al.* (2003).

Code	QikProp descriptors
# Stars	Drug-likeness
# Rotor	Number of rotatable bonds
# rctvFG	Number of reactive functional groups
CNS	Predicted central nervous system activity on a -2 (inactive) to 2 (active) scale
MW	Molecular weight of the molecule
Dipole	Dipole moment for the molecule
SASA	Total solvent-accessible surface area in $\text{\AA}^2$
FOSA	Hydrophobic component of the SASA (saturated carbons and attached hydrogen)
FISA	Hydrophilic component of the SASA (SASA on N, O and H on heteroatom)
PISA	Pi (carbon and attached hydrogen) component of the SASA
WPSA	Weakly polar component of the SASA (halogens, P and S)
Volume	Total solvent-accessible volume in $\text{\AA}^3$
DonorHB	Estimated number of hydrogen bonds that would be donated by solute to water molecules in an aqueous solution
AcptHB	Estimated number of hydrogen bonds that would be accepted by solute from water molecules in an aqueous solution
Dip $\wedge$ 2/V	Kirkwood–Onsager dipole solvent index [(dipole moment) ^{2} /volume]
AC $\times$ DN $\wedge$.5/SA	Index of cohesive interactions in solids
Glob	Globularity ($4 \times \pi \times r^2$ /SASA)
QPpolrz	Predicted polarizability in $\text{\AA}^3$
QPlogPC	Predicted log of the hexadecane/gas partition coefficient
QPlogPoc	Predicted log of the octanol/gas partition coefficient
QPlogPw	Predicted log of the water/gas partition coefficient
QPlogPo/w	Predicted log of the octanol/water partition coefficient
QPlogS	Predicted aqueous solubility

(Continued)

Table Q1 (Continued)

Code	QikProp descriptors
BIPcaco	Predicted apparent Caco-2 cell permeability in nm/s. Boehringer–Ingelheim scale
AffyPCaco	Predicted apparent Caco-2 cell permeability in nm/s. Affymax scale
QPlogBB	Predicted log of the brain/blood partition coefficient
AffyPMDCK	Predicted apparent MDCK cell permeability in nm/s. Affymax scale.
QPlogKp	Predicted log of the skin permeability
IP(eV)	Calculated ionization potential
EA(eV)	Calculated electron affinity
# Metabol	Number of likely metabolic reactions that are listed in the propout file

- **Q polarity index** → electric polarization descriptors
- **QSAR** → Structure/Response Correlations
- **QSPR** → Structure/Response Correlations
- **quadratic index** → Zagreb indices
- **quadratic indices** → TOMOCOMD descriptors
- **quadratic mean** $\equiv$ *root mean square* → statistical indices ($\odot$ indices of central tendency)
- **quadrupole moment** → electric polarization descriptors
- **quantiles** → statistical indices
- **quantitative information analysis** → Structure/Response Correlations
- **quantitative molecular similarity analysis** → similarity/diversity
- **quantitative shape–activity relationships** → Structure/Response Correlations
- **quantitative similarity–activity relationships** → similarity/diversity
- **quantitative structure–activity relationships** → Structure/Response Correlations
- **quantitative structure–chromatographic relationships** → Structure/Response Correlations
- **quantitative structure–enantioselective retention relationships** → Structure/Response Correlations
- **quantitative structure–property relationships** → Structure/Response Correlations
- **quantitative structure–reactivity relationships** → Structure/Response Correlations
- **quantitative structure/response correlations** → Structure/Response Correlations
- **quantitative structure–toxicity relationships** → Structure/Response Correlations

■ quantum-chemical descriptors

Quantum-chemical descriptors ($\rightarrow$ *computational chemistry*) are calculated by solving the Schrödinger equation:

$$\mathcal{H}\psi_i = E_i\psi_i$$

where $\mathcal{H}$ is the many-electron Hamiltonian operator [Streitwieser, 1961; Salem, 1966; Dewar, 1969; Kier, 1971; Leach, 1996; Clark, 2003b], ψ_i is the wave function corresponding to the i th electronic state and E_i is the energy of the i th electronic state. The spectrum of the energy levels is itself a molecular descriptor. Moreover, from the wave function ψ_i several important physical properties are obtained. The Schrödinger equation can be solved exactly (analytically) only for the H atom or the H_2^+ molecule. For polyelectronic atoms and molecules, approximate methods are needed. In *ab initio methods*, the first approximation of the Schrödinger equation relies on the wave function that is defined as a **Slater determinant** of monoelectronic functions ϕ_i (i.e.,

molecular orbitals):

$$\Psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_1(1) & \varphi_2(1) & \cdots & \varphi_N(1) \\ \varphi_1(2) & \varphi_2(2) & \cdots & \varphi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(N) & \varphi_2(N) & \cdots & \varphi_N(N) \end{vmatrix} \quad \varphi_i = \sum_{\mu} c_{\mu i} \cdot \phi_{\mu}$$

where molecular orbitals φ_i in turn are determined as a linear combination of monoelectronic atomic functions ϕ_{μ} (i.e., *atomic orbitals*).

In the second equation, c are the coefficients of the linear combination of atomic orbitals defining each i th molecular orbital φ_i . The application of the variational principle to a wave function in the form of Slater determinant leads to the set of the *Hartree–Fock equations* (HF):

$$\hat{f}\varphi_i = \varepsilon_i\varphi_i$$

which can be solved by the Self Consistent Field–Molecular Orbital (SCF–MO) method. In the equation above, $\hat{f}$ is the monoelectronic Fock operator, φ_i is the i th molecular orbital, and ε_i is the corresponding orbital energy.

In the monodeterminant HF wave function, any electron moves in a spherical average potential generated by all other electrons. The energy calculated at the HF level (E_{HF}) is therefore affected by an error due to the instantaneous repulsion of the electrons. The difference between the exact (nonrelativistic) energy E_0 and E_{HF} is called *correlation energy* E_{corr} and it is due to the fact that the correlated motion of the electron is not being properly taken into account in the HF scheme. Accurate wave functions and energies, in which a fraction of E_{corr} is recovered can be calculated through the expression of the wave function as a linear combination of Slater determinants generated from the reference HF wave function by exciting electrons from occupied to virtual orbitals. These methods are generally indicated as *correlated* or *post HF* methods and are generally much more computational demanding than the HF method itself. Examples of these approaches are the Configuration Interaction scheme (CI), in which the variational principle is applied to the multideterminant wave function, and the Møller–Plesset scheme, in which the perturbation treatment is applied to the HF wave function, by considering the set of Fock operators as the zeroth order Hamiltonian [Szabo and Ostlund, 1996; Jorgensen, Olsen *et al.*, 2000].

In the opposite direction with respect to the *ab initio post-HF* methods are the *semiempirical* approaches, which, in the most recent implementations, solve the HF-SCF set of equations by using an Hamiltonian parameterized to calculate properties fitting experimental data. In these approaches, only valence electrons are considered explicitly, and a limited number of atomic orbitals is used. Most of the *semiempirical* methods are based on the *zero differential overlap* (ZDO) approximation. In the ZDO approximation, the products of the atomic orbitals centered on different atoms are set to zero. Due to this approximation, a lot of two electron integrals are neglected. The remaining integrals are generally parameterized with reference to experimental data. Several models, which differ for the level of approximation and the set of parameters such as the *neglect of diatomic differential overlap* (NDDO) model, the *intermediate neglect of differential overlap* (INDO) model, the *modified neglect of differential overlap* (MNDO) model, the *Austin Model 1* (AM1) model and the *Parameterized Model 3* (PM3) model has been proposed.

Semiempirical methods are significantly less computational demanding than *ab initio* approaches, enabling calculations on molecular systems of very large size. Results from semiempirical calculation can be of quality as good as that of *post-HF* methods due to the parameterization procedure with reference to experimental data, but can be very poor when applied to class of molecules or for molecular properties not considered in the parameterization.

Finally, **Density Functional Theory** (DFT) methods [Parr and Yang, 1989; Geerlings, De Proft *et al.*, 1996, 2003; Geerlings, Langenaeker *et al.*, 1996; Koch and Holthausen, 2001] are based on the two Hohenberg–Kohn theorems, which state that the ground-state electronic energy is uniquely determined by the electron density, or in other words, that the electronic energy is a functional of the electron density; $E_0 = F[\rho]$. However, the exact form of this functional is not known and the finding of approximate functionals for accurate calculations is a current research topic. In the widely used Kohn–Sham implementation of DFT, molecular orbitals are introduced to calculate the electron density. In this scheme, the energy for the nuclei–electrons interaction and the Coulomb interaction is calculated exactly from the classical expressions, the kinetic energy is calculated exactly for a system of noninteracting electrons (as in HF), and the exchange and correlation energy is calculated using an approximate functional $E_{xc}[\rho]$. In the last contribution is also included the part of the kinetic energy not considered in the noninteracting electrons model. $E_{xc}[\rho]$ is therefore the term containing everything cannot be calculated exactly, and it is the focus of the current research. The mathematical development of the Kohn–Sham scheme leads to a set of equations that are known as the Kohn–Sham (KS) equations:

$$\hat{h}_{KS}\varphi_i = \varepsilon_i\varphi_i$$

Although KS equations look like very similar to HF equations and are solved using the same SCF iterative method, they are based on a different physical basis. In fact, as discussed above the KS scheme relies on the electron density, whereas the HF scheme relies on the wave function. This also means that the KS orbitals φ_i and orbital energies ε_i should not have the physical meaning they have in the HF scheme (see below Koopman’s theorem). It should be noted, however, that the physical meaning of KS molecular orbitals is still a matter of debate, and that recent results show that they are reliable for qualitative analysis and explanation of chemical reactivity. We should also outline that, contrary to HF, the KS scheme allows the calculation of the correlation energy through the inclusion of the exchange–correlation functional. This is one of the main advantages of the DFT–KS method; results of quality comparable to those of *ab initio post-HF* methods can be obtained with a computational cost comparable to that of HF. On the other hand, the empirical nature of the exchange–correlation functionals proposed so far, does not allow a way to systematically improve the accuracy of the calculations. In this respect, DFT methods are more similar to the *semiempirical* schemes.

Several different kinds of quantum-chemical descriptors have been defined and these can be broadly divided into *energy-based descriptors*, *orbital energies descriptors*, *local quantum-chemical properties*, descriptors based on the analysis of the wave function, frontier orbital electron densities, superdelocalizability indices, polarizabilities, and derived from the Density Functional Theory [Cartier and Rivail, 1987; Bergmann and Hinze, 1996; Karelson, Lobanov *et al.*, 1996].

The most common quantum-chemical descriptors are listed below.

Molecular energies calculated by $\rightarrow$ *computational chemistry* methods are fundamental descriptors commonly used in QSAR models; moreover, energies are used as cutoff values for the selection of the most important conformation(s). Besides the energy levels calculated from quantum-mechanical calculations and the molecular total energy, several other kinds of

energies are defined such as heat of formation, heat of solvation, heat of vaporization, steric energy, and so on.

→ *Substituent front strain*, → *steric energy difference*, → *Joshi steric descriptor* and → *Joshi electronic descriptors* are examples of molecular descriptors calculated from the standard enthalpy of formation and the steric energy.

The whole set of orbital energies, the set of the eigenvalues $\{E_i\}$, is of fundamental importance to explain the global and local behavior of a molecule.

The **energy moments** of k th order are molecular descriptors defined as [Burdett, 1995]

$${}^k\mu = \sum_i E_i^k = \text{tr}(\mathbf{H}^k)$$

where $\mathbf{H}$ is the Hamiltonian matrix raised to the k th power and tr denotes the trace of the matrix. The k th energy moment may simply be interpreted as a weighted sum over all the → *self-returning walks* of the orbitals.

Two other important and useful molecular energy measures are ionization potential and electron affinity. The **ionization potential** IP (or **ionization energy**) is defined as the energy needed to extract one electron from a chemical system, that is,

$$\text{IP} = E(N_{\text{el}}) - E(N_{\text{el}} - 1)$$

where N_{el} is the number of electrons of the system. Ionization potential is a measure of the capability of a molecule to give the corresponding positive ion.

The **electron affinity** EA is defined as the gain in energy of the chemical system when an electron is captured from the system, that is,

$$\text{EA} = E(N_{\text{el}}) - E(N_{\text{el}} + 1)$$

Electron affinity is a measure of the capability of a molecule to give the corresponding negative ion.

Ionization potential and electron affinity are the basic quantities that give rise to the definition of → *atom electronegativity*.

Other fundamental **energy-based descriptors** have been defined in the framework of the computational chemistry.

Molecular orbital energies give information about reactivity/stability of specific regions of the molecule. Among the molecular orbitals, a fundamental role is played by the **frontier orbitals** [Fleming, 1990; Clare, 1994; Huang, Kong *et al.*, 1996], that are the orbitals involved in a transition state, that is, the **highest occupied molecular orbital** (HOMO), the **lowest unoccupied molecular orbital** (LUMO), and the **singly occupied molecular orbital** (SOMO) for radicals.

These orbitals play a major role in governing many chemical reactions and are responsible for the formation of many charge-transfer complexes.

The main descriptors based on molecular orbital energies are the following.

- **highest occupied molecular orbital energy** (ϵ_{HOMO})

The energy of the highest energy level containing electrons in the molecule. Molecules with high HOMO energy values can donate their electrons more easily compared to molecules with low HOMO energy values, and hence are more reactive. Therefore, within the validity of the Koopman's theorem, the ϵ_{HOMO} descriptor is related to the → *ionization potential* IP

($IP = -\epsilon_{\text{HOMO}}$), is a measure of the nucleophilicity of a molecule and is important in modeling molecular properties and reactivity, in particular for radical reactions.

The energy of second (HOMO-1) and third (HOMO-2) highest occupied orbital has been also used in reaction modeling [Barone, Camilo Jr. *et al.*, 1996; Braga, Barone *et al.*, 1999; Chumakov, Terletskaia *et al.*, 2000; Braga and Galvão, 2003; Aptula, Roberts *et al.*, 2005a]. Moreover, the **Activation Energy Index** (AEI) is a mechanism-based molecular orbital parameter defined as [Aptula, Roberts *et al.*, 2005b]

$$AEI = \epsilon_{\text{HOMO}} + \epsilon_{\text{HOMO-1}}$$

where the energy terms are orbital energies of the electrophiles and the intermediates for Michael addition of *n*-butylamine. The implicit assumption is that as the structure of electrophile is changed, most of the variation in energy difference between electrophile and intermediate is reflected in the ΔE_{HOMO} and $\Delta E_{\text{HOMO-1}}$. In other words, for the π -orbital below HOMO-1, although the energy changes on going from electrophile to intermediate, the energy changes do not vary greatly as the structure of electrophile changes.

- **lowest unoccupied molecular orbital energy** (ϵ_{LUMO})

The energy of the lowest energy level containing no electrons in the molecule. Molecules with low LUMO energy values are more able to accept electrons than molecules with high LUMO energy values. Therefore, within the validity of the Koopman's theorem, the ϵ_{LUMO} energy is related to the $\rightarrow$ *electron affinity* EA ($EA = -\epsilon_{\text{LUMO}}$), is a measure of the electrophilicity of a molecule and is important in the modeling of molecular properties and reactivity, in particular for radical reactions.

📖 [Debnath, Compadre *et al.*, 1991]

- **HOMO–LUMO energy gap** (*GAP*)

The difference between the HOMO and LUMO energies:

$$GAP = \epsilon_{\text{LUMO}} - \epsilon_{\text{HOMO}}$$

where ϵ_{LUMO} and ϵ_{HOMO} are the energies of the lowest unoccupied molecular orbital and the highest occupied molecular orbital, respectively.

GAP is an important stability index, a large *GAP* being related to the high stability of a molecule with its low reactivity in chemical reactions. It is an approximation of the lowest excitation energy of the molecule and can be used for the definition of absolute and activation hardness.

- **HOMO/LUMO energy fraction** ($f_{H/L}$)

A stability index defined as the ratio between the HOMO and LUMO energies:

$$f_{H/L} = \frac{\epsilon_{\text{HOMO}}}{\epsilon_{\text{LUMO}}}$$

where ϵ_{LUMO} and ϵ_{HOMO} are the energies of the lowest unoccupied molecular orbital and the highest occupied molecular orbital, respectively. Low values of $f_{H/L}$ are related to high stability of the molecule.

Applications of HOMO and LUMO and derived descriptors found in literature are [Chastrette, Rajzmann *et al.*, 1985; Mekenyan, Roberts *et al.*, 1997; Vaes, Urrestarazu Ramos *et al.*, 1998; Benigni, Passerini *et al.*, 1999a; Vendrame, Braga *et al.*, 1999; Warne, Boyd *et al.*, 1999; Sullivan, Jones *et al.*, 2000; Trohalaki, Gifford *et al.*, 2000; Agatonovic-Kustrin, Beresford *et al.*, 2001; Chen, Quan *et al.*, 2001c; Garg and Achenie, 2001; Rugutt, Rugutt *et al.*, 2001; Aptula, Netzeva *et al.*, 2002; Bhattacharjee, Kyle *et al.*, 2002; Clare, 2002; El-Taher, El-sawy *et al.*, 2002; Fitch, McGregor *et al.*, 2002; Harju, Andersson *et al.*, 2002; Qi, Zhang *et al.*, 2002; Xu, Yang *et al.*, 2002; Braga and Galvão, 2003; Katritzky, Oliferenko *et al.*, 2003b; Mattioni and Jurs, 2003; Mekenyan, Nikolova *et al.*, 2003; Pinheiro, Kiralj *et al.*, 2003; Safarpour, Hemmateenejad *et al.*, 2003; Braga and Galvão, 2004; Chen, Yao *et al.*, 2004; Clare, 2004; Estrada and Patlewicz, 2004; Lind, Lopes *et al.*, 2004; Turabekova and Rasulev, 2004; Benigni, 2005; Bhat, Hayik *et al.*, 2005; de Lima Ribeiro and Castro Ferreira, 2005; Glossman-Mitnik, 2005; Kelly, Spillane *et al.*, 2005; Mazzatorta, Smiesko *et al.*, 2005; Rajkó, Körtévlyesi *et al.*, 2005; Simón-Manso, 2005; Villanueva-García, Gutiérrez-Parra *et al.*, 2005; Zhang, Qu *et al.*, 2005; Holder, Ye *et al.*, 2006a; Liu, Xiang *et al.*, 2007]

From quantum theory, a number of **local quantum-chemical properties** are defined at each point $\mathbf{r}$ of the molecule space. The most important ones are listed below.

- **electron density**

A local electronic descriptor calculated by quantum-mechanical methods. The electron density $\rho(\mathbf{r})$ at a point $\mathbf{r}$ can be defined by the equation:

$$\rho(\mathbf{r}) = N_{\text{el}} \cdot \int |\Psi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)|^2 d\mathbf{s}_1 d\mathbf{x}_2 \dots d\mathbf{x}_N$$

where N_{el} is the total number of electrons. $\rho(\mathbf{r})$ corresponds to the probability of finding an electron in the volume $d\mathbf{r}$ independently of the position of all other electrons. It can be observed that the integral of $\rho(\mathbf{r})$ over all space equals the number of electrons N_{el} in the system:

$$N_{\text{el}} = \int \rho(\mathbf{r}) d\mathbf{r}$$

In the case of monodeterminant wave functions, $\rho(\mathbf{r})$ can be calculated as the sum of squares of the molecular orbitals ϕ at point $\mathbf{r}$ for all occupied molecular orbitals N_{OCC} ; for a closed-shell system of N_{el} electrons occupying $N/2$ orbitals it is defined as

$$\rho(\mathbf{r}) = 2 \cdot \sum_{i=1}^{N_{\text{OCC}}} |\phi_i(\mathbf{r})|^2$$

The **average local ionization energy** was introduced [Sjöberg, Murray *et al.*, 1990] to estimate the average energy required to remove an electron located at the point $\mathbf{r}$ from the molecule and is defined as

$$\bar{I}(\mathbf{r}) = \frac{\sum_i \rho_i(\mathbf{r}) \cdot |\epsilon_i|}{\rho(\mathbf{r})}$$

where the sum runs over the molecular orbitals, $\rho_i(\mathbf{r})$ is the electronic density of the i th molecular orbital at the point $\mathbf{r}$, and ϵ_i the i th orbital energy.

- **composite nuclear potential**

For a given configuration of the nuclei of a molecule, the composite nuclear potential is defined as

$$v(\mathbf{r}) = \sum_{a=1}^A \frac{Z_a}{|\mathbf{r} - \mathbf{R}_a|}$$

where Z_a are the nuclear charges at positions $\mathbf{R}_a$. Assuming that the electronic density is removed without changing the nuclear configuration, the composite nuclear potential is the potential experienced by a unit charge at location $\mathbf{r}$.

- **Somoyai function**

A special representation of the difference between the $\rightarrow$ *electronic density* $\rho(\mathbf{r})$ at a point $\mathbf{r}$ and the $\rightarrow$ *composite nuclear potential* $v(\mathbf{r})$ at the same point, defined as

$$S(\mathbf{r}, s) = \rho(\mathbf{r}) - s \cdot v(\mathbf{r})$$

The Somoyai parameter s has the physical dimension of bohr^{-2} . Since a notable part of the electronic density is mimicked by the composite nuclear potential, it can be assumed that only their difference provides a description of the chemical bonding. Then, for any fixed value of s , the Somoyai function gives information about the role of the electronic density in the chemical bonding.

- **Molecular Electrostatic Potential (MEP)**

An important reactivity descriptor [Bonaccorsi, Scrocco *et al.*, 1970], giving the interaction energy of a molecule with a unit positive charge at position $\mathbf{r}$, and defined as

$$V(\mathbf{r}) = \sum_{a=1}^A \frac{Z_a}{|\mathbf{r} - \mathbf{R}_a|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \cdot d\mathbf{r}'$$

where Z_a is the nuclear charge of the a th atom in position $\mathbf{R}_a$. In other words, the electrostatic potential at a certain point around the molecule is the work needed to bring a unit positive charge from infinite to that point and it is usually calculated by quantum-mechanical approaches of various degree of approximation.

MEPs give detailed information for studies on chemical reactivity or biological activity of a compound. The spatial distribution and the values of the electrostatic potential determine the attack of an electrophilic or a nucleophilic agent as the primary event of a chemical reaction [Gasteiger, Li *et al.*, 1994a].

MEP is among the most used $\rightarrow$ *molecular interaction fields*.

- **Interaction Index (II)**

Also called **index of charge and orbital controlled interaction (ICOCI)**, this was proposed in the framework of the $\rightarrow$ *Electron Conformational method* to account for electronic properties of atoms, including both orbital and charge reactivity properties of atoms in their donor/acceptor interaction with a receptor [Bersuker, Bahçeci *et al.*, 2000b; Rosines, Bersuker *et al.*, 2001]. The interaction index of a given atom in a given molecular environment is defined as proportional to the maximum energy of interaction of that atom with a target atom.

For the a th atom, the interaction index is calculated as

$$II_a = n_a \cdot \exp[-(2 \cdot IP_a)^{1/2} \cdot R_0]$$

where n is the electron population of the outer orbital (1s for H, np for second and third row elements) in the valence state of the molecule, IP is the $\rightarrow$ *ionization potential*, and R_0 is the distance from the maximum density to the point of assumed maximum overlap with the wave function of the target atom ($R_0 = 2$ au). All the three parameters stand for the orbital contribution to the ligand–receptor interaction and, among them, the IP value is strongly dependent on the atomic charge.

The **population analysis** of the wave function obtained by a quantum-chemical calculation allow to assign **atomic charges** q (or **net atomic charges**) and evaluating **bond orders**. The charges measure the extent of electronic density localization in a molecule. Negative q_a values mean that excess electronic charge is at center a while positive values mean that center a is electron-deficient.

Several schemes for the analysis of the wave function have been proposed. The most commonly used are those proposed by Mulliken and Löwdin, those based on natural bond orbital theory (NBO), the Bader AIM theory, and the fitting of the electrostatic potential.

For closed shell systems, it is useful to define a **charge density matrix** $\mathbf{P}$ as

$$P_{\mu\nu} = 2 \cdot \sum_{i=1}^{N_{\text{OCC}}} c_{\mu i} \cdot c_{\nu i}$$

so that the electron density, at each point $\mathbf{r}$, is expressed as

$$\rho(\mathbf{r}) = \sum_{\mu} \sum_{\nu} P_{\mu\nu} \cdot \phi_{\mu}(\mathbf{r}) \cdot \phi_{\nu}(\mathbf{r})$$

The concept of bond order and atomic charge was first introduced by Coulson in the framework of Hückel semiempirical approach. In the density matrix, Coulson defined the off-diagonal element $P_{\mu\nu}$ as the **bond order** [Coulson, 1939], and the diagonal element $P_{\mu\mu}$ as the **partial atomic charge** of the μ th atomic orbital:

$$P_{\mu\nu} = \sum_{i=1}^{N_{\text{OCC}}} n_i \cdot c_{\mu i} \cdot c_{\nu i} \quad \text{and} \quad P_{\mu\mu} = \sum_{i=1}^{N_{\text{OCC}}} n_i \cdot c_{\mu i}^2 = q_{\mu}$$

where the summation runs over all the occupied molecular orbitals N_{OCC} .

For those methods in which atomic orbitals are not orthogonal, the overlap between orbitals must be considered as in the **Mulliken population analysis** [Mulliken, 1955a, 1955b]. In this case, the equation

$$\rho(\mathbf{r}) = \sum_{\mu} \sum_{\nu} P_{\mu\nu} \cdot \phi_{\mu}(\mathbf{r}) \cdot \phi_{\nu}(\mathbf{r})$$

is integrated over all space, leading to

$$\begin{aligned} N_{\text{el}} &= \int \sum_{\mu} \sum_{\nu} P_{\mu\nu} \cdot \phi_{\mu}(r) \cdot \phi_{\nu}(r) dr = \sum_{\mu} \sum_{\nu} P_{\mu\nu} \cdot \int \phi_{\mu}(r) \cdot \phi_{\nu}(r) dr \\ &= \sum_{\mu} \sum_{\nu} P_{\mu\nu} S_{\nu\mu} = \text{tr}[\mathbf{PS}] \end{aligned}$$

where

$$S_{\mu\nu} = \int \phi_{\mu}(r) \cdot \phi_{\nu}(r) dr$$

are the elements of the **overlap matrix S**. The element $[\mathbf{PS}]_{\mu\mu}$ corresponds to the electron population in the orbital ϕ_{μ} . Therefore, the atomic charge q of an atom A can be calculated by the sum of the contributions given by all of the atomic orbitals centered on A:

$$q_A = Z_A - \sum_{\mu \in A} [\mathbf{PS}]_{\mu\mu}$$

where Z_A is the effective nuclear charge of the atom A.

Each single element of the summation represents the charge of each single atomic orbital and are also useful descriptors.

The off-diagonal terms of the **PS** matrix between pairs of atomic orbitals centered on different atoms can be used to define the **Mulliken bond order** between atom A and atom B (B_{AB}):

$$B_{AB} = 2 \cdot \sum_{\mu \in A} \sum_{\nu \in B} P_{\mu\nu}^{AB} S_{\mu\nu}^{AB}$$

As with other schemes of partitioning the electron density in molecules, Mulliken population analysis is arbitrary and is strongly dependent on the particular basis set employed. However, the comparison of population analyses for a series of molecules is useful for a quantitative description of intramolecular interactions, chemical reactivity, and structural information.

The **Löwdin population analysis** [Löwdin, 1970] is similar to that proposed by Mulliken, but the atomic orbitals are first transformed into an orthogonal set, as are the molecular orbital coefficients.

Due to the properties of **P** and **S** matrices, the total number of electrons can also be calculated as

$$N_{el} = \sum_{\mu} [\mathbf{S}^{1/2} \mathbf{PS}^{1/2}]_{\mu\mu} = \sum_{\mu} [\mathbf{P}']_{\mu\mu}$$

where $P'_{\mu\mu}$ corresponds to the population of symmetrically orthogonalized atomic orbitals. As for the Mulliken analysis, contributions of the orbitals centered on atom A can be summed up to give the atomic charge of A:

$$q_A = Z_A - \sum_{\mu \in A} [\mathbf{P}']_{\mu\mu}$$

The **Natural Bond Orbital analysis** of Weinhold [Foster and Weinhold, 1980; Reed, Weinstock *et al.*, 1985; Reed, Curtiss *et al.*, 1988] generates, departing from canonical MOs, a set of localized one center (core, lone pairs) and two center (π and σ bonds) strongly occupied orbitals, and a set of one center (Rydberg) and two center (σ^* , π^*) weakly occupied orbitals: the NBOs. The Natural Bond Orbitals (NBOs) are obtained by a sequence of transformations from the input basis to give, first, the Natural Atomic Orbitals (NAOs), then the Natural Hybrid Orbitals (NHOs), and finally the Natural Bond Orbitals (NBOs). For NAOs, atomic charges can be calculated as a summation of contributions given by orbitals localized on each atom; moreover, from NBOs, bond order can be also calculated.

Other different definitions of atomic charge and bond order have been suggested by other authors: based on the valence bond theory [Pauling, 1939], or corrected by the eigenvalues of the secular equation [Ham and Ruedenberg, 1958a, 1958b; Ham, 1958; Ruedenberg, 1958]. Moreover, atomic charges can be calculated as point charges defined to fit the $\rightarrow$ *molecular electrostatic potential* obtained from the wave function.

Some other molecular descriptors related to the charge density matrix have been proposed [Salem, 1966; Mayer, 2007]. Some of them are listed below.

The **bond index** B_{ij} (also called **Wiberg index**) was proposed to measure the multiplicity of bonds between two atoms [Wiberg, 1968; Trindle, 1969]. It is defined between two bonded atoms i and j as the square of the off-diagonal elements $P_{\mu\nu}$ of the charge density matrix between atomic orbitals μ on the i th atom and ν on the j th atom, summed over all such distinct orbitals:

$$B_{ij} = \sum_{\mu \in i} \sum_{\nu \in j} P_{\mu\nu}^2$$

It should be noted that the bond index is a function of the square of the charge density matrix elements whereas the bond order is defined in terms of the charge density matrix elements themselves. It is a measure of the extent of electron sharing between two atoms, that is, it reflects the strength of a bond, but it has a disadvantage that it is always positive and hence cannot describe antibonding situations.

Variants and generalizations of the bond index were proposed by Mayer [Mayer, 1986b, 1986a; Wang and Werstiuk, 2003; Ponc and Cooper, 2005]. For example, in the framework of the $\rightarrow$ *AIM theory*, starting from the correlation function characterizing the extent of the electron sharing and defined as

$$C(\mathbf{r}_1, \mathbf{r}_2) = 2 \cdot \rho(\mathbf{r}_1, \mathbf{r}_2) - \rho(\mathbf{r}_1) \cdot \rho(\mathbf{r}_2)$$

in which $\rho(\mathbf{r}_1)$ and $\rho(\mathbf{r}_2)$ denote the ordinary first order densities and $\rho(\mathbf{r}_1, \mathbf{r}_2)$ the corresponding pair density, the **shared electron distribution index** (SEDI) for atoms A and B was proposed [Ponc and Cooper, 2005] as the sum of k_{AB} and k_{BA} , k_{AB} being defined as

$$k_{AB} = \int_A \rho(\mathbf{r}_1) d\mathbf{r}_1 \cdot \int_B \rho(\mathbf{r}_2) d\mathbf{r}_2 - 2 \cdot \int_A d\mathbf{r}_1 \times \int_B \rho(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_2$$

and k_{BA} the complementary quantity. The integration is performed on the basis of the two considered atoms A and B.

The **atomic valency index** V_i represents the valency of the i th atom as the sum of the valencies of its atomic orbitals [Armstrong, Perkins *et al.*, 1973; Gopinathan and Jug, 1983b]. Thus, the valency V_i of the i th atom is given by

$$V_i = \sum_{j \neq i} \sum_{\mu \in i} \sum_{\nu \in j} P_{\mu\nu}^2 = \sum_{\mu \in i} 2 \cdot P_{\mu\mu} - \sum_{\mu \in i} \sum_{\nu \in i} P_{\mu\nu}^2 = \sum_{j \neq i} B_{ij}$$

where the first summation in the first term and the last one runs over all the atoms different from the i th atom, $P_{\mu\nu}^2$ are the squares of the off-diagonal charge density matrix elements, and B_{ij} the bond index.

For closed shell systems, the previous expression of atom valency is the following:

$$V_i = \sum_{\mu \in i} (2 \cdot q_{\mu} - q_{\mu}^2)$$

where the summation goes over all the atomic orbitals of the i th atom and q indicates orbital occupancy. This quantity has a value zero when q_{μ} , the occupancy of the orbital ϕ_{μ} , is either 2 or 0; it has the maximum value of 1 when $q_{\mu} = 1$.

The **molecular valency index** V_M has also been defined as

$$V_M = \frac{1}{2} \cdot \sum_{i=1}^A V_i$$

where the summation runs over all the molecule atoms and V_i is the i th atomic valency index.

The first derivatives of the molecular valency index were also proposed as quantum descriptors [Balawender, Komorowski *et al.*, 1998], where the left-hand-side derivative describes the effect of the electrophilic attack and the right-hand-side derivative measures reactivity toward a nucleophilic attack. This last one is also related to the aromatic character of a molecule, measured by the $\rightarrow$ *diamagnetic susceptibility exaltation*.

The **free valence index** was proposed as a measure of the residual valency of the i th atom in π -electron molecular orbitals [Coulson, 1946]; it is defined as

$$F_i = \pi_i^{\max} - \pi_i^*$$

where $\pi_i^{\max}$ is the maximum bond order of the i th atom and π_i^* is the sum of the bond orders of the bonds connecting the i th atom to all its neighbors. $\pi_i^{\max}$ is usually taken as $\sqrt{3}$, the value for the central carbon atom in trimethylene methane. Topological formulas for the free valence index were proposed by [Gutman, 1978a]. Moreover, a generalization of the original free valence index accounting for σ electrons is the **general free valence index** defined as [Gopinathan and Jug, 1983a]

$$F'_i = V'_i - V_i$$

where V_i is the valency index of the i th atom and V'_i is the “reference valency” of the i th atom chosen as the integer value around which the computed valency of the atom is distributed in a large number of compounds. This general free valence index can also be defined as a percentage:

$$F'_i\% = \frac{V'_i - V_i}{V'_i} \cdot 100$$

representing the residual covalent binding capacity of the i th atom.

Other molecular descriptors derived from the charge density matrix are the **average atom charge density** P_1 and the **average bond charge density** P_2 , defined as

$$P_1 = \frac{\sum_{\mu, \nu} P_{\mu\nu}}{A} \quad \text{and} \quad P_2 = \frac{\sum'_{\mu, \nu} P_{\mu\nu}}{B}$$

where A and B are the number of atoms and bonds, respectively, and the second summation is restricted to bonds [Balasubramanian, 1994].

Frontier orbital electron densities also involve the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), providing useful measures of donor–acceptor interactions in the molecular space.

The main descriptors based on molecular orbital electron densities of the a th atom are the following.

- **electrophilic atomic frontier electron density** (f_a^-)

A descriptor defined as the $\rightarrow$ *electron density* of the HOMO orbital as

$$f_a^- \equiv q^E \equiv EFD_a = \sum_{\mu \in a} (c_{\text{HOMO},\mu})^2$$

where the summation runs over all the HOMO atomic orbitals and c are the coefficients of their linear combination; q^E is usually called **electrophilic charge**.

To compare reactivity of atoms of different molecules, the **electrophilic frontier electron density index** is defined as

$$F_a^- = \frac{f_a^-}{|\epsilon_{\text{HOMO}}|}$$

where the electrophilic frontier electron density is normalized by the energy of the corresponding frontier molecular orbital.

- **nucleophilic atomic frontier electron density** (f_a^+)

A descriptor defined as the $\rightarrow$ *electron density* of the LUMO orbital as

$$f_a^+ \equiv q^N \equiv NFD_a = \sum_{\mu \in a} (c_{\text{LUMO},\mu})^2$$

where the summation runs over all the LUMO atomic orbitals and c are the coefficients of their linear combination; q^N is usually called **nucleophilic charge**.

To compare the reactivity of atoms of different molecules, the **nucleophilic frontier electron density index** is defined as

$$F_a^+ = \frac{f_a^+}{|\epsilon_{\text{LUMO}}|}$$

where the nucleophilic frontier electron density is normalized by the energy of the corresponding frontier molecular orbital.

Superdelocalizabilities indices are $\rightarrow$ *dynamic reactivity indices* of occupied and unoccupied molecular orbitals, provide information about molecular interactions and allow comparison between corresponding atoms in different molecules.

The **superdelocalizability** of an atom is related to the contribution of the atom to the stabilization energy in the formation of a charge-transfer complex with a second molecule or to the ability of a reactant to form bonds through charge transfer.

- **electrophilic superdelocalizability** (S_a^-)

Also known as the **acceptor superdelocalizability**, it is defined as the sum over all occupied molecular orbitals (N_{OCC}) and the atomic orbitals μ of the a th atom of the square atomic orbital

coefficient c , divided by the energy of the corresponding molecular orbital, that is,

$$S_a^- \equiv ESD_s = 2 \cdot \sum_{i=1}^{N_{\text{OCC}}} \frac{\sum_{\mu} c_{i\mu,a}^2}{|\epsilon_i|}$$

It is a measure of the availability of electrons in the a th atom. If the transition states are mainly controlled by the frontier orbital, the electrophilic superdelocalizability is calculated on the highest occupied molecular orbital (HOMO).

The **maximum electrophilic superdelocalizability** among all the atoms in a molecule is a molecular descriptor defined as

$$S_{\text{max}}^- = \max_a(S_a^-)$$

The **total electrophilic superdelocalizability** is the sum over all the atoms of the electrophilic superdelocalizability [Cartier and Rivail, 1987], that is,

$$S_{\text{TOT}}^- \equiv ESD_{\text{TOT}} = \sum_a S_a^-$$

The **average electrophilic superdelocalizability** is derived from the total electrophilic superdelocalizability as

$$\bar{S}^- = \frac{S_{\text{TOT}}^-}{A}$$

where A is the number of atoms.

 [Bearden and Schultz, 1997]

• **nucleophilic superdelocalizability (S_a^+)**

Also known as the **donor superdelocalizability**, it is defined as the sum over all the unoccupied molecular orbitals ($N_{\text{MO}} - N_{\text{OCC}}$), N_{MO} being the total number of molecular orbitals, and over the atomic orbitals μ of the a th atom of the square atomic orbital coefficient c , divided by the energy of the corresponding molecular orbital, that is,

$$S_a^+ \equiv NSD_a = 2 \cdot \sum_{i=N_{\text{OCC}}+1}^{N_{\text{MO}}} \frac{\sum_{\mu} c_{i\mu,a}^2}{|\epsilon_i|}$$

It is a measure of the availability for additional electron density on the a th atom.

If the transition states are mainly controlled by the frontier orbital, the nucleophilic superdelocalizability is calculated on the lowest unoccupied molecular orbital (LUMO).

The **maximum nucleophilic superdelocalizability** among all the atoms in a molecule is a molecular descriptor defined as

$$S_{\text{max}}^+ = \max_a(S_a^+)$$

The **total nucleophilic superdelocalizability** is the sum over all the atoms of the nucleophilic superdelocalizability, that is,

$$S_{\text{TOT}}^+ \equiv NSD_{\text{TOT}} = \sum_a S_a^+$$

The **average nucleophilic superdelocalizability** is derived from the total nucleophilic superdelocalizability as

$$\bar{S}^+ = \frac{S_{\text{TOT}}^+}{A}$$

where A is the number of atoms.

• **radical superdelocalizability** (S_a^0)

This is given by both contributions of the acceptor and donor superdelocalizability, defined as

$$S_a^0 \equiv RSD_a = \sum_{i=1}^{N_{\text{OCC}}} \frac{\sum_{\mu} c_{i\mu,a}^2}{|\epsilon_i|} + \sum_{i=N_{\text{OCC}}+1}^{N_{\text{MO}}} \frac{\sum_{\mu} c_{i\mu,a}^2}{|\epsilon_i|}$$

where N_{MO} and N_{OCC} are the numbers of molecular orbitals and occupied molecular orbitals, respectively; c are the orbital coefficients.

The **total radical superdelocalizability** is the sum over all the atoms of the radical superdelocalizability, that is,

$$S_{\text{TOT}}^0 \equiv RDS_{\text{TOT}} = \sum_a S_a^0$$

The **average radical superdelocalizability** is derived from the total radical superdelocalizability as

$$\bar{S}^0 = \frac{S_{\text{TOT}}^0}{A}$$

where A is the number of atoms.

• **atom polarizability** (π)

The atom polarizability is the ability of the charge of the a th atom to give a linear response to a change of the Coulomb integral of the b th atom, that is, as the derivative of the charge of the a th atom q_a over the derivative of Coulomb integral of the b th atom C_b [Brown and Simas, 1982; Kang and Jhon, 1982; Cartier and Rivail, 1987; Nagle, 1990; Langenaeker and Liu, 2001]:

$$\pi_{ab} = \frac{\partial q_a}{\partial C_b}$$

Descriptors of atom polarizability are among the $\rightarrow$ *electric polarization descriptors*.

DFT-based descriptors are those derived from the Density Functional Theory and the most important are chemical potential, molecular electronegativity, hardness and softness indices, and Fukui functions.

Papers about DFT and/or DFT descriptors are [Boon, De Proft *et al.*, 1998; Pérez and Contreras, 1998; Chang, Jalbout *et al.*, 2003; Parthasarathi, Subramanian *et al.*, 2003; Arulmozhiraja and Morita, 2004; Deka, Roy *et al.*, 2004; Meneses, Tiznado *et al.*, 2004; Wan, Zhang *et al.*, 2004; Morell, Grand *et al.*, 2005; Senthilkumar and Kolandaivel, 2005; Simón-Manso, 2005; Zhai, Wang *et al.*, 2005; Chatterjee, Balaji *et al.*, 2006; Cavalli, Carloni *et al.*, 2007].

- **electronic chemical potential**

For a molecule of N_{el} electrons, the electronic chemical potential μ is defined as

$$\mu = \left(\frac{\partial E}{\partial N_{\text{el}}} \right)_{v(\mathbf{r})}$$

where E is the ground-state energy and $v(\mathbf{r})$ is the $\rightarrow$ *composite nuclear potential* at the point $\mathbf{r}$.

It is a measure of the escaping tendency of an electronic cloud and corresponds to the negative of the electronegativity χ . In effect, as defined by [Iczkowski and Margrave, 1961], **molecular electronegativity** χ is

$$\chi = - \left(\frac{\partial E}{\partial N_{\text{el}}} \right)_{v(\mathbf{r})} = -\mu$$

where μ is the electronic chemical potential. This definition is restricted to the ground state and results in the same value of electronegativity (and electronic chemical potential) everywhere, for molecule, atoms, solid or molecular regions.

Considering a finite difference approximation and a quadratic dependence of energy on the number of electrons, the electronegativity approximates the loss or gain in charge by the $\rightarrow$ *ionization potential* IP and $\rightarrow$ *electron affinity* EA, that is,

$$\chi^{\text{MU}} \approx \frac{\text{IP} + \text{EA}}{2} = -\mu$$

which corresponds to the common definition of Mulliken electronegativity.

The **orbital electronegativity** χ_v of the μ th atomic orbital is an atomic descriptor defined as

$$\chi_\mu = - \left(\frac{\partial E}{\partial n_\mu} \right) = \left(\frac{\partial E}{\partial q_\mu} \right)$$

where n_μ and q_μ are the occupation number and the partial atomic charge of the μ th atomic orbital.

- **hardness indices**

The second derivative of the energy with respect to the number of electrons is called **absolute hardness** η (or **chemical hardness**) [Parr and Pearson, 1983], which for a molecule with N_{el} electrons is defined as

$$\eta = \frac{1}{2} \left(\frac{\partial^2 E}{\partial N_{\text{el}}^2} \right)_{v(\mathbf{r})} = \frac{1}{2} \left(\frac{\partial \mu}{\partial N_{\text{el}}} \right)_{v(\mathbf{r})} = \int h(\mathbf{r}) d\mathbf{r} = \frac{1}{2 \cdot S}$$

where μ is the electronic chemical potential, $h(\mathbf{r})$ is called **local hardness** (or **hardness density**), and S is the *total softness* (defined below).

From the frontier orbital energies, an approximated absolute hardness is also obtained. Under a finite difference approximation and a quadratic dependence of the energy on the number of electrons, absolute hardness is defined as

$$\eta \simeq \frac{\text{IP} - \text{EA}}{2} = \frac{\text{GAP}}{2} \equiv \frac{\epsilon_{\text{LUMO}} - \epsilon_{\text{HOMO}}}{2}$$

where IP and EA are the $\rightarrow$ *ionization potential* and the $\rightarrow$ *electron affinity*, respectively, and GAP denotes the $\rightarrow$ *HOMO–LUMO energy gap*. ϵ_{LUMO} and ϵ_{HOMO} are, respectively, the energies of the lowest unoccupied molecular orbital and the highest occupied molecular orbital. Such hardness is that within the framework of Koopmans' theorem; high values of hardness are related to the stability of a molecule as well as the LUMO–HOMO energy gap (see below).

The **η – χ diagram** is a scatter plot of absolute values of the hardness η against absolute values of the molecular electronegativity χ [Kobayashi, Shinohara *et al.*, 2006]. It was proposed as a QSAR tool to highlight groups of molecules with different electronic structures and, accordingly, different reactivities.

Both chemical potential μ and hardness η participate to the energy change ΔE due to the charge transfer ΔN :

$$\Delta E = \mu \cdot \Delta N + \frac{1}{2} \eta \cdot \Delta N^2$$

If $\eta > 0$ and $\Delta E < 0$, the charge transfer process is energetically favorable and the quantity

$$\omega = \frac{\mu^2}{2 \cdot \eta}$$

has been proposed as a measure of the electrophilic power of the molecule and called **electrophilicity index** [Parr, Szentpály *et al.*, 1999; Chattaraj and Roy, 2007]. Depending on the state from which the energy charge transfer is evaluated, the two quantities, ω_{gs} for the ground state and ω_{vs} for the valence state, have been defined and used in QSAR modeling:

$$\omega_{\text{gs}} = \frac{1}{8} \frac{(\text{IP} + \text{EA})^2}{(\text{IP} - \text{EA})} \quad \omega_{\text{vs}} = \frac{1}{4} \frac{(\text{IP} + \text{EA})^2}{(\text{IP} - \text{EA})}$$

where IP and EA are the $\rightarrow$ *ionization potential* and the $\rightarrow$ *electron affinity*, respectively.

The **activation hardness** is a measure of dynamic reactivity ($\rightarrow$ *dynamic reactivity indices*) obtained as the difference between absolute hardness of reactant (R) and transition state (T):

$$\Delta\eta = \eta_{\text{R}} - \eta_{\text{T}}$$

The activation hardness is sensitive to the reactivity at different molecule sites and orientation effects.

• softness indices

The **total softness** S is defined as

$$S = \left(\frac{\partial N_{\text{el}}}{\partial \mu} \right)_{v(\mathbf{r})} = \int s(\mathbf{r}) d\mathbf{r} = \frac{1}{2\eta}$$

where η is the $\rightarrow$ *absolute hardness*, μ the $\rightarrow$ *electronic chemical potential*, and $s(\mathbf{r})$ the **local softness** (or **softness density**) that is related to the Fukui function $f(\mathbf{r})$, via total softness, by the relationship

$$s(\mathbf{r}) = S \cdot f(\mathbf{r})$$

indicating that the Fukui function distributes the total softness among different regions of space.

Under a finite difference approximation and a quadratic dependence of energy on the number of electrons, the total softness can also be calculated as

$$S \simeq \frac{1}{IP-EA}$$

where IP and EA are the $\rightarrow$ ionization potential and the $\rightarrow$ electron affinity, respectively.

In terms of local softness, three local softness indices centered on the a th atom (s_a^+ , s_a^- , and s_a^0) were defined as [Roy, 2004b]

$$\begin{aligned} s_a^+(\mathbf{r}) &= [p_a(N+1) - p_a(N)] \cdot S = f_a^+ \cdot S \\ s_a^-(\mathbf{r}) &= [p_a(N) - p_a(N-1)] \cdot S = f_a^- \cdot S \\ s_a^0(\mathbf{r}) &= [p_a(N+1) - p_a(N-1)] \cdot S = f_a^0 \cdot S \end{aligned}$$

where S is the total softness and $p_a(N)$, $p_a(N+1)$, and $p_a(N-1)$ represent the condensed electronic populations of atom a for neutral, anionic, and cationic systems, respectively. So, s_a^+ , s_a^- , and s_a^0 represent the condensed local softness values indicating that atom a is more susceptible toward attack by a nucleophile, electrophile, and a radical on it.

• EIM descriptors

The EIM descriptors are $\rightarrow$ electronic descriptors based on critical values for energy separation involving frontier orbitals and descriptors derived from the electronic density of states in the framework of the **Electronic Indices Methodology** (EIM) [Barone, Camilo Jr. *et al.*, 1996; Braga, Barone *et al.*, 1999; Vendrame, Braga *et al.*, 1999, 2001; Vendrame and Takahata, 1999; Braga and Galvão, 2003, 2004]. This methodology uses simple Boolean rules based on electronic descriptors and has been applied in QSAR to classification of active/inactive compounds [Vendrame, Coluci *et al.*, 2002; Vendrame, Ferreira *et al.*, 2002; Coluci, Vendrame *et al.*, 2002].

The electronic **Density Of States (DOS)** is defined as the number of electronic states per energy unit. The related concept of **Local Density Of States (LDOS)**, that is, the DOS calculated over a specific molecular region, was introduced to also describe the spatial distribution of the electronic states over the system under consideration [Barone, Braga *et al.*, 2000; Santos, Contreras *et al.*, 2002; Santos, Chamorro *et al.*, 2004]. For the LDOS calculations, the contribution of each atom to an electronic level is weighed by the square of the (real) molecular orbital coefficient, that is, by the probability density corresponding to the level in that site [Vendrame, Coluci *et al.*, 2002]. The summation is carried over the selected atomic orbitals (n_i to n_j), leading to the following expression:

$$LDOS(E_i) = 2 \cdot \sum_{m=n_i}^{n_j} |c_{mi}|^2$$

The factor 2 comes from the Pauli exclusion principle maximum of two electrons per electronic level. This results in a discrete modulation that allows a direct comparison of DOS and LDOS obtained by any method for the calculation of linear combination of atomic orbitals (LCAO).

Moreover, based on the same principles, the relative contribution difference between the most relevant molecular electronic levels and the identified molecular region most responsible

for the biological activity is defined as [Braga, Vendrame *et al.*, 2000]

$$\eta = 2 \cdot \sum_{m=n_i}^{n_f} (|c_{m\text{Level}1}|^2 - |c_{m\text{Level}2}|^2)$$

where, for example, Level 1 and Level 2 may be HOMO and its adjacent lower level HOMO-1.

In carcinogenic studies on aromatic compounds, the ring with the highest bond order was suggested as the most informative region for applying LDOS approach [Barone, Braga *et al.*, 2000].

• Fukui functions

The Fukui functions $f(\mathbf{r})$ are local electronic descriptors of reactivity that find their origin within Density Functional Theory (DFT) and are defined as [Fukui, 1982; Parr and Yang, 1989]

$$f(\mathbf{r}) = \left(\frac{\partial \rho(\mathbf{r})}{\partial N_{\text{el}}} \right)_{v(\mathbf{r})} = \left(\frac{\partial \mu}{\partial v(\mathbf{r})} \right)_{N_{\text{el}}}$$

that is, as the first derivative of the $\rightarrow$ *electronic density* $\rho(\mathbf{r})$, at a point $\mathbf{r}$, with respect to the number of electrons N_{el} of the system, at a given external potential $v(\mathbf{r})$. The Fukui function corresponds to the first derivative of the $\rightarrow$ *electronic chemical potential* μ with respect to the external potential $v(\mathbf{r})$ for a given number of electrons. The Fukui function indicates the regions in a molecule where the charge density changes during a reaction, and measures how sensitive a chemical potential is to external perturbation at a specific point.

Because of the discontinuity of this derivative, a **backward Fukui function** $f^-(\mathbf{r})$ and a **forward Fukui function** $f^+(\mathbf{r})$ are defined, corresponding to local descriptors for electrophilic and nucleophilic attack, respectively. In terms of the finite difference approximation, both functions can be written as

$$f^-(\mathbf{r}) = \rho_N(\mathbf{r}) - \rho_{N-1}(\mathbf{r}) \quad f^+(\mathbf{r}) = \rho_{N+1}(\mathbf{r}) - \rho_N(\mathbf{r})$$

where $\rho_N(\mathbf{r})$, $\rho_{N+1}(\mathbf{r})$, and $\rho_{N-1}(\mathbf{r})$ are the electron densities of the N_{el} , $N_{\text{el}} + 1$, and $N_{\text{el}} - 1$ electron systems, respectively, all calculated at the same external potential $v(\mathbf{r})$ of the N_{el} electron system.

To define a reactivity index for radical attack, an **average Fukui function** $f^0(\mathbf{r})$ is also defined as

$$f^0(\mathbf{r}) = \frac{f^+(\mathbf{r}) + f^-(\mathbf{r})}{2} = \frac{\rho_{N+1}(\mathbf{r}) + \rho_{N-1}(\mathbf{r})}{2}$$

In practice, in the case of two different sites with similar disposition for reacting with a given reagent, the reagent prefers the one that is associated with the maximum response of the electronic chemical potential of the system.

Using a population analysis method, an atom-localized version of the Fukui functions is defined as [Yang and Mortier, 1986]

$$\begin{aligned} f_a^- &= q_a(N_{\text{el}}) - q_a(N_{\text{el}} - 1) \\ f_a^+ &= q_a(N_{\text{el}} + 1) - q_a(N_{\text{el}}) \\ f_a^0 &= \frac{q_a(N_{\text{el}} + 1) - q_a(N_{\text{el}} - 1)}{2} \end{aligned}$$

where $q_a(N_{el})$, $q_a(N_{el} + 1)$, and $q_a(N_{el} - 1)$ are the charges on the a th atom in the N_{el} , $N_{el} + 1$, and $N_{el} - 1$ electron systems.

A **nucleophilicity–electrophilicity index** was defined as [Morell, Grand *et al.*, 2005]

$$\Delta f(\mathbf{r}) = f^+(\mathbf{r}) - f^-(\mathbf{r}) = \rho_{LUMO}(\mathbf{r}) - \rho_{HOMO}(\mathbf{r}) \quad -1 \leq \Delta f(\mathbf{r}) \leq +1$$

If $\Delta f(\mathbf{r}) > 0$, then the site is favored for a nucleophilic attack, whereas if $\Delta f(\mathbf{r}) < 0$, then the site could hardly be susceptible to undertake a nucleophilic attack but it may be favored for an electrophilic attack.

Two other reactivity descriptors called **relative electrophilicity** RE and **relative nucleophilicity** RN were defined as [Roy, 2004b]

$$RE = \frac{f_a^+}{f_a^-} \quad RN = \frac{f_a^-}{f_a^+}$$

which represent the most preferred atom to be attached by a nucleophile and an electrophile, respectively.

Other quantum-chemical descriptors are $\rightarrow$ *TAE descriptors* based on the Bader's quantum theory of $\rightarrow$ *Atoms In Molecules* (AIM).

📖 Additional references are collected in the thematic bibliography (see Introduction).

■ quantum similarity

Quantum similarity (or **molecular quantum similarity**) is an approach to the analysis of similarity/diversity of molecules, mostly based on $\rightarrow$ *quantum-chemical descriptors*.

For molecules described by properties distributed in the molecular space such as a lattice, **Molecular Quantum Similarity Measures** (MQSM) were proposed to quantify the comparison between the fields representing two previously superimposed molecules [Carbó and Calabuig, 1992c, 1992a, 1992b; Besalú, Carbó *et al.*, 1995; Carbó, Besalú *et al.*, 1995; Carbó-Dorca and Besalú, 1996, 1998; Lobato, Amat *et al.*, 1997, 1998; Robert and Carbó-Dorca, 1998a; Robert, Amat *et al.*, 1999].

In general, a MQSM between the two molecules s and t is defined as

$$s_{st} = \iint \rho_s(\mathbf{r}_1) \Omega(\mathbf{r}_1, \mathbf{r}_2) \rho_t(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2$$

where ρ_s and ρ_t are density functions for molecules s and t , $\mathbf{r}_1$ and $\mathbf{r}_2$ are two points in the space, and $\Omega(\mathbf{r}_1, \mathbf{r}_2)$ is a definite positive operator. The most commonly encountered operators are

- (1) $\Omega(\mathbf{r}_1, \mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$ *overlap-like MQSM*
- (2) $\Omega(\mathbf{r}_1, \mathbf{r}_2) = |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ *Coulomb-like MQSM*
- (3) $\Omega(\mathbf{r}_1, \mathbf{r}_2) = |\mathbf{r}_1 - \mathbf{r}_2|^{-2}$ *gravitational-like MQSM*
- (4) $\Omega(\mathbf{r}_1, \mathbf{r}_2) = \rho_C(\mathbf{r}_1) \delta(\mathbf{r}_1 - \mathbf{r}_2)$ *triple-density MQSM*

where δ is the delta Dirac function. To calculate MQSM, appropriate $\rightarrow$ *alignment rules* are required.

By different mathematical transformations, **Molecular Quantum Similarity Indices** (MQSI) are derived from molecular quantum similarity measures. They are divided into two main classes: C-class indices, referred to as *correlation-like indices* ranging from 0 (maximum

dissimilarity) to 1 (maximum similarity), and D-class indices, referred to as *distance-like indices* ranging from 0 (maximum similarity) to infinite (maximum dissimilarity). C-class indices s_{st} can be transformed into D-class indices d_{st} by the following:

$$d_{st} = \sqrt{1 - s_{st}^2}$$

It can be noted that distances calculated in this way range from zero to one.

Based on the molecular quantum similarity measures, **Molecular Quantum Self-Similarity Measures** (MQS-SM) were proposed as molecular descriptors calculated by comparing each molecule with itself and all the others, and using appropriate Hermitian operators Ω associated to each molecular property [Ponec, Amat *et al.*, 1999].

The most popular molecular quantum similarity indices based on the overlap-like MQSM operator are reported below, together with other proposed similarity indices of distributed properties [Carbó and Calabuig, 1990; Good, Peterson *et al.*, 1993; Constans and Carbó, 1995; Good, 1995; Carbó *et al.*, 1996; Carbó, Besalú *et al.*, 1996; Good and Richards, 1998]. They are expressed in a form suitable for a discrete molecular space represented by a $\rightarrow$ grid of scalar values; N is the total number of grid points, P_{sk} is the property value (density function ρ) for the s th molecule in the k th grid point.

- **Carbó similarity index (C)**

Initially proposed to compare molecules in terms of their $\rightarrow$ electron density, it can be applied to compare any property between molecules s and t [Carbó *et al.*, 1980; Carbó, Leyda *et al.*, 1980; Carbó and Domingo, 1987]. It is defined as

$$C_{st} = \frac{\sum_{k=1}^N P_{sk} \cdot P_{tk}}{\left(\sum_{k=1}^N P_{sk}^2\right)^{1/2} \cdot \left(\sum_{k=1}^N P_{tk}^2\right)^{1/2}}$$

The Carbó index is sensitive to the shape of the property distributions, but not to their magnitude.

- **Hodgkin similarity index (H)**

With respect to the Carbó index, it is less sensitive to the shape of the property distribution but more sensitive to its magnitude [Hodgkin and Richards, 1987, 1988]. It is defined as

$$H_{st} = \frac{2 \cdot \sum_{k=1}^N P_{sk} \cdot P_{tk}}{\sum_{k=1}^N P_{sk}^2 + \sum_{k=1}^N P_{tk}^2} = 1 - \frac{d_{st}^2}{\sum_{k=1}^N P_{sk}^2 + \sum_{k=1}^N P_{tk}^2}$$

where d_{st}^2 is the square $\rightarrow$ Euclidean distance between the molecules s and t . This index is particularly important for calculating the $\rightarrow$ molecular electrostatic potential (MEP) and the molecular electric field (MEF) similarity because these properties may be of similar shape for a pair of molecules, while their absolute values are significantly different.

- **linear similarity index (L)**

A measure of the similarity between molecules s and t defined as [Good, 1992]

$$L_{st} = \frac{1}{N} \cdot \sum_{k=1}^N \left(1 - \frac{|P_{sk} - P_{tk}|}{\max(|P_{sk}|, |P_{tk}|)} \right)$$

- **exponential similarity index (E)**

A measure of the similarity between molecules s and t defined as [Good, 1992]

$$E_{st} = \frac{1}{N} \cdot \sum_{k=1}^N \exp^{-d_{st,k}}$$

where $d_{st,k}$ is a distance measure between the molecules s and t at the k th grid point, defined as

$$d_{st,k} = \frac{|P_{sk} - P_{tk}|}{\max(|P_{sk}|, |P_{tk}|)}$$

- **Meyer–Richards similarity index (S)**

A modified version of the Carbó similarity index [Meyer and Richards, 1991], defined as

$$S_{st} = \frac{N_{st}}{(N_s \cdot N_t)^{1/2}}$$

where N_{st} is the number of grid points falling inside both molecules and N_s and N_t are the total number of grid points falling inside each individual molecule.

- **similarity score (A_F)**

A similarity measure between two molecules s and t in any relative orientation to each other, used in CoMSIA ($\rightarrow$ comparative molecular similarity analysis) and defined as [Kearsley and Smith, 1990]

$$A_F = - \sum_{i=1}^{A_s} \sum_{j=1}^{A_t} w_{ij} \cdot e^{-a \cdot r_{ij}^2}$$

where A_s and A_t are the number of atoms of the two molecules, the parameter a defines the distance dependence, r_{ij} is the interatomic distance between atoms i and j , w_{ij} is the total contribution due to the property values of the two considered atoms, defined as

$$w_{ij} = w_E \cdot q_i \cdot q_j + w_S \cdot V_i^{vdw} \cdot V_j^{vdw} + \dots$$

where $w_E, w_S, \dots$ are user-defined values to give different weights to electrostatic (E), steric (S), hydrophobic and hydrogen-bonding properties; q and V^{vdw} are partial charges and van der Waals volumes of the atoms, respectively [Kubinyi, Hamprecht *et al.*, 1998].

- **Tanimoto coefficient (T)**

Derived from the $\rightarrow$ Jaccard/Tanimoto coefficient for continuous variables and applied to 3D distributed properties in the form

$$T_{st} = \frac{\sum_{k=1}^N P_{sk} \cdot P_{tk}}{\sum_{k=1}^N P_{sk}^2 + \sum_{k=1}^N P_{tk}^2 - \sum_{k=1}^N P_{sk} \cdot P_{tk}}$$

A modified Tanimoto coefficient was also proposed to measure the degree of relatedness of the shapes of two structures [Hahn, 1997].

📖 [Carbó-Dorca and Mezey, 1998; Gironés *et al.*, 2000; Gironés, Gallegos *et al.*, 2000; O'Brien and Popelier, 2001; Podlipnik and Koller, 2001; Gironés and Carbó-Dorca, 2002b; Gironés *et al.*, 2002; Gironés, Amat *et al.*, 2002; Bultinck and Carbó-Dorca, 2003; Boon, Van Alsenoy *et al.*, 2005; Gallegos Saliner and Gironés, 2005]

➤ **quartile deviation** → statistical indices (☉ indices of dispersion)

■ **QuaSAR descriptors** ($\equiv$ *MOE descriptors*)

These are molecular descriptors calculated via the QuaSAR descriptors module present in the MOE package [MOE – Chemical Computing Group, Inc., 1999]. MOE (*Molecular Operating Environment*) is a software that allows not only the calculation of several well-known molecular descriptors, but also definition of custom descriptors using MOE's built-in Scientific Vector Language (SVL). MOE also contains a semiempirical force field as well as tools for multivariate analysis such as → *Principal Component Analysis*, classification, clustering, filtering and predictive model construction methods for both biological activities and → *ADME properties*.

QuaSAR descriptors [QuaSAR – Chemical Computing Group, Inc., 2007] include several types of traditional molecular descriptors: Kier–Hall → *connectivity indices*, → *structural keys*, → *E-state indices*, descriptors of → *physico-chemical properties* (such as log *P*, molecular weight and molar refractivity), 3D molecular features (such as potential energy descriptors, surface area, volume and → *shape descriptors*, conformation dependent partial charge descriptors), and some → *pharmacophore-based descriptors*.

📖 [Labute, 2000; Xue, Godden *et al.*, 2000; Xue and Bajorath, 2000; Mazza, Sukumar *et al.*, 2001; Deretey, Feher *et al.*, 2002; Yuan and Parrill, 2002; Byvatov, Fechner *et al.*, 2003; Tugcu, Song *et al.*, 2003; Tugcu, Ladiwala *et al.*, 2003; Kovatcheva, Golbraikh *et al.*, 2004; Schefzik, Kibbey *et al.*, 2004; Xiao, Varma *et al.*, 2004; Kriegl, Arnhold *et al.*, 2005a; Kriegl, Arnhold *et al.*, 2005b; de Cerqueira Lima, Golbraikh *et al.*, 2006; Givchchi, Bender *et al.*, 2006; Vedani, Dobler *et al.*, 2005; Prathipati and Saxena, 2006; Wegner *et al.*, 2006; Wegner, Fröhlich *et al.*, 2006]

- **quasi-Euclidean distances** → Laplacian matrix
- **quasi-Euclidean matrix** → Laplacian matrix
- **quasi-length carbon chain** → electric polarization descriptors (☉ Polarizability Effect Index)
- **quasi-Wiener index** → Laplacian matrix
- **QUIK rule** → validation techniques
- **quotient Balaban index of the first kind** → distance matrix
- **quotient Balaban index of the second kind** → distance matrix
- **quotient map matrix** → biodescriptors (☉ proteomics maps)
- **quotient matrices** → matrices of molecules

R

- **radial centric information index** → centric indices
- **radial distribution function** → molecular transforms
- **radial distribution function descriptors** $\equiv$ *RDF descriptors* → molecular transforms
- **radical superdelocalizability** → quantum-chemical descriptors
- **radius-corrected connectivity index** → connectivity indices (⊙ Kupchik modified connectivity indices)
- **radius of gyration** → size descriptors
- **radius-diameter diagram** → shape descriptors (⊙ Petitjean shape indices)
- **Raeviski H-bond indices** → hydrogen-bonding descriptors
- **RAI** $\equiv$ *Relative Alkylation Index*
- **ramification index** → vertex degree
- **ramification pair indices** → vertex degree
- **Randić atomic path code** $\equiv$ *vertex path code* → path counts
- **Randić chirality index** → chirality descriptors
- **Randić connectivity ID number** → ID numbers
- **Randić connectivity index** → connectivity indices
- **Randić-like indices** → connectivity indices
- **Randić–Plavsic complexity index** → molecular complexity
- **Randić–Razinger index** → connectivity indices (⊙ walk connectivity indices)
- **range** → statistical indices (⊙ indices of dispersion)
- **random walk** $\equiv$ *walk* → graph
- **random walk count** → walk counts
- **random walk Markov matrix** → walk counts
- **random walk matrices** $\equiv$ *walk matrices*
- **rank distance** $\equiv$ *Rouvray index* → distance matrix
- **rank distance** → similarity/diversity
- **ranking methods** → chemometrics
- **Rao similarity coefficient** $\equiv$ *Russell–Rao similarity coefficient* → similarity/diversity (⊙ Table S9)
- **Rashevsky complexity index** → molecular complexity
- **RC index** → delocalization degree indices
- **R-connectivity index** → GETAWAY descriptors
- **RDCHI index** → distance matrix
- **RDF descriptors** → molecular transforms

- **RDSQ index** → distance matrix
- **RDSUM index** → distance matrix
- **RDSUM index** → Harary indices
- **reaction constant** → Hammett equation
- **reaction MOLMAPs** → MOLMAP descriptors

■ reaction rate coefficient

The rate coefficient or rate constant of a chemical reaction is the coefficient that precedes reactant concentrations in a simple rate equation.

For a chemical reaction $nA + mB \rightarrow C + D$, with a rate equation

$$r = k(T) \cdot [A]^n \cdot [B]^m$$

where $k(T)$ is the reaction rate coefficient or the reaction rate constant and T the temperature of the reaction.

Rate coefficient includes all factors that affect reaction rate, except for concentration, which is explicitly accounted for. Rate coefficient is therefore not constant; because of that reason the name *reaction rate coefficient* is preferred over *reaction rate constant*. The rate coefficient is mainly affected by temperature as described by Arrhenius equation but also, ionic strength, surface area of the adsorbent (for heterogeneous reactions), light irradiation, and other → *physico-chemical properties*, depending on the considered reaction.

■ reactivity indices

Reactivity indices are molecular descriptors encoding information about the behavior of molecules in chemical reactions and are usually categorized as either **electrophilic indices** or **nucleophilic indices**, depending on whether the reaction of interest involves electrophilic or nucleophilic attack. Moreover, **static reactivity indices**, such as charges, describe isolated molecules in their ground state, while **dynamic reactivity indices** refer to molecules in their transition states during a reaction.

→ *Charge descriptors* and several → *quantum-chemical descriptors* such as → *molecular orbital energies* and → *superdelocalizability indices* are examples of reactivity indices. → *Reaction MOLPMAPs* are descriptors specifically designed to describe chemical reactions.

📖 [Lall, 1981c; Jug, 1984; Simon, Ciubotariu *et al.*, 1985; Gasteiger, Hutchings *et al.*, 1987; Gasteiger, Röse *et al.*, 1988; Jiang and Zhang, 1990; Rabinowitz and Little, 1991; Mekenyan and Basak, 1994; Mekenyan and Veith, 1994; Balaban, 1994b; Geerlings, Langenaeker *et al.*, 1996; Bakken and Jurs, 1999a]

- **recall** $\equiv$ *true positive rate* → classification parameters
- **Receiver Operator Characteristic curve** → classification parameters
- **receptor** → drug design
- **receptor cavity** $\equiv$ *binding site cavity* → drug design
- **receptor mapping** → drug design
- **receptor mapping techniques** → structure/response correlations
- **reciprocal Barysz distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **reciprocal Cluj matrices** → Cluj matrices

- **reciprocal complementary distance matrix** → distance matrix
- **reciprocal complement Barysz distance matrix** → weighted matrices ($\odot$ weighted distance matrices)
- **reciprocal detour-distance combined matrix** → detour matrix
- **reciprocal detour matrix** → detour matrix
- **reciprocal distance complement matrix** → distance matrix
- **reciprocal distance-detour combined matrix** → detour matrix
- **reciprocal distance matrix** → distance matrix
- **reciprocal distance-path matrix** → distance-path matrix
- **reciprocal distance polynomial** → characteristic polynomial-based descriptors
- **reciprocal distance sum** → distance matrix
- **reciprocal edge distance matrix** → edge distance matrix
- **reciprocal geometry matrix** → geometry matrix
- **reciprocal matrices** → matrices of molecules
- **reciprocal resistance matrix** $\equiv$ *conductance matrix* → resistance matrix
- **reciprocal reverse Wiener matrix** → distance matrix
- **reciprocal Schultz indices** → Schultz molecular topological index
- **reciprocal spanning-tree density** → Laplacian matrix
- **reciprocal square distance matrix** → distance matrix
- **reciprocal Szeged matrices** → Szeged matrices
- **reciprocal topographic matrix** → molecular geometry
- **reciprocal walk matrix** → walk matrices
- **reciprocal Wiener matrix** → Wiener matrix
- **RDI** $\equiv$ *ring degree-distance index* → Cao–Yuan indices
- **reduced graph** → molecular graph
- **redundancy index** $\equiv$ *Brillouin redundancy index* → information content
- **redundant information content** → indices of neighborhood symmetry
- **reentrant surface** → molecular surface ($\odot$ solvent-accessible molecular surface)
- **reference compound** → drug design
- **reference polynomial** $\equiv$ *matching polynomial* → Hosoya Z-index
- **reference structure** $\equiv$ *reference compound* → drug design
- **refractive index** → physico-chemical properties
- **refractive index function** → physico-chemical properties ($\odot$ refractive index)
- **regression analysis** → chemometrics

■ regression parameters

→ *Statistical indices* used for the evaluation of the performance of regression models. They are derived from two kinds of statistics, called *goodness of fit* and *goodness of prediction*. The former pays more attention to the fitting ability of models, while the latter to the prediction power of models, it being based on → *validation techniques*.

The **goodness of fit** statistic measures how well a model fits the data of the → *training set*, for example, how well a regression model (or a classification model) accounts for the variance of the response variable.

The most important indices are listed below. The quantity df_E refers to the degrees of freedom of the error, that is, to $n - p'$, where n is the number of objects (samples) and p' the number of model parameters (for example, $p' = p + 1$ for a linear regression model with p variables plus

intercept). df_M and df_T refer to the degrees of freedom of the model and the total degrees of freedom, respectively.

Residual sum of squares (RSS) ($\equiv$ error sum of squares). The sum of square differences between the observed (y) and estimated response ($\hat{y}$) over all the sample objects

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

This quantity is minimized by the least square estimator. A complementary quantity is the **Model sum of squares, MSS**, defined as the sum of the square differences between the estimated response and the average observed response:

$$MSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

This is a part of the total variance explained by the regression model as opposed to the residual sum of squares RSS . Moreover, a reference quantity is the **total sum of squares, TSS** (also denoted by SSY), defined as the sum of the square differences between the observed response and the average observed response:

$$TSS \equiv SSY = \sum_{i=1}^n (y_i - \bar{y})^2$$

TSS represents the total variability of the response that a regression model has to explain and is used as a reference quantity to calculate standard quality parameters such as the coefficient of determination. For a given data set of n samples, TSS has a constant value. The values of TSS , MSS , and RSS depend on the unit of measure of the y response and are related by the following equation:

$$TSS = MSS + RSS$$

Coefficient of determination (R^2). The square multiple correlation coefficient that is the part of total variance of the response explained by a regression model and is often expressed as percentage. It can be calculated from the model sum of squares MSS or the residual sum of squares RSS as

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad 0 \leq R^2 \leq 1$$

where TSS is the total sum of squares around the mean. A value of one indicates perfect fit, or, in other words, a model with zero error.

A related quantity is the **multiple correlation coefficient R** defined as the squared root of R^2 . It is a measure of linear association between the observed response and the estimated response.

Note. By definition, the R values are always greater (or equal) than zero. In some papers, for univariate regression models, that is, $y = b_0 - b_1x$, a negative value of R is reported: this is formally wrong and due to the confusion between the multiple correlation coefficient R and the

bivariate correlation r_{xy} between the pairs of variables x – y , which is characterized by the signed regression coefficient b_1 .

A quantity complementary to R^2 is the **coefficient of nondetermination** (or **coefficient of alienation**) defined as

$$cnd = 1 - R^2$$

Mean square error, MSE (or **residual mean square, s^2** , or **expected square error**). The estimate s^2 of the error variance σ^2 , defined as

$$MSE \equiv s^2 = \frac{RSS}{df_E} \quad 0 \leq s^2 < \infty$$

where RSS is the residual sum of squares and df_E indicates the error degrees of freedom ($n - p'$).

The squared root of the residual mean square is called **residual standard deviation, RSD** or **standard error of estimate, SE** or s (or **error standard deviation**, or **residual standard error**). It is an estimate of the model error σ , defined as

$$SE \equiv RSD \equiv s = \sqrt{\frac{RSS}{df_E}} \quad 0 \leq s < \infty$$

where RSS is the residual sum of squares and df_E indicates the error degrees of freedom ($n - p'$).

Root mean square error (RMSE) (or **root mean square deviation, RMSD**). Also known as **standard error in calculation (SEC)** or **standard deviation error in calculation (SDEC)**, it is a function of the residual sum of squares:

$$RMSE \equiv RMSD \equiv SDEC \equiv SEC = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{RSS}{n}} \quad 0 \leq RMSE < \infty$$

where RSS is the residual sum of squares and n the number of samples.

Normalized root mean square error, denoted as **NRMSE** (or **normalized root mean square deviation, NRMSD**), is defined dividing $RMSE$ (or $RMSD$) by the range of the response variable:

$$NRMSE \equiv NMRSD = \frac{RMSE}{y_{\max} - y_{\min}}$$

Note. The terms *mean square error*, *root mean square error*, *mean square deviation*, *root mean square deviation*, and *expected square error* are also often defined using the word “squared” instead of the word “square”.

Relative error in calculation. The relative error calculated from the root mean square error in calculation, defined as

$$REC\% = \frac{100}{\bar{y}} \cdot \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} = \frac{100}{\bar{y}} \cdot RMSE$$

where $\bar{y}$ is the average response.

F-ratio test. Among the most known statistical tests, the F-ratio test is defined as the ratio of the model sum of squares MSS over the residual sum of squares RSS , relatively to their

corresponding degrees of freedom:

$$F = \frac{MSS/df_M}{RSS/df_E} = \frac{MSS \cdot (n-p-1)}{RSS \cdot p} = \frac{(n-p-1) \cdot R^2}{p \cdot (1-R^2)}$$

The calculated F value is compared with the critical value F_{crit} provided the same degrees of freedom. Since the F test compares the model explained variance with the residual variance, high values of the F -ratio test should indicate reliable models.

Note. In spite of its very common use in regression analysis, the F -ratio test is a weak test for multivariate models. In effect, the null hypothesis of the F test states that all of the regression coefficients are zero, while the alternative hypothesis states that *at least one* regression coefficient is different from zero. However, in modern QSAR, multivariate models are usually produced where the variables should be *all* relevant. But, even very high F values mean only that at least one variable is significant and not that the whole model is significant. Validation tools are more suitable to give information about the relevance of all the variables in a model.

Adjusted R^2 (R^2_{adj}). A parameter adjusted for the degrees of freedom, so that it can be used for comparing models with different numbers of predictor variables:

$$R^2_{adj} = 1 - \frac{RSS/df_E}{TSS/df_T} = 1 - (1-R^2) \cdot \left(\frac{n-1}{n-p'} \right) \quad 0 \leq R^2_{adj} < 1$$

where RSS and TSS are the residual sum of squares and the total sum of squares, respectively; R^2 is the coefficient of determination.

Exner statistic (Ψ^2). Another parameter adjusted to the degrees of freedom, so that it can be used for comparing models with different numbers of predictor variables:

$$\Psi^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \cdot \frac{n}{n-p} = \frac{RSS}{TSS} \cdot \frac{n}{n-p} = (1-R^2) \cdot \frac{n}{n-p}$$

Kubinyi fitness function (FIT). A parameter used to compare models with a different number of p variables [Kubinyi, 1994a], defined as a modified $\rightarrow F$ -ratio test:

$$FIT = \frac{(n-p-1) \cdot R^2}{(n+p^2) \cdot (1-R^2)} = F \cdot \frac{p}{n+p^2}$$

where R^2 is the coefficient of determination. This function is used to perform $\rightarrow$ variable selection by using the method called $\rightarrow$ MUltation and Selection Uncover Models.

Several regression parameters are measures suitable for choosing models to predict a dependent variable y from a potentially large set of independent variables. The **generalized final prediction error criteria** (FPE criteria) include many known parameters and are defined as [Krieger and Zhang, 2006]

$$FPE(k, \alpha_{n,k}) = \frac{RSS_k}{RSS_p/(n-p')} + k \cdot \alpha_{n,k} = \frac{RSS_k}{RSS_p} \cdot (n-p') + k \cdot \alpha_{n,k} \quad \alpha_{n,k} \geq 0$$

where RSS_k is the residual sum of squares based on a model with k parameters, $RSS_p/(n-p')$ is the estimated $\hat{\sigma}^2$ of the full model with p' parameters, and $\alpha_{n,k}$ is the penalty factor for model complexity, which depends on the sample size n and/or the number of k fitted parameters.

There are many suggested choices for $\alpha_{n,k}$, some of them providing more parsimonious models and other more complex models. The most known FPE criteria are shown in Table R1.

Table R1 Final prediction errors derived from the generalized final prediction error criteria.

FPE criteria	Symbol	Parameters
Akaike information content	AIC	$\alpha_{n,k} = 2$
Mallows C_p	C_p	$\alpha_{n,k} = (2k - n)/k$
Schwarz–Bayesian information criterion	BIC	$\alpha_{n,k} = \log n$
Hannan–Quinn ϕ -criterion	ϕ	$\alpha_{n,k} = c \cdot \log \log n \quad c > 2$
George–Foster criterion	GFC	$\alpha_{n,k} = 2 \cdot \log k$
Hurvish–Tsai criterion	HTC	$\alpha_{n,k} = 2n/(n - k - 2)$

k , the number of model parameters; n , the number of samples.

It has been demonstrated that, although these parameters are all based on the goodness of fit RSS , they are related to the prediction ability of a model as it can be usually estimated from $\rightarrow$ *validation techniques*. For example, in the case of the v -fold cross-validation parameter $PRESS(k, v)$, where v is the v -dimensional subset of objects in turn deleted from the training set, $PRESS(k, v)$ is asymptotically equivalent, for $n \rightarrow \infty$, to FPE with

$$\alpha_{n,k} = \frac{2 \cdot n - v}{n - v}$$

Similar results can be obtained from the bootstrap method.

Among the final prediction error criteria, the most popular are detailed below together with some other criteria.

Akaike Information Criterion (AIC). A parameter used to choose among models with different parameters [Akaike, 1974] and defined as

$$AIC_p = -2 \cdot L_p + 2 \cdot p'$$

where p' is the number of model parameters and L_p is the maximized log-likelihood. For regression models, the optimal complexity according to the Akaike criterion is obtained by minimizing the following:

$$\min(AIC_p) = \frac{n + p' + 1}{n - p' - 1} \cdot s^2$$

where s^2 is the residual mean square.

Mallows C_p . Closely related to the adjusted R^2 , C_p statistic [Mallows, 1973] is used to compare regression models obtained with a small subset of variables with the full least squares regression model:

$$C_p = \frac{RSS_p}{s^2} + 2p' - n$$

where RSS_p is the residual sum of squares of the biased model with p' parameters and s^2 is the residual mean square of the full least squares model. If an equation with p' parameters is adequate, that is, does not suffer from lack of fit, the expected value (E) of the ratio RSS_p/s^2 is $n - p'$ and $E(C_p) = p'$. It follows that a plot of C_p versus p' show up the best models as points fairly

close to the $C_p = p'$ line, that is,

$$\min_p (|C_p - p'|)$$

In Ordinary Least Squares regression, $C_p = p'$.

J_p statistics. A function of the residual mean square s^2

$$J_p = \frac{(n + p') \cdot s^2}{n} = \frac{RSS}{n} \cdot \frac{(n + p')}{(n - p')}$$

where n is the total number of observations, p' the number of parameters in the model, and RSS the residual sum of squares.

S_p statistics. Also called *Hocking statistic*, it is a function of the residual mean square s^2 , defined as

$$S_p = \frac{s^2}{df_E} = \frac{RSS}{(n - p')^2}$$

where df_E is the error degrees of freedom and RSS the residual sum of squares.

Ant Colony fitness function. A parameter proposed for the ant colony optimization (ACO) algorithm, defined as [Shen, Jiang *et al.*, 2005]

$$F = -\log \left[\frac{RSS_p}{s_M^2} + 2 \cdot p \right]$$

where RSS is the residual sum of squares of a model obtained from p independent variables and s_M^2 is the residual mean square corresponding to the minimum number M of PLS latent variables when a further increase of the latent variables does not cause a significant reduction in RSS .

Note. A *quality factor* (Q) is often (mis)used to measure the quality of the QSAR regression models [Pogliani, 1994a]. It is defined as the ratio of the multiple correlation coefficient R over the residual standard deviation s , that is, $Q = R/s$. However, since s depends on the measure unit of the response according to RSS definition, Q ranges between zero and infinite and thus it should not be used as a quality measure of the regression models. Moreover, in several papers, the Q quality factor was improperly confounded with the predictive parameter Q , that is, the cross-validated R [Agrawal, Chaturvedi *et al.*, 2003; Khadikar, Singh *et al.*, 2003; Agrawal, Agrawal, Srivastava *et al.*, 2004; Jaiswal and Khadikar, 2004; Thakur, Thakur *et al.*, 2004a; Gupta *et al.*, 2005; Khadikar, Diudea *et al.*, 2006; Verma and Hansch, 2007]. In this quality factor, any information about the model predictive ability is lacking.

The **goodness of prediction** statistic measures how well a model can be used to estimate future (test) data, that is, how well a regression model (or a classification model) estimates the response variable given a set of values for predictor variables. This statistics is obtained using $\rightarrow$ *validation techniques*.

Besides some special parameters adopted in $\rightarrow$ *variable selection*, such as *Friedman's lack-of-fit function* (LOF), the most important indices of goodness of prediction are listed below.

Prediction error sum of squares (PRESS). The sum of the squared differences between the observed and estimated response by $\rightarrow$ *validation techniques* [Allen, 1971, 1974]:

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{i/i})^2$$

where the summation goes over the n objects and $\hat{y}_{i/i}$ denotes the response of the i th object estimated by using a model obtained without using the i th object. Use of validation techniques minimizes this quantity. The **uncertainty in the prediction** is given by

$$s_{PRESS} = \sqrt{\frac{PRESS}{n-p'}}$$

where p' is the number of parameters in the model and n the number of objects.

Cross-validated R^2 (R^2_{cv}) (or Q^2 or q^2). The explained variance in prediction:

$$R^2_{cv} \equiv Q^2 = 1 - \frac{PRESS}{TSS} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad Q^2 \leq 1$$

where $PRESS$ is the $\rightarrow$ *prediction error sum of squares* and TSS the $\rightarrow$ *total sum of squares*. It must be noted that Q^2 can assume negative values for regression models with poor prediction ability.

The quantities RSS and $PRESS$ from the leave-one-out validation procedure, as well as R^2 and Q^2_{LOO} (see $\rightarrow$ *asymptotic Q^2 rule*), are asymptotically related when the number of samples n tends to infinite, as [Miller, 1990a]

$$PRESS \approx RSS \cdot \left(\frac{n}{n-p'}\right)^2 \quad \text{if } n \rightarrow \infty \quad Q^2_{LOO} \approx Q^2_{ASYM} = 1 - (1 - R^2) \cdot \left(\frac{n}{n-p'}\right)^2 \quad \text{if } n \rightarrow \infty$$

These relationships highlight how the leave-one-out procedure may give an over-optimistic prediction ability when the number of samples is high enough. Then, adjusted R^2 can be viewed as an estimate of Q^2_{LOO} for an infinite number of samples:

$$Q^2_{LOO} \approx 1 - (1 - R^2) \cdot \left(\frac{n}{n-p'}\right)^2 \approx R^2_{adj} \cdot \left(\frac{n}{n-p'}\right)^2 \quad \text{if } n \rightarrow \infty$$

Predictive square error (PSE). The average value of the prediction error sum of squares:

$$PSE = \frac{PRESS}{n} = \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{n}$$

where $PRESS$ is the $\rightarrow$ *prediction error sum of squares* and n the number of objects.

Root mean square error in prediction (RMSEP) (or **root mean square deviation in prediction, RMSDP**). Also known as **standard error in prediction (SEP)** or **standard deviation error in prediction (SDEP)**, is a function of the prediction residual sum of squares $PRESS$, defined as

$$RMSEP \equiv RMSDP \equiv SEP \equiv SDEP = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{n}} = \sqrt{\frac{PRESS}{n}} = \sqrt{PSE}$$

where PSE is the predictive square error.

Relative error in prediction. The relative error calculated from the standard deviation error in prediction:

$$REP\% = \frac{100}{\bar{y}} \cdot \sqrt{\frac{\sum_{i=1}^n (\hat{y}_{i/i} - y_i)^2}{n}} = \frac{100}{\bar{y}} \cdot RMSEP$$

where $\bar{y}$ is the average observed response and $RMSEP$ the root mean square error in prediction.

Based on the definitions given above, several combined functions of these basic quantities were proposed to assess the predictive ability of regression models or to rank regression models obtained by some $\rightarrow$ *variable selection* procedure. Examples of these functions are listed below.

Jurs cost function. A function based on the $\rightarrow$ *mean square error* s^2 of both the training and test set [Sutter, Peterson *et al.*, 1997]:

$$Cost = s_{TR}^2 + 0.4 \cdot |s_{TR}^2 - s_{CV}^2| \quad \text{or} \quad Score = R^2 - 0.4 \cdot |R^2 - Q^2|$$

where the subscripts TR and CV stand for training and cross-validation set, respectively; R^2 and Q^2 are the coefficient of determination and the cross-validated R^2 , respectively.

Exponential cost function. A function to be minimized, defined as [Casalegno, Benfenati *et al.*, 2005]

$$ExpCost = \sum_{i=1}^n \exp(|\hat{y}_i - y_i| + 1) - \exp(1)$$

where the summation runs over all the training set objects and $\hat{y}_i$ denotes the calculated response of the i th object. The exponential form was introduced to maximally uniform the deviations of the calculated $\hat{y}$ values from the observed y values, the big differences being exponentially magnified.

Hou fitness function. A parameter which combines the $\rightarrow$ *multiple correlation coefficient* R and the leave-one-out Q [Hou, Wang *et al.*, 1999]:

$$R_P = R \cdot Q_{LOO}$$

Depczynski fitness function. A parameter based on the $\rightarrow$ *standard deviation error* both in calculation $SDEC$ and prediction $SDEP$ [Depczynski, Frost *et al.*, 2000]:

$$\eta = \sqrt{\frac{(n_C - p - 1) \cdot SDEC^2 + n_T \cdot SDEP^2}{n_C + n_T - p - 1}}$$

where n_C and n_T are the number of objects of the training and test sets, respectively, and p , the number of variables.

Tarko-Ivanciuc fitness function. A measure of predictive ability based on maximizing the product of the leave-one-out Q_{LOO}^2 and the leave-one-out $\rightarrow$ *Kendall rank correlation coefficient* τ_{LOO} [Tarko and Ivanciuc, 2001]:

$$F = Q_{LOO}^2 \cdot \tau_{LOO}$$

Daren fitness function. A parameter used in PLS approach [Daren, 2001], defined as

$$MQ^2 = \frac{1}{u} \cdot \left(1 + \frac{100 - M}{100 + M} \cdot d \right) \cdot Q_{LMO}^2$$

where M is the number of significant PLS latent variables and d and u two adjustable parameters (values of 0.317 and 1.298 were proposed, respectively, for three significant components). Q_{LMO}^2 is the Q^2 obtained by the leave-more-out validation technique.

Golbraikh–Tropsha statistics. These are a set of conditions that a model has to satisfy to be considered a predictive model [Golbraikh and Tropsha, 2002a]:

$$\begin{aligned} Q_{LOO}^2 &> 0.5 & \text{and} & R^2 > 0.6 \\ \frac{R^2 - R_0^2}{R^2} &< 0.1 & \text{or} & \frac{R^2 - R_0'^2}{R^2} < 0.1 \\ 0.85 \leq b \leq 1.15 & \text{or} & 0.85 \leq b' \leq 1.15 \end{aligned}$$

where R_0^2 is the coefficient of determination of predicted versus observed responses, $R_0'^2$ is the coefficient of determination of observed versus predicted responses, and b and b' the two corresponding slopes of the regression lines through the origin, being the slope for an ideal model equal to 1 and the intercept b equal to 0.

Asymptotic Q^2 rule. This rule is based on the comparison between the actual Q_{LOO}^2 and the Q_{ASYM}^2 values, and, specifically, regression models could be accepted if, provided an acceptable value of the Q_{LOO}^2 , the following condition is also verified:

$$Q_{LOO}^2 - Q_{ASYM}^2 > \delta Q \quad Q_{ASYM}^2 = 1 - (1 - R^2) \cdot \left(\frac{n}{n - p'} \right)^2$$

where δQ is a threshold value. It is implicitly assumed that a model with an actual Q_{LOO}^2 value lower than or very near the asymptotic value of a quantity δQ does not guarantee its future predictive ability. The simplest threshold value is $\delta Q = 0$, but a more conservative value could, for example, be -0.005 , while a less conservative value could be 0.005 .

RQK statistics. This is a multicriteria fitness function defined in terms of Q_{LOO}^2 statistics with additional constraints thought of to avoid some regression model pathologies [Todeschini, Consonni *et al.*, 2004a, 2004b]. By using the RQK function in the evaluation of a regression model or in generating an optimal model population by a $\rightarrow$ *variable selection* algorithm, one should maximize Q_{LOO}^2 and accept models only if the following criteria are all contemporarily satisfied:

1. $Q_{LOO}^2 > Q_0$ otherwise unacceptable predictive ability
2. $K_{XY} - K_X > \delta K$ otherwise high predictor multicollinearity
3. $Q_{LOO}^2 - Q_{ASYM}^2 > \delta Q$ otherwise doubtful predictive ability
4. $R^P > t^P$ otherwise redundancy in explanatory variables
5. $R^N > t^N$ otherwise overfitting due to noisy variables

where Q_0 is a user-defined threshold for the acceptable predictive ability (e.g., >0.5). Criterion 2 is the $\rightarrow$ *QUIK rule*, criterion 3 is the $\rightarrow$ *asymptotic Q^2 rule*, criterion 4 and 5 are the $\rightarrow$ *R function based rules*.

The general form of the RQK statistics is

$$RQK = Q_{LOO}^2 \cdot \prod_k I_k$$

where I_k is a binary function assuming a value equal to 1 if the k th criterion is satisfied, and zero otherwise. A family of RQK functions can be defined, depending on which criteria are adopted

and their threshold values. A reference value RQK^0 can be based on the less demanding threshold values for the five criteria:

$$Q_0 > 0.5 \quad \delta K = 0 \quad \delta Q = 0 \quad t^P = 0 \quad \text{and} \quad t^N = 0$$

R function-based rules. These were proposed to evaluate (a) the possible redundancy of a variable in a model (R^P function) and (b) the possible irrelevant contribution of a variable in a model (R^N function) [Todeschini, Consonni *et al.*, 2004a, 2004b].

The two functions are defined as

$$R^P = \prod_{j=1}^{p^+} \left(1 - M_j \cdot \left(\frac{p}{p-1} \right) \right) \quad \forall M_j > 0 \quad 0 \leq R^P \leq 1$$

$$R^N = \sum_{j=1}^{p^-} M_j \quad \forall M_j < 0 \quad -1 < R^N \leq 0$$

where p is the number of model variables and M_j is the following quantity:

$$M_j = \frac{R_{jy}}{R} - \frac{1}{p} \quad -\frac{1}{p} \leq M_j \leq \frac{p-1}{p}$$

where R_{jy} is the correlation coefficient between the j th model variable and y response. p^+ and p^- are the number of cases for which M_j has positive and negative values, respectively.

Regression models could be considered acceptable if for the two quantities R^P and R^N the following conditions hold:

$$R^P \geq t^P \quad R^N \geq t^N$$

where t^P and t^N are user-defined thresholds, depending on the data. t^P ranges from 0.01 to 0.1 and a suggested value for it is 0.05. R^P accounts for an excess of modeling variables without a global gain in the total R^2 . For example, a model having a total correlation of 0.71 constituted of two variables both having a pairwise correlation of 0.70 with the response seems to have a redundant variable.

R^N accounts for an excess of nonmodeling or useless variables and can be thought of to be a measure of overfitting due to noisy variables. It takes the maximum value equal to zero when no noisy variables are in the model. An estimate of the threshold can be performed assuming that all the nonmodeling variables had almost the same small correlation with response equal to ϵ , that is, each of such variables gives a contribution to R^N as

$$M_j = \frac{\epsilon}{R} - \frac{1}{p} = \frac{p \cdot \epsilon - R}{p \cdot R} \quad \epsilon \ll R$$

The value of ϵ can be tuned by the user, depending on the knowledge of the y response noise. Moreover, it can be assumed that no more than one noisy variable is allowed in the model, thus a threshold value for R^N can be estimated as

$$t^N(\epsilon) = \frac{p \cdot \epsilon - R}{p \cdot R}$$

The choice of eventually accepting one variable with low correlation with the response derives from the impossibility of knowing *a priori* if the variable is only noise or is useful in modeling residuals. The assumption of $\epsilon = 0$ gives a default threshold equal to $-1/p$.

- **regressive decremental distance sums** → layer matrices
- **regressive distance sums** → layer matrices
- **regressive incremental distance sums** $\equiv$ *regressive distance sums* → layer matrices
- **regressive vertex degrees** → layer matrices
- **regular graph** → graph

■ **Relative Alkylation Index (RAI)**

Proposed to quantify the skin sensitization of chemicals, *RAI* is a quantifier of the relative degree of carrier protein haptenation by a dose D of the compound under consideration. Besides being dependent on the dose given, the degree of carrier haptenation also depends on the compound's chemical reactivity (modeled by the $\log k_{rel}$ term) toward carrier protein nucleophiles and on its partitioning properties, modeled by the $\rightarrow \log P$ term [Roberts and Williams, 1982; Roberts, Fragnals *et al.*, 1991]. It is defined as

$$RAI = \log D + a \cdot \log k_{rel} + b \cdot \log P$$

The values of the coefficients a and b should be constant for a given series of compounds all based on the same molecular mechanism and all tested by the same protocol.

📖 [Betso, Carreon *et al.*, 1991; Franot, Roberts *et al.*, 1994]

- **relative atom-type count** → count descriptors
- **relative atomic moments** → self-returning walk counts
- **Relative Chirality Index** → chirality descriptors
- **relative electrophilicity** → quantum-chemical descriptors (⊙ Fukui functions)
- **relative entropy** $\equiv$ *Kullback–Leibler divergence* → information content
- **relative error in calculation** → regression parameters
- **relative error in prediction** → regression parameters
- **relative hydrophobic surface area** → charged partial surface area descriptors
- **relative negative charge** → charged partial surface area descriptors
- **relative negative charge surface area** → charged partial surface area descriptors
- **relative nucleophilicity** → quantum-chemical descriptors (⊙ Fukui functions)
- **relative polar surface area** → charged partial surface area descriptors
- **relative positive charge** → charged partial surface area descriptors
- **relative positive charge surface area** → charged partial surface area descriptors
- **relative retention time** → chromatographic descriptors (⊙ retention time)
- **relative topological distance** → bond order indices (⊙ conventional bond order)
- **relative valence connectivity indices** → combined descriptors
- **relative vertex distance complexity** → topological information indices
- **Ren vertex degree** → vertex degree
- **repeated evaluation set technique** → validation techniques (⊙ training/evaluation set splitting)
- **repulsive steric effects** → minimal topological difference
- **residual mean square** $\equiv$ *mean square error* → regression parameters
- **residual standard deviation** → regression parameters
- **residual standard error** $\equiv$ *residual standard deviation* → regression parameters
- **residual sum of squares** → regression parameters
- **resistance degree** → resistance matrix

- **resistance distance** → resistance matrix
- **resistance distance-detour distance combined matrix** → matrices of molecules (⊙ Table M3)
- **resistance distance/detour distance quotient matrix** → matrices of molecules (⊙ Table M2)
- **resistance distance-geometric distance combined matrix** → matrices of molecules (⊙ Table M3)
- **resistance distance/geometric distance quotient matrix** → matrices of molecules (⊙ Table M2)
- **resistance distance hyper-Wiener index** → Wiener matrix
- **resistance distance matrix** $\equiv$ *resistance matrix*
- **resistance/distance quotient matrix** → resistance matrix
- **resistance distance-topographic distance combined matrix** → matrices of molecules (⊙ Table M3)
- **resistance distance/topographic distance quotient matrix** → matrices of molecules (⊙ Table M2)
- **resistance distance-topological distance combined matrix** → matrices of molecules (⊙ Table M3)
- **resistance distance/topological distance quotient matrix** $\equiv$ *resistance/distance quotient matrix* → resistance matrix

■ **resistance matrix (Ω)** ($\equiv$ *resistance distance matrix*)

A square $A \times A$ symmetric matrix, A being the number of nonhydrogen atoms, whose off-diagonal entries are given by the resistance distance Ω_{ij} between any pair of vertices in the $\rightarrow H$ -depleted molecular graph G as [Klein and Randić, 1993]

$$[\Omega]_{ij} = \begin{cases} 0 & \text{if } i = j \\ \Omega_{ij} & \text{if } i \neq j \end{cases}$$

The **resistance distance** is based on the electrical network theory and is defined as the effective electrical resistance between two vertices (nodes) when a battery is connected across them and each graph edge is considered as a resistor taking a value of 1 ohm.

The resistance distance Ω_{ij} between two adjacent vertices v_i and v_j is obviously equal to 1 if the edge defined by the two vertices does not belong to a cycle, that is, there is a single path between them. The inverse of the resistance distance is the *conductance* σ_{ij} between two vertices v_i and v_j , which may be viewed as a sort of efficacy of communication between two sites; it is calculated as the following [Klein and Ivanciuc, 2001]:

$$\sigma_{ij} = \frac{1}{\Omega_{ij}} = \sum_{p_{ij}} |p_{ij}|^{-1}$$

where the sum runs over all the paths p_{ij} connecting the two considered vertices and $|p_{ij}|$ is the length of the considered path p_{ij} . Multiple connections between two vertices decrease the distance between them because the difficulty of transport decreases when there is more than one possibility for their communication. The **conductance matrix** (or **electrical conductance matrix**) is therefore the **reciprocal resistance matrix**, whose elements are the conductance values σ_{ij} between two vertices v_i and v_j [Ivanciuc, 2000h].

For any pair of nonadjacent vertices, the resistance distance is the effective resistance calculated according to the two classical Kirchhoff laws for series and parallel electrical circuits (Example R1).

The resistance distance between any pair of vertices can also be calculated by the $\rightarrow$ Laplacian matrix as the following:

$$\Omega_{ij} = (\mathbf{d}^{ij})^T \cdot \mathbf{\Gamma}^+ \cdot \mathbf{d}^{ij} = \mathbf{u}_i^T \cdot \mathbf{\Gamma}^+ \cdot \mathbf{u}_i - \mathbf{u}_i^T \cdot \mathbf{\Gamma}^+ \cdot \mathbf{u}_j - \mathbf{u}_j^T \cdot \mathbf{\Gamma}^+ \cdot \mathbf{u}_i + \mathbf{u}_j^T \cdot \mathbf{\Gamma}^+ \cdot \mathbf{u}_j$$

where $\mathbf{d}^{ij}$ is the vector obtained as the difference between $\mathbf{u}_i$ and $\mathbf{u}_j$ that are column vectors with every element equal to zero except the i th and j th element, respectively, $\mathbf{\Gamma}^+$ is the Moore–Penrose generalized inverse of the $\rightarrow$ Laplacian matrix.

It has been also shown [Xiao and Gutman, 2003] that the elements of the resistance matrix can be calculated from three other different matrices, denoted as $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$.

Let A be the number of atoms, B the number of bonds, $\mathbf{A}$ the $\rightarrow$ adjacency matrix and $\mathbf{V}$ the $\rightarrow$ vertex degree matrix, that is, a diagonal matrix whose diagonal elements are the $\rightarrow$ vertex degrees of the atoms. Moreover, $\mathbf{I}$ and $\mathbf{U}$ are the $\rightarrow$ identity matrix and the $\rightarrow$ unit matrix, respectively.

Then, the three matrices are defined as the following:

$$\mathbf{X} = \left(\mathbf{V} - \mathbf{A} + \frac{1}{A} \cdot \mathbf{U} \right)^{-1} \quad \mathbf{Y} = \left(\mathbf{V} - \mathbf{A} + \frac{1}{2B} \cdot \mathbf{V} \cdot \mathbf{U} \right)^{-1} \quad \mathbf{Z} = \left(\mathbf{I} - \mathbf{V}^{-1} \mathbf{A} + \frac{1}{2B} \cdot \mathbf{U} \cdot \mathbf{V} \right)^{-1}$$

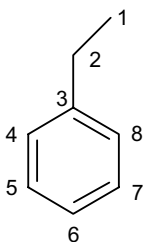
and the resistance matrix elements are accordingly calculated as

$$\Omega_{ij} = x_{ii} + x_{jj} - 2 \cdot x_{ij} \quad \Omega_{ij} = y_{ii} + y_{jj} - y_{ij} - y_{ji} \quad \Omega_{ij} = \frac{z_{ii}}{\delta_i} + \frac{z_{jj}}{\delta_j} - 2 \cdot \frac{z_{ij}}{\delta_j}$$

where δ is the vertex degree.

Example R1

Resistance matrix Ω for ethylbenzene.



$$\Omega_{15} = \Omega_{12} + \Omega_{23} + \Omega_{35} = 1 + 1 + 1/(1/2 + 1/4) = 2 + 4/3 = 3.333$$

$$\Omega_{28} = \Omega_{23} + \Omega_{38} = 1 + 1/(1 + 1/5) = 1 + 5/6 = 1.833$$

$$\Omega_{34} = 1/(1 + 1/5) = 5/6 = 0.833$$

Atom	1	2	3	4	5	6	7	8
1	0	1	2	2.833	3.333	3.500	3.333	2.833
2	1	0	1	1.833	2.333	2.500	2.333	1.833
3	2	1	0	0.833	1.333	1.500	1.333	0.833
4	2.833	1.833	0.833	0	0.833	1.333	1.500	1.333
5	3.333	2.333	1.333	0.833	0	0.833	1.333	1.500
6	3.500	2.500	1.500	1.333	0.833	0	0.833	1.333
7	3.333	2.333	1.333	1.500	1.333	0.833	0	0.833
8	2.833	1.833	0.833	1.333	1.500	1.333	0.833	0

The sum of the resistance distances between all pairs of vertices in the graph was proposed [Klein and Randić, 1993] as a molecular descriptor analogue to the $\rightarrow$ *Wiener index* and successively called **Kirchhoff number** Kf [Bonchev, Balaban *et al.*, 1994] and is defined as

$$Kf \equiv Wi(\Omega) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \Omega_{ij}$$

where Wi is the $\rightarrow$ *Wiener operator* and A the number of vertices in the graph. For acyclic graphs the Kirchhoff number coincides with the Wiener index as the resistance distances coincide with the classical topological distances for any pair of vertices. Therefore, it is mainly related to molecular size, like the Wiener index. Moreover, among the isomeric series it was proposed as better index of $\rightarrow$ *molecular cyclicity* than the Wiener index due to its low degeneracy; it decreases when the number of cycles and their centrality decrease.

The Kirchhoff number can also be directly calculated from the $\rightarrow$ *Laplacian matrix* by the following:

$$Kf = A \cdot tr(\Gamma^+)$$

where $tr(\Gamma^+)$ is the trace of the Moore–Penrose generalized inverse of the Laplacian matrix [Mardia, Kent *et al.*, 1988]. It was demonstrated that the Kirchhoff number and the $\rightarrow$ *quasi-Wiener index* coincide [Gutman and Mohar, 1996].

From the resistance matrix the **Balaban-like resistance index** was also defined as [Babic, Klein *et al.*, 2002]

$$J_{\Omega} = \frac{B}{C+1} \cdot \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot (\omega_i \cdot \omega_j)^{-1/2}$$

where ω_i is the **resistance degree** of the i th vertex calculated as the row sum of the resistance matrix:

$$\omega_i = \sum_{j=1}^A \Omega_{ij}$$

B and C are the number of vertices and the $\rightarrow$ *cyclomatic number*, respectively, and a_{ij} the elements of the $\rightarrow$ *adjacency matrix*, taking values equal to 1 for pairs of adjacent vertices, and zero otherwise. Other molecular descriptors were calculated by applying several different matrix operators [Ivanciuc, 2000h], such as $\rightarrow$ *spectral indices*, $\rightarrow$ *Hosoya-like indices*, and $\rightarrow$ *Balaban-like information indices*.

When there are more than one path (even of different lengths) between two vertices, the resistance distance is strictly less than the topological distance; therefore, the *resistance deficit* $d_{ij} - \Omega_{ij}$, or, otherwise, the *conductance excess* $\sigma_{ij} - 1/d_{ij}$, is related to the presence of cycles in the region of the graph between vertices v_i and v_j regardless of whether v_i and v_j are themselves in a cycle [Klein and Ivanciuc, 2001]. Based on this concept, two **global cyclicity indices** were proposed as

$$\Omega_C = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot (\sigma_{ij} - d_{ij}^{-1}) \quad \Omega_{\mu} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot (d_{ij} - \Omega_{ij})$$

where d_{ij} is the topological distance between vertices v_i and v_j , and a_{ij} are the elements of the adjacency matrix, taking value equal to 1 only for pairs of adjacent vertices. The first index is the **total excess bond conductance index**, while the second one is the **total bond resistance deficit index**.

Two $\rightarrow$ *quotient matrices* were derived from the resistance matrix [Babic, Klein *et al.*, 2002]. Namely, the **distance/resistance quotient matrix D/Ω** (or **topological distance/resistance distance quotient matrix**) was obtained by dividing the off-diagonal elements of the $\rightarrow$ *distance matrix D* by the corresponding elements of the resistance matrix Ω :

$$[D/\Omega]_{ij} = \begin{cases} \frac{d_{ij}}{\Omega_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} are the elements of the topological distance matrix.

The reciprocal of this matrix, obtained by inverting its off-diagonal elements, is the **resistance/distance quotient matrix Ω/D** (or **resistance distance/topological distance quotient matrix**), defined as

$$[\Omega/D]_{ij} = \begin{cases} \frac{\Omega_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

From the matrix D/Ω , the **D/Ω index** (also called **Wiener sum D/Ω index**) was defined as [Babic, Klein *et al.*, 2002]

$$D/\Omega = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \frac{d_{ij}}{\Omega_{ij}}$$

while, from the matrix Ω/D , the **Ω/D index** (also called **Kirchhoff sum index, KfS**) was defined as [Babic, Klein *et al.*, 2002]

$$KfS \equiv \Omega/D = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \frac{\Omega_{ij}}{d_{ij}}$$

📖 [Ivanciuc, Ivanciuc *et al.*, 1997; Lukovits, Nikolić *et al.*, 1999, 2000; Palacios, 2001; Klein, 2002; Fowler, 2002; Ivanciuc, 2002d; Xiao, 2004]

- **resonance** $\rightarrow$ delocalization degree indices
- **resonance constant** $\rightarrow$ electronic substituent constants ($\odot$ field/resonance effect separation)
- **resonance effect** $\rightarrow$ electronic substituent constants
- **resonance electronic constants** $\rightarrow$ electronic substituent constants
- **resonance energy** $\rightarrow$ delocalization degree indices
- **resonance indices** $\rightarrow$ delocalization degree indices
- **resonance polar effect** $\rightarrow$ electronic substituent constants ($\odot$ resonance electronic constants)

- **resonance-weighted edge adjacency matrix** → edge adjacency matrix
- **resonance-weighted edge connectivity index** → edge adjacency matrix
- **response variables** $\equiv$ *dependent variables* → data set
- **response-variable correlation cutoff** → variable selection
- **restricted random walk matrix** → walk matrices
- **restricted walk ID number** → walk matrices
- **retardation factor** → chromatographic descriptors ($\odot$ Bate–Smith–Westall retention index)
- **retention factor** $\equiv$ *capacity factor* → chromatographic descriptors
- **retention time** → chromatographic descriptors
- **reversed Balaban index** $\equiv$ *complement Balaban index* → distance matrix
- **reversed distance matrix** $\equiv$ *distance complement matrix* → distance matrix
- **reverse detour index** → detour matrix
- **reverse detour matrix** → detour matrix
- **reverse-distance sum** → distance matrix
- **reversed Wiener index** $\equiv$ *complement Wiener index* → distance matrix
- **reverse Wiener index** → distance matrix
- **reverse Wiener matrix** → distance matrix
- **reversible decoding** → structure/response correlations
- **revised Wiener index** → Szeged matrices
- **REX descriptors** → substructure descriptors
- **R function-based rules** → regression parameters
- **RHTA index** → charged partial surface area descriptors
- **rigid bond number** → flexibility indices ($\odot$ rotatable bond number)
- **R⁺ index** → distance matrix
- **R indices** → GETAWAY descriptors
- **R* indices** → distance matrix
- **ring bridges** → ring descriptors
- **ring complexity index** → ring descriptors
- **ring degree** → Cao–Yuan indices
- **ring degree-distance index** → Cao–Yuan indices

■ **ring descriptors** ($\equiv$ *cyclicity indices*)

Ring descriptors are numerical quantities encoding information about the presence of rings in a molecule. The most known and simple ring descriptor is the **cyclomatic number** (or **ring number**), derived from the → *Euler formula* and proposed by Kurnakov [Kurnakov, 1928] and later by Frèrejacque [Frèrejacque, 1939]. It is defined as the number of independent molecule cycles C , that is, the number of nonoverlapping cycles [Lipkus, 2001; Wilson, 1972]. Widely used in the description of ring systems, the cyclomatic number is often also denoted by μ and is calculated as the cardinality of the set of independent rings called the **Smallest Set of Smallest Rings (SSSR)**. The cyclomatic number of a polycyclic graph is equal to the minimum number of edges that must be removed from the graph to transform it to the related acyclic graph. It is equal to zero for trees and to one for monocyclic graphs.

The cyclomatic number is the simplest descriptor that discriminates cyclic compounds from acyclic ones and is related, via the → *Euler formula* applied to a graph, to the number of bonds B and atoms A in a molecule as

$$C \equiv \mu = B - A + 1$$

where B and A are the total numbers of bonds and atoms, respectively. The cyclomatic number is the usual way that a chemist counts rings in a molecular structure. This descriptor appears as a part of other molecular descriptors such as the $\rightarrow$ *Balaban distance connectivity index* [Balaban, 1976d; Hanser, Jauffret *et al.*, 1996] and several $\rightarrow$ *Balaban-like indices*.

A more general expression for the cyclomatic number is

$$C \equiv \mu = B - A + D$$

where D is the number of disconnected fragments into the whole molecular structure.

The cyclomatic number must not be confused with the graph $\rightarrow$ *cyclicity* C^+ , that is, the number of all possible cycles in a graph. Thus, for example, naphthalene has a cyclomatic number equal to two (the two benzene rings) and a cyclicity equal to three (the two benzene rings plus the more external 10-atom ring).

The maximal number of rings in the **Extended Set of Smallest Rings (ESSR)** is given by [Downs, Gillet *et al.*, 1989a]

$$C^{\text{MAX}} = \begin{cases} C & \text{if } 2 \cdot C \leq A \\ C + 1 & \text{if } 2 \cdot C > A \end{cases}$$

where C is the cyclomatic number and A the number of atoms. For most molecules, $C^{\text{MAX}} = C$; however, for example, in cyclocubane ($A = 8$ and $B = 12$), since the cyclomatic number C is equal to 5 ($12 - 8 + 1$), then twice the cyclomatic number is greater than the number of atoms ($2 \times 5 > 8$), and thus C^{MAX} is equal to 6. In other words, ESSR contains all the chemically meaningful cycles.

The ratio of the cyclomatic number C over the number of atoms was proposed as a size-independent measure for the complexity of polycyclic molecules [Rücker and Rücker, 2000].

A **ring system** is defined as the system including all the cycles sharing at least one edge (fused or spiro connections between two atoms).

The **number of ring systems (NRS)** is calculated as [Feher and Schmidt, 2003]

$$NRS = (B - B_R) - (A - A_R) + 1$$

where B and A are the total numbers of bonds and atoms, respectively; B_R and A_R are the number of atoms and bonds belonging to rings, respectively. The **normalized number of ring systems** (NRS^*) is defined as [Feher and Schmidt, 2003]

$$NRS^* = \frac{NRS}{C}$$

where C is the cyclomatic number.

Moreover, the reciprocal of NRS^* , called **ring fusion degree (RFD)**, was also defined as [Xu and Stevenson, 2000]

$$RFD = \frac{C}{NRS}$$

Note that both NRS^* and RFD are not defined for acyclic structures.

The sum of the ring size of all the single cycles of all the ring systems is called **total ring size** and denoted as R :

$$R = \sum_{r=1}^C A_R(r)$$

where $A_R(r)$ is the number of atoms forming each r th ring and C the cyclomatic number; in this case, the atoms belonging to fused connections are counted twice.

Two simple ring descriptors are the **two-degree cyclic atom count** (2_dgc) that is the number of atoms in a cycle having a vertex degree equal to 2, and **three-degree cyclic atom count** (3_dgc) that is the number of atoms in a cycle having a vertex degree equal to 3 [Lin and Tsai, 2003].

The **ring perimeter** R_P represents the perimeter of all the rings present in the molecule, while the **ring bridges** R_B represents the number of bridge edges [Lipkus, 2001]:

$$R_P = 2 \cdot B_R - R \quad R_B = R - B_R$$

where B_R is the sum of bonds of the ring systems and R the total ring size; by definition $B_R = R_P + R_B$. Using these two descriptors R_P and R_B together with the cyclomatic number C , ring structures can be classified according to their shape in the 3D space [Lipkus, 2001].

The **ring complexity index** was defined as [Gasteiger and Jochum, 1979]

$$C_R = \frac{R}{A_R}$$

where R is the total ring size and A_R is the total number of atoms belonging to any ring system. For isolated rings $C_R = 1$, for fused or bridged ring systems $C_R > 1$, for molecules with no rings $C_R = 0$.

The **molecular cyclized degree** (MCD) was defined as [Lin and Tsai, 2003]

$$MCD = \frac{A_R}{A}$$

where A_R is the total number of atoms belonging to any ring system.

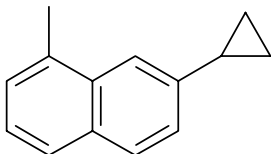
Another measure of the ring characteristics, called **ring fusion density**, denoted as RFA , is here defined as [Authors, This Book]:

$$RFA = 2 \cdot \frac{R_B}{A_R}$$

where R_B is the number of ring bridges, A_R the total number of atoms belonging to ring systems; the coefficient 2 allows to reach a value of 1 for coronene ($R_B = 12$ and $A_R = 24$). $RFA = 0$ for all acyclic and monocyclic molecules; $RFA = 0.5$ for bicyclobutane, $RFA = 0.20$ for naphthalene, $RFA = 0.286$ for phenanthrene and anthracene, $RFA = 0.333$ for triphenylene, $RFA = 0.625$ for pyrene.

Example R2

Calculation of some ring descriptors.



$$MCD = 13/14 = 0.929$$

$$RFD = 3/2 = 1.5$$

$$RF\Delta = (2 \times 1)/13 = 0.154$$

$$C^{MAX} = 3$$

$$A = 14$$

$$B = 16$$

$$A_R = 10 + 3 = 13$$

$$B_R = 11 + 3 = 14$$

Number of smallest rings in the smallest set: $C = 16 - 14 + 1 = 3$

Number of all possible rings: $C^+ = 4$

Number of ring systems:

$$NRS = (16 - 14) - (14 - 13) + 1 = 2$$

$$NRS^* = 2/3 = 0.667$$

$$R = 6 + 6 + 3 = 15$$

$$2\_dgc = 8 \quad 3\_dgc = 5$$

$$R_p = 2 \times 14 - 15 = 13$$

$$R_B = 15 - 14 = 1$$

$$C_R = 15/13 = 1.154$$

Information about atoms and bonds belonging to cycles is usually encoded by the $\rightarrow$ *vertex-cycle incidence matrix* and the $\rightarrow$ *edge-cycle incidence matrix*.

The presence of cycles in a molecule, that is, molecular cyclicity, is a component in the evaluation of the whole $\rightarrow$ *molecular complexity*, together with molecular size and branching. Since rings in molecules are structural features very important in determining physico-chemical properties and biological activities, several algorithms were developed for computerized ring perception [Randić and Wilkins, 1980; Fujita, 1988; Downs, Gillet *et al.*, 1989a, 1989b; Sadowski, Gasteiger *et al.*, 1994; Figueras, 1996; Hanser, Jauffret *et al.*, 1996; Xu, 1996, 2003; Lipkus, 1997, 2001; Dury, Latour *et al.*, 2001; García, Ruiz *et al.*, 2002; Badreddin Abolmaali, Wegner *et al.*, 2003; Downs, 2003].

- **ring fusion degree** $\rightarrow$ ring descriptors
- **ring fusion density** $\rightarrow$ ring descriptors
- **ring ID number** $\rightarrow$ ID numbers
- **ring number** $\equiv$ *cyclomatic number* $\rightarrow$ ring descriptors
- **ring perimeter** $\rightarrow$ ring descriptors
- **ring system** $\rightarrow$ ring descriptors
- **R-matrix leading eigenvalue** $\rightarrow$ GETAWAY descriptors
- **Roberts–Moreland inductive constant** $\rightarrow$ electronic substituent constants ($\odot$ inductive electronic constants)
- **ROC curve** $\equiv$ *Receiver Operator Characteristic curve* $\rightarrow$ classification parameters
- **Rogers–Tanimoto similarity coefficient** $\rightarrow$ similarity/diversity (Table S9)
- **rooted tree** $\rightarrow$ graph ($\odot$ tree)
- **root mean square** $\rightarrow$ statistical indices ($\odot$ indices of central tendency)
- **root mean square deviation** $\equiv$ *root mean square error* $\rightarrow$ regression parameters

- **root mean square deviation in prediction** $\equiv$ *root mean square error in prediction* $\rightarrow$ regression parameters
- **root mean square error in prediction** $\rightarrow$ regression parameters
- **root mean square Wiener index** $\rightarrow$ Wiener index
- **rotamer factors** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **rotamer modification number** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **rotatable bond fraction** $\rightarrow$ flexibility indices
- **rotatable bond number** $\rightarrow$ flexibility indices
- **rotational invariance** $\rightarrow$ molecular descriptors ($\odot$ invariance properties of molecular descriptors)
- **roughness** $\equiv$ *ovality index* $\rightarrow$ shape descriptors
- **Rouvray index** $\rightarrow$ distance matrix
- **row sum operator** $\rightarrow$ algebraic operators
- **row sum vector** $\rightarrow$ algebraic operators ($\odot$ row sum operator)
- **RQK statistics** $\rightarrow$ regression parameters
- **RSAA index** $\rightarrow$ charged partial surface area descriptors
- **RSAH index** $\rightarrow$ charged partial surface area descriptors
- **RSAM index** $\rightarrow$ charged partial surface area descriptors
- **RSHM index** $\rightarrow$ charged partial surface area descriptors
- **R total index** $\rightarrow$ GETAWAY descriptors
- **Ruch's chirality functions** $\rightarrow$ chirality descriptors
- **rugosity** $\rightarrow$ grid-based QSAR techniques ($\odot$ VolSurf descriptors)
- **rule-of-five** $\equiv$ *Lipinski drug-like index* $\rightarrow$ property filters ($\odot$ drug-like indices)
- **rule of six** $\equiv$ *six position number* $\rightarrow$ steric descriptors ($\odot$ number of atoms in substituent specific positions)
- **rule-of-three** $\rightarrow$ property filters ($\odot$ lead-like indices)
- **rule-of-unity** $\rightarrow$ property filters ($\odot$ drug-like indices)
- **Russell–Rao similarity coefficient** $\rightarrow$ similarity/diversity ($\odot$ Table S9)
- **R weighting scheme** $\rightarrow$ weighting schemes

S

- **Sachs graph** → graph
- **Sadhana index** → empirical indices
- **SAF index** → scoring functions (⊙ substructural analysis)
- **SAL Index** → Structure/Response Correlations
- **Sanderson group electronegativity** → atomic electronegativity
- **SAR Index** → Structure/Response Correlations
- **SCAA₁ index** ≡ *HACA index* → charged partial surface area descriptors
- **SCAA₂ index** → charged partial surface area descriptors (⊙ *HACA index*)
- **scalar product of vectors** → algebraic operators
- **Schläfli indices** → Euler's formula
- **Schläfli topological form index** → Euler's formula
- **Schrödinger equation** → quantum-chemical descriptors

■ Schultz molecular topological index (*MTI*)

A topological index, originally called by the author **Molecular Topological Index**, derived from the → *adjacency matrix* **A**, the → *distance matrix* **D**, and the *A*-dimensional column vector **v** constituted by the → *vertex degree* δ of the *A* atoms in the → *H-depleted molecular graph* [Schultz, 1989]. The Schultz index is defined as

$$MTI \equiv S = \sum_{i=1}^A [(\mathbf{A} + \mathbf{D}) \cdot \mathbf{v}]_i = \sum_{i=1}^A t_i$$

where t_i are the elements, called **intricity numbers**, of the *A*-dimensional column vector **t** obtained as the following:

$$\mathbf{t} = (\mathbf{A} + \mathbf{D}) \cdot \mathbf{v}$$

that is, matrices **D** and **A** are summed and then multiplied by the vector **v**. The matrix **A** + **D** is called **adjacency-plus-distance matrix** and denoted ^{AD}**Σ** [Schultz, 1989; Müller *et al.*, 1990b; Müller, Szymanski *et al.*, 1990b].

Intricity numbers measure the combined influence of valence, adjacency, and distance for each comparable set of vertices; the lower the intricity number, the more intricate or complex the vertex [Schultz and Schultz, 1993].

The Schultz index can be decomposed in two distinct parts corresponding to two independent descriptors:

$$MTI = \mathbf{u}^T [\mathbf{A} \cdot (\mathbf{A} + \mathbf{D})] \mathbf{u} = \mathbf{u}^T (\mathbf{A}^2 + \mathbf{A} \cdot \mathbf{D}) \mathbf{u} = \mathbf{u}^T \mathbf{A}^2 \mathbf{u} + \mathbf{u}^T (\mathbf{A} \cdot \mathbf{D}) \mathbf{u} = M_2 + MTI'$$

$$M_2 = \mathbf{u}^T \mathbf{A}^2 \mathbf{u} = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{A}^2]_{ij} = \sum_{i=1}^A \delta_i^2$$

$$MTI' \equiv S' = \mathbf{u}^T (\mathbf{A} \cdot \mathbf{D}) \mathbf{u} = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{A} \cdot \mathbf{D}]_{ij}$$

where $\mathbf{u}$ is an A -dimensional unit column vector, M_2 is just the $\rightarrow$ *second Zagreb index*, and S' is the nontrivial part of MTI , originally denoted by MTI' and called **MTI' index** [Müller, Szymanski *et al.*, 1990b; Mihalić, Nikolić *et al.*, 1992].

The **S' index** can also be written as

$$MTI' \equiv S' = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{A} \cdot \mathbf{D}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A (\delta_i + \delta_j) \cdot d_{ij} = \sum_{i=1}^A \delta_i \cdot \sum_{j=1}^A d_{ij} = \sum_{i=1}^A \delta_i \cdot \sigma_i \equiv D'$$

where d_{ij} is the $\rightarrow$ *topological distance* and σ_i is the i th $\rightarrow$ *vertex distance degree* [Gutman, 1994b; Schultz, Schultz *et al.*, 1995]. From the above relationships follows the coincidence between the S' index and the **degree distance of the graph** D' , later proposed by Dobrynin [Dobrynin and Kochetova, 1994], simply summing over all graph vertices the local invariants called **vertex degree distance** defined as

$$D'_i = \delta_i \cdot \sigma_i$$

Since the S' index is a sum of δ -weighted vertex distance degrees, it can be considered a vertex-valency-weighted analogue $\rightarrow$ *Wiener index*.

Reciprocal Schultz indices were defined as [Schultz and Schultz, 1998]

$$RS = \sum_{i=1}^A [(\mathbf{A} + \mathbf{D}^{-1}) \cdot \mathbf{v}]_i \quad \text{and} \quad RS' = \sum_{i=1}^A [\mathbf{D}^{-1} \cdot \mathbf{v}]_i = \sum_{i=1}^A \delta_i \cdot RDS_i$$

where $\mathbf{D}^{-1}$ is the $\rightarrow$ *reciprocal distance matrix* and RDS is the $\rightarrow$ *reciprocal distance sum*.

A Schultz-type topological index – **Gutman molecular topological index** S_G – was also defined by Gutman [Gutman, 1994b], as

$$S_G = \sum_{i=1}^A \sum_{j=1}^A \delta_i \cdot \delta_j \cdot d_{ij}$$

where $\delta_i \cdot \delta_j \cdot d_{ij}$ is the topological distance between the vertices v_i and v_j weighted by the product of the endpoint vertex degrees. Like the S' index, the S_G index is a vertex-valency-weighted analogue Wiener index, whereas the weighting factor is multiplicative instead of additive.

The Schultz MTI was demonstrated [Klein, Mihalić *et al.*, 1992; Plavšić, Nikolić *et al.*, 1993a] to be strongly correlated to the Wiener index W according to the following formal relation:

$$MTI = 4 \cdot W + 2 \cdot N_2 - (A-1) \cdot (A-2)$$

where N_2 is the $\rightarrow$ connection number and A is the number of graph vertices. This relation is only true if the molecular graph G is a tree.

An **edge-Schultz index** has been derived from the $\rightarrow$ edge adjacency matrix ${}^E\mathbf{A}$ and the $\rightarrow$ edge distance matrix ${}^E\mathbf{D}$ [Estrada and Gutman, 1996; Estrada and Rodriguez, 1997]:

$${}^E\text{MTI} = \sum_{b=1}^B [({}^E\mathbf{A} + {}^E\mathbf{D}) \cdot \mathbf{v}]_b$$

where $\mathbf{v}$ in this case is a B -dimensional column vector whose elements are the $\rightarrow$ edge degrees ϵ .

The matrix ${}^E\mathbf{A} + {}^E\mathbf{D}$ is called **edge-adjacency-plus-edge-distance matrix** and denoted as ${}^{\text{EAD}}\mathbf{\Sigma}$. In the same way an **edge-Gutman index** can be defined as

$${}^E S_G = \sum_{i=1}^B \sum_{j=1}^B \epsilon_i \cdot \epsilon_j \cdot [{}^E\mathbf{D}]_{ij}$$

where ϵ_i and ϵ_j are the edge degrees of the two considered edges and $[{}^E\mathbf{D}]_{ij}$ is the topological distance between them.

A generalization of the Schultz molecular topological index to account for heteroatoms and multiple bonds was proposed based on the $\rightarrow$ Barysz distance matrix.

The **3D-Schultz index** ${}^3\text{D} \text{MTI}$ is derived from the $\rightarrow$ geometry matrix $\mathbf{G}$ as [Mihalić, Nikolić *et al.*, 1992]

$${}^3\text{D} \text{MTI} = \sum_{i=1}^A [({}^b\mathbf{A} + \mathbf{G}) \cdot \mathbf{v}]_i$$

where ${}^b\mathbf{A}$ is the $\rightarrow$ bond length-weighted adjacency matrix whose entries corresponding to bonded atoms are $\rightarrow$ bond distances and $\mathbf{v}$ is an A -dimensional column vector constituted by the $\rightarrow$ vertex degree δ of the A atoms in the H-depleted molecular graph. Analogously, the **3D-MTI' index** is defined as

$${}^3\text{D} \text{MTI}' = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{A} \cdot \mathbf{G}]_{ij} = \sum_{i=1}^A [\mathbf{v}^T \cdot \mathbf{G}]_i$$

where $\mathbf{v}^T$ is the transposed vector $\mathbf{v}$ defined above.

By analogy with the Schultz molecular topological index, $\rightarrow$ Schultz-type indices were also proposed.

📖 [Jurić, Gagro *et al.*, 1992; Klavžar and Gutman, 1996; Gutman and Klavžar, 1997]

■ Schultz-type indices

Defined by analogy with the $\rightarrow$ Schultz molecular topological index MTI , Schultz-type indices are $\rightarrow$ molecular descriptors based on a product of square $A \times A$ matrices, the $\rightarrow$ adjacency matrix $\mathbf{A}$ being obligatory [Diudea, Pârv *et al.*, 1997b; Diudea and Gutman, 1998].

The general formula of Schultz-type indices is the following:

$$\text{MTI}_{(\mathbf{M}_1, \mathbf{A}, \mathbf{M}_2)} = \mathbf{u}^T [\mathbf{M}_1 \cdot (\mathbf{A} + \mathbf{M}_2)] \mathbf{u} = \mathbf{u}^T [\mathbf{M}_1 \cdot \mathbf{A} + \mathbf{M}_1 \cdot \mathbf{M}_2] \mathbf{u} = S(\mathbf{M}_1 \cdot \mathbf{A}) + S(\mathbf{M}_1 \cdot \mathbf{M}_2)$$

where $\mathbf{u}$ is a unit column vector of size A , A being the number of graph vertices; $\mathbf{M}_1$ and $\mathbf{M}_2$ are two generic square $A \times A$ matrices, and S is the sum of all the matrix elements.

Selecting different combinations of $\mathbf{M}_1$ and $\mathbf{M}_2$ matrices leads to the derivation of several Schultz-type indices. The original Schultz molecular topological index MTI is obtained for $\mathbf{M}_1 = \mathbf{A}$ and $\mathbf{M}_2 = \mathbf{D}$, where $\mathbf{D}$ is the topological $\rightarrow$ distance matrix. Typical Schultz indices are derived from $(\mathbf{D}, \mathbf{A}, \mathbf{D})$, $(\mathbf{D}^{-1}, \mathbf{A}, \mathbf{D}^{-1})$, $(\mathbf{W}, \mathbf{A}, \mathbf{D})$, $(\mathbf{W}^{-1}, \mathbf{A}, \mathbf{D}^{-1})$, $(\mathbf{W}, \mathbf{A}, \mathbf{W})$, $(\mathbf{UCJ}, \mathbf{A}, \mathbf{UCJ})$, $(\mathbf{USZ}, \mathbf{A}, \mathbf{USZ})$, where $\mathbf{D}^{-1}$ is the $\rightarrow$ reciprocal distance matrix, $\mathbf{W}$ is one among $\rightarrow$ walk matrices, $\mathbf{W}^{-1}$ is the $\rightarrow$ reciprocal walk matrix, $\mathbf{UCJ}$ and $\mathbf{USZ}$ the unsymmetrical Cluj and Szeged matrices, respectively.

Schultz-type indices based on eigenvectors were proposed [Medeleanu and Balaban, 1998] as $\rightarrow$ local vertex invariants and molecular descriptors on the basis of the eigenvector of adjacency and distance matrices associated with the lowest (largest negative) eigenvalue. The LOVIs are defined as the following A -dimensional column vector:

$$\begin{aligned} \mathbf{v}_1 &= (\mathbf{A} + \mathbf{D}) \cdot \mathbf{v}_\mathbf{A} & \mathbf{v}_2 &= (\mathbf{A} + \mathbf{D}) \cdot \mathbf{v}_\mathbf{D} \\ \mathbf{v}_3 &= \mathbf{A} \cdot \mathbf{v}_\mathbf{A} & \mathbf{v}_4 &= \mathbf{A} \cdot \mathbf{v}_\mathbf{D} & \mathbf{v}_5 &= \mathbf{D} \cdot \mathbf{v}_\mathbf{A} & \mathbf{v}_6 &= \mathbf{D} \cdot \mathbf{v}_\mathbf{D} \end{aligned}$$

where $\mathbf{A}$ is the adjacency matrix, $\mathbf{D}$ the distance matrix, $\mathbf{v}_\mathbf{A}$ the eigenvector of the adjacency matrix, and $\mathbf{v}_\mathbf{D}$ the eigenvector of the distance matrix. $\mathbf{v}_1$ – $\mathbf{v}_6$ are vectors containing different local vertex invariants; the highest values of $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_5$, and $\mathbf{v}_6$ LOVIs correspond to vertices of lower degree and farther from the graph center, $\mathbf{v}_3$ LOVIs show the opposite trend as do the coefficients of the eigenvector $\mathbf{v}_\mathbf{A}$, and the variation of $\mathbf{v}_4$ LOVIs is the least regular.

From these sets of LOVIs, two different types of topological index were derived:

$$\text{XMTn} = \sum_{i=1}^A [\mathbf{v}_\mathbf{n}]_i \quad \text{XMTnR} = \sum_{b=1}^B ([\mathbf{v}_\mathbf{n}]_i \cdot [\mathbf{v}_\mathbf{n}]_j)^{-1/2}_b$$

where the summation is over all of the LOVIs in each vector $\mathbf{v}_\mathbf{n}$, $\mathbf{v}_\mathbf{n}$ being equal to $\mathbf{v}_1$ – $\mathbf{v}_6$. The second index is based on a Randić-like formula, where the summation runs over the edges in the graph, B being the total number of graph edges, and the product is between the LOVIs of the two vertices incident to the considered edge.

📖 [Diudea, 1995a; Diudea and Pop, 1996; Diudea and Randić, 1997]

- **Schultz-type indices based on eigenvectors** $\rightarrow$ Schultz-type indices
- **Schultz weighted distance matrices** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **Schwarz Bayesian Information Criterion** $\rightarrow$ regression parameters ($\odot$ Table R1)
- **score matrix** $\rightarrow$ Principal Component Analysis

■ scoring functions ($\equiv$ drug-like scores)

Scoring functions are molecular descriptors encoding information about drug likeness of compounds [Oprea, Gottfries *et al.*, 2000; Walters and Murcko, 2002; Muegge, 2003; Leach, Hann *et al.*, 2006]. As the strictly related $\rightarrow$ property filters, they are mainly applied in the design of combinatorial libraries and screening of virtual libraries. They aim at reducing the number of compounds to be synthesized and tested, allowing the selection of those compounds that have desired properties to be good drug candidates.

The scoring functions measure the similarity of compounds to existing drugs by either structural features or by chemical properties or by both structural and chemical properties; their values are ideally distributed between zero (i.e., nondrug-like) and one (i.e., drug-like).

Scoring functions can be considered as molecular descriptors that enable molecules to be ranked according to their likelihood of exhibiting activity [Gillet, Willett *et al.*, 1998].

Several scoring functions have been proposed so far that differ for the molecular properties they are based on and the statistical and machine-learning approaches used to relate molecular descriptors to the drug-like property.

Drug-like scores are frequently the result of binary classification models, which evaluate the potential biological behavior of a molecule on the basis of a number of contemporarily occurring structural features and molecular property values.

Some scoring functions are listed below.

- **substructural analysis**

Substructural analysis is substructure searching where weights are calculated, relating the presence of a specific substructure moiety in a molecule to the probability that the molecule is active in some biological test system [Cramer III, Redl *et al.*, 1974; Hodes, Hazard *et al.*, 1977; Ormerod, Willett *et al.*, 1989, 1990; Craig, 1990; Klopman, 1992; Gillet, Willett *et al.*, 1998].

This analysis was designed specifically for large structurally diverse data sets resulting useful for lead discovery and drug design. The two basic assumptions in substructural analysis are (a) a weight (**SAF index**) for each substructure can be calculated, which characterizes its differential occurrence in active and inactive compounds [Cramer III, Redl *et al.*, 1974]:

$$SAF_j = \frac{N_j^{AC}}{N_j^{AC} + N_j^{IN}}$$

where N_j^{AC} and N_j^{IN} are the number of active and inactive compounds containing the j th substructure in the data set, respectively; and (b) the overall probability of activity of a compound can be calculated by summing (or otherwise combining) the weights for its constituent substructures.

The additive scheme implicitly assumes that the contributions of the different substructures to the molecule activity are statistically independent of each other.

- **Hodes statistical–heuristic method**

Hodes statistical–heuristic method is an approach similar to the $\rightarrow$ *substructural analysis* in that it is concerned with the calculation of weights from the relative frequency of substructure occurrences in known active and inactive molecules, but these weights are used to select likely active compounds and predict biological activity [Hodes, Hazard *et al.*, 1977; Hodes, 1981a, 1981b]. For each present substructure in a large data set of molecules, a weight for activity and a weight for inactivity are derived by calculating separately the incidence of each substructure in active and inactive compounds. The binomial distribution is assumed with mean m^B and standard deviation s^B given by

$$p_j = \frac{N_j}{n} \quad m_j^B = N^{AC} \cdot p_j \quad s_j^B = \sqrt{N^{AC} \cdot p_j \cdot (1 - p_j)}$$

where N_j^{AC} is the number of active compounds, N_j the number of compounds (both active and inactive) containing the j th substructure, n the total number of compounds, and p_j the incidence of the j th substructure in the whole set of compounds, that is, its probability. Therefore, each weight w_j^{AC} is calculated as logarithm of the inverse of the probability that the number of standard deviations away from the mean can get by chance:

$$w_j^{\text{AC}} = \log\left(\frac{1}{p(z_j)}\right) \quad z_j = \frac{N_j^{\text{AC}} - m_j^{\text{B}}}{s_j^{\text{B}}}$$

where $p(z_j)$ is the one-tailed probability obtained using the normal approximation to the binomial probability and N_j^{AC} the number of active compounds containing the j th substructure.

For example, a substructure occurs 17.7% in the whole data set, that is, $p = 0.177$. In the subset of 33 active compounds ($N_j^{\text{AC}} = 33$), 5.84 compounds ($m_j^{\text{B}} = 0.177 \times 33$) are expected to have this substructure, assuming that it has nothing to do with activity. This number is assumed as the mean of the binomial distribution, with a standard deviation equal to 2.19 ($s_j^{\text{B}} = \sqrt{33 \times 0.177 \times (1 - 0.177)}$). If the actual number of active compounds containing the considered substructure is 11, the number of standard deviations away from the mean is

$$z = \frac{11 - 5.84}{2.19} = 2.36$$

The one-tailed probability obtained using the normal approximation to the binomial probability is 0.0091 and the logarithm of the inverse of this probability value is used as activity weight w_j^{AC} for the considered substructure, that is, $\log(109.89) = 2.04$, the smaller the probability, the larger the significance. The inactivity weight w_j^{IN} is analogously calculated by using the subset of inactive compounds.

The drug-like score s of a compound is finally calculated by adding the weights for each substructural feature present in the considered compound as

$$s = \sum_j \log(w_j^{\text{AC}}) = \log\left(\prod_j w_j^{\text{AC}}\right)$$

where the logarithm is used to give convenient magnitudes to the scores. The score is an estimate of the probability that the compound belongs to the active set.

Note. It must be noted that for a number of active compounds in the data set significantly smaller than the expected mean value, leading to negative values of z ($N_j^{\text{AC}} < m_j^{\text{B}}$), an incorrect large activity weight is obtained. In the previous numerical example, if no active compounds instead of 11 contained the considered substructure, the active weight would be 2.42; with 6 active compounds, the active weight would be 0.326! Then, in this case ($z < 0$), a correct weight is assigned by using the complementary probability, that is, $1 - p(z_j)$. In the given example, if no active compounds contained the considered j th substructure, the final active weight would be $\log(1/(1 - 0.0038)) = 1.6 \times 10^{-3}$.

• average binding energy

The average binding energy was originally proposed [Andrews, Craik *et al.*, 1984] to measure the free energy of interaction that a small organic molecule might be expected to express in an interaction with a biological macromolecule. Then, it was used as drug-like score with the name

of **maximal binding energy** [Leach, Bradshaw *et al.*, 1999]. It is estimated by summing the intrinsic binding energies of specific functional groups present in the molecule and subtracting an entropic factor ($T\Delta S = -14$ kcal/mol) and a term related to the degrees of internal rotational freedom (-0.7). The intrinsic binding energies were calculated for 10 functional groups by regression analysis applied on a training set of 200 drug molecules (Table S1). Compounds of a library can be then sorted on the basis of the binding energy values; the higher the binding energy of a compound, the higher is its drug likeness.

Table S1 Intrinsic binding energies by [Andrews, Craik *et al.*, 1984].

Functional group	Kcal/mol	Functional group	Kcal/mol
Csp ³	0.8	O, S (ethers)	1.1
Csp ²	0.7	Halogens	1.3
N	1.2	CO ₂ ⁻	8.2
N ⁺	11.5	C=O	3.4
OH	2.5	PO ₄ ²⁻	10.0

• Klopman–Henderson cumulative substructure count

This is a QSAR graph theory-based method involving a heuristic processing of $\rightarrow$ *H-depleted molecular graphs* represented by common substructure counts [Klopman and Henderson, 1991]. The method consists in extracting $\rightarrow$ *substructure descriptors* from the data set and then searching for their significance in correlating the biological activity by the aid of an approach similar to the $\rightarrow$ *substructural analysis*.

From the frequencies of the row entries of the $\rightarrow$ *distance matrix* of each molecule, the $\rightarrow$ *vertex distance code* is defined as

$$\{^1f_i, ^2f_i, ^3f_i, \dots, \eta_i f_i\}$$

where $^1f_i, ^2f_i, ^3f_i, \dots$ are the $\rightarrow$ *vertex distance counts* indicating the frequencies of distances equal to 1, 2, 3, ..., respectively, from vertex v_i to any other vertex and η_i is the i th $\rightarrow$ *atom eccentricity*.

A cumulative path matrix of dimension $p \times D$ is calculated, where each row represents one among the p -derived fragment descriptors relative to the considered molecule and D being the $\rightarrow$ *topological diameter* of the molecule. Each fragment descriptor is represented by a vertex distance code called **structural environment vector (SEV)**, which identifies a particular atom-centered fragment that can occur more than once in the molecule. In other words, the different vertex distance codes in the molecule are collected in the cumulative path matrix, each identifying a particular atom-centered fragment. Then, all the different structural environment vectors in the data set are selected as fragment descriptors related to the studied biological response through a function of likeliness instead of the classical multivariate regression analysis.

Therefore, the potential influence for the biological activity of **SEV** is evaluated by using a likeliness function defined as

$$w_j(\alpha) = 1 - 4 \cdot \frac{(N_j^{\text{AC}} \cdot \alpha) \cdot (N_j^{\text{IN}} \cdot \alpha)}{(N_j^{\text{AC}} + N_j^{\text{IN}} + 2 \cdot \alpha)^2}$$

where N_j^{AC} and N_j^{IN} are the number of active and inactive compounds, containing the j th fragment represented by the j th SEV, respectively, in the data set; α is an adjustable parameter whose optimal value was determined to be in the 0.01–10 range. The larger the $w(\alpha)$ value of a fragment descriptor, the stronger its association is with the biological activity.

To establish the statistical significance of the likeliness function of each j th SEV, the binomial probability p_j is calculated as

$$p_j(m > N_j^{\text{AC}}) = \sum_{m=N_j^{\text{AC}}}^{N_j} \frac{N_j!}{m!(N_j-m)!} p^m q^{N_j-m}$$

where N_j^{AC} is the number of active compounds in which the j th fragment occurs and N_j the total number of compounds containing the j th fragment; the probabilities p and q are defined as

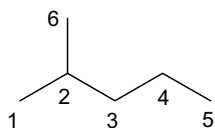
$$p = \frac{N^{\text{AC}}}{N^{\text{AC}} + N^{\text{IN}}} \quad q = 1 - p$$

where N^{AC} and N^{IN} are the total number of active and inactive compounds, respectively.

The general procedure is to examine all the possible structural features within the framework of an expert system. The potential structural features are ordered according to importance on the basis of the statistical parameters defined above. A heuristic algorithm is then followed, which selects the optimal topological structural features that account for the activity of compounds in which they occur. Moreover, once the significant SEVs most related to the biological activity have been found, $\rightarrow$ reversible decoding can be easily performed.

Example S1

Vertex distance counts and SEVs for 2-methylpentane. ${}^m f_i$ indicates the number of occurrences of distances equal to m from the i th vertex.



					2-methylpentane SEVs				
Atom	${}^1f_i$	${}^2f_i$	${}^3f_i$	${}^4f_i$	Fragment	${}^1f_i$	${}^2f_i$	${}^3f_i$	${}^4f_i$
1	1	2	1	1	1	1	2	1	1
2	3	1	1	0	2	3	1	1	0
3	2	3	0	0	3	2	3	0	0
4	2	1	2	0	4	2	1	2	0
5	1	1	1	2	5	1	1	1	2
6	1	2	1	1					

Note that vertices 1 and 6 have the same vertex distance count and, then, only five different SEVs are considered for further analysis.

• LUDI energy function

LUDI is a knowledge-based ligand design system based on a set of geometric rules derived from a statistical analysis of a series of small-molecule crystal structures [Böhm, 1992a, 1992b, 1994a, 1994b, 1998]. This system was designed to estimate the free energy of binding affinity of any ligand–receptor complex when the 3D structure of the complex is known or can be approximated. It can also be used to sort the compounds of a library according to their binding affinities, which are measure of the drug-like character of compounds.

The LUDI score is the binding affinity measure, calibrated using the binding constants from a database of 45 ligand–protein complexes and defined as

$$LUDI\ score \equiv S = -73.33 \cdot \Delta G$$

where

$$\Delta G = 1.3 - 1.1 \cdot \sum_{h\text{-bonds}} f_1(\Delta r) \cdot f_2(\Delta \alpha) - 2.0 \cdot \sum_{ionic} f_1(\Delta r) \cdot f_2(\Delta \alpha) - 0.040 \cdot SA_{lip} + 0.33 \cdot RBN$$

ΔG and the other equation coefficients are in kcal/mol units. The *h-bond* term represents the overall contribution of all the possible ligand–protein H-bonds, whereas the ionic term represents the contribution of all the possible ligand–protein ionic interactions. The SA_{lip} term is proportional to the lipophilic contact surface between the ligand and the receptor. The RBN term is the contribution to the free binding energy due to the freezing of internal degrees of freedom in the ligand, RBN being the number of rotatable bonds in the ligand [So and Karplus, 1999]. The offset term 1.3 kcal/mol represents the contribution due to nonspecific interactions of the molecule with the receptor (e.g., loss of translational and rotational entropy of the ligand).

The functions $f_1(\Delta r)$ and $f_2(\Delta \alpha)$ scale the strength of ligand–receptor interactions according to the deviation from ideal hydrogen-bond length (r) or angle (α). The two functions are defined as

$$f_1(\Delta r) = \begin{cases} 1 & \text{if } \Delta r \leq 0.25 \text{ \AA} \\ 1 - (\Delta r - 0.25 / 0.4) & \text{if } 0.25 < \Delta r \leq 0.65 \text{ \AA} \\ 0 & \text{if } \Delta r > 0.65 \text{ \AA} \end{cases}$$

$$f_2(\Delta \alpha) = \begin{cases} 1 & \text{if } \Delta \alpha \leq 30^\circ \\ 1 - (\Delta \alpha - 30 / 50) & \text{if } 30^\circ < \Delta \alpha \leq 80^\circ \\ 0 & \text{if } \Delta \alpha > 80^\circ \end{cases}$$

where Δr is the deviation of the H $\cdots$ O/N hydrogen-bond length from 1.85 Å and $\Delta \alpha$ is the deviation of the hydrogen-bond angle from its ideal value of 180°.

The LUDI score S ($\times 100$) corresponds to a binding constant K_i value of 10^{-S} M at 300 K.

A **modified LUDI energy function** was also proposed to estimate the free energy of binding for a protein–ligand complex [Eldridge, Murray *et al.*, 1997; Murray, Auton *et al.*, 1998].

This, as the LUDI energy function, includes simple terms to estimate lipophilic and hydrogen-bond contributions and a term that penalizes flexibility. Unlike the LUDI function, a term related to metal–ligand contributions is also accounted for.

The coefficients of each term are obtained by a regression analysis based on 82 ligand–receptor complexes for which the binding affinity is known. Before calculating this modified LUDI function, atom types are assigned to all ligand atoms and receptor atoms in

contact with the ligand. The selected atom types are (1) *lipophilic* (chlorine, bromine, and iodine atoms that are not ions; sulfurs that are not acceptor or polar types; and carbons that are not polar type); (2) *H-bond donor* (nitrogens with hydrogen attached, and hydrogens attached to oxygen or nitrogen); (3) *H-bond donor/acceptor* (oxygens attached to hydrogen atoms; special case of imine nitrogen (i.e., C=NH nitrogen)); (4) *H-bond acceptor* (oxygen not attached to hydrogen; nitrogens with no hydrogens attached and one or two connections; halogens that are ions; and sulfurs with only one connection (e.g., thioureas)); (5) *polar* (nitrogens with no hydrogens attached and more than two connections; phosphorus; sulfurs attached to one or more polar atoms (including H-bonding atoms and not including polar carbon atoms or fluorine atoms); carbons attached to two or more polar atoms (including H-bonding atoms and not including polar carbon atoms or fluorine atoms); carbons in nitriles or carbonyls; nitrogens with no hydrogens and four connections; and fluorine atoms); and (6) *metal* (i.e., metal atoms).

The obtained model is the following:

$$\Delta G = -5.48 - 3.34 \cdot \sum_{h\text{-bonds}} f_1(\Delta r) \cdot f_2(\Delta \alpha) - 6.03 \cdot \sum_{metal} f_3(\Delta r) - 0.117 \cdot \sum_{lip} f_4(\Delta r) + 2.56 \cdot H_{rot}$$

The term *h-bond* is that previously defined; the term *metal* is calculated for all acceptor and acceptor/donor atoms in the ligand and any metal atom in the receptor, the f_3 function being a contact term depending on the distance; and the term *lip* is calculated considering all the contributions of all the lipophilic atoms of the ligand and the receptor, the f_4 function being used to modulate the long-range interactions. The term H_{rot} is used to estimate the flexibility penalty for molecules possessing frozen rotatable bonds:

$$H_{rot} = 1 + \left(1 - \frac{1}{RBN}\right) \cdot \frac{\sum_r [P_{nl}(r) + P'_{nl}(r)]}{2}$$

where *RBN* is the number of rotatable bonds, the summation runs over the frozen rotatable bonds, and P_{nl} and P'_{nl} are the percentages of nonlipophilic heavy atoms on either side of each rotatable bond. *RBN* counts any sp^3-sp^3 and sp^2-sp^3 bonds, excluding CH_3 , CF_3 , NH_2 , and NH_3 groups. Moreover, bonds are considered frozen if atoms on both sides of the rotatable bond are in contact with the receptor, that is, the distance between any two heavy atoms is less than the sum of the relevant van der Waals radii plus 0.5 Å.

📖 [Botzki, Salmen *et al.*, 2005]

• biological activity profile score

This is a drug-like score derived by a $\rightarrow$ *substructural analysis* approach used to calculate *biological activity profiles*, which contain weights that describe the differential occurrences of selected molecular properties in active and inactive molecules [Gillet, Willett *et al.*, 1998].

Molecular properties are those thought of to affect (either positively or negatively) the tendency of a molecule to exhibit biological activity. Namely, these are the molecular weight (MW), the $\rightarrow$ *Kier shape descriptor* ($^2\kappa_\alpha$), the number of aromatic rings (AR), the number of rotatable bonds (*RBN*), the number of hydrogen-bond donors (*HBD*), and the number of hydrogen-bond acceptors (*HBA*). According to $\rightarrow$ *cell-based methods*, the range of values of each molecular property is divided into 20 equally sized bins. The bin size is set to unity for counting

descriptors such as *AR*, *RBN*, *HBD*, and *HBA*, so that each bin is associated a specific occurrence number of a feature; the first bin is associated a value of zero, whereas the last bin is associated all occurrence numbers equal to or greater than 19. For the molecular weight and the shape index, each bin represents a range of values: a bin size of 75 is chosen for the molecular weight and of 2 for Kier shape index.

Then, distributions for the occurrences of the various molecular properties in two sets of molecules, one of active molecules and the other of inactive molecules, are evaluated. From these frequency distributions, weights are calculated using one of the two weighting schemes, each of which seeks to quantify the differential occurrences of the defined ranges of property values in active (AC) and inactive (IN) molecules:

$$w_{jk} = \frac{N_{jk}^{AC}}{N_{jk}^{AC} + N_{jk}^{IN}} \quad w_{jk} = \ln \exp \left(\frac{N_{jk}^{AC}/N^{AC}}{N_{jk}^{IN}/N^{IN}} \right)$$

where w_{jk} is the weight for the k th bin of the j th property, N_{jk}^{AC} and N_{jk}^{IN} the number of active and inactive molecules containing the k th bin of the j th property, and N^{AC} and N^{IN} the total number of active and inactive molecules, respectively.

Each property is then represented by 20 weights and the vector containing the weights of all the considered molecular properties is the biological activity profile. A high weight indicates a high likelihood that a molecule having the given property value range will be an active molecule rather than an inactive molecule.

Finally, the drug-like score of the molecule is generated by first calculating the values of the molecular properties in the molecule and then retrieving the appropriate weights from the biological activity profile. Then, the score is obtained by summing the weights over all the considered properties. This summation implicitly assumes that the properties are statistically independent of each other.

A different scoring function was also proposed by other authors [Ajay, Walters *et al.*, 1998], still based on MW, $^2\kappa_a$, *AR*, *RBN*, *HBA*, and *HBD*, to which $\log P$ and $\rightarrow$ *ISIS keys* were added. This scoring function was developed by training a Bayesian neural network on a large database of known drugs and nondrug compounds.

Another simple scoring function for substituents, called **substituent drug-likeness index**, was proposed according to the following formula [Ertl, 2003]:

$$w_j = \log \left(\frac{N_j^{AC}/N^{AC}}{N_j^{IN}/N^{IN}} \right)$$

where w is the score for the j th substituent, N_j^{AC} and N_j^{IN} the number of active and inactive molecules containing the j th substituent, and N^{AC} and N^{IN} the total number of active and inactive processed molecules, respectively.

• Binary QSAR analysis

Binary QSAR analysis is an approach to screening of chemical libraries aimed at identifying possible lead compounds on the basis of a probability distribution function for active and inactive compounds [Labute, 1999; Gao, Williams *et al.*, 1999; Gao and Bajorath, 1999; Stahura, Godden *et al.*, 2000; Gao, 2001; Gao, Lajiness *et al.*, 2002; Stahura, Godden *et al.*, 2002; Godden and Bajorath, 2003].

In binary QSAR, the biological activity, expressed in a binary form (1 for active and 0 for inactive) is correlated with molecular descriptors of compounds, and a probability distribution for active and inactive compounds in a training set is estimated. The derived binary QSAR model can subsequently be used to predict the probability of new compounds to be active against a given biological target.

The probability density $p(Y = 1 | \mathbf{X} = \mathbf{x})$ of a compound, represented by the descriptor vector $\mathbf{x}$, to be active is calculated according to

$$p(Y = 1 | \mathbf{X} = \mathbf{x}) \approx \left[1 + \frac{p(Y = 0)}{p(Y = 1)} \cdot \prod_{m=1}^M \frac{p(Z_m = z_m | Y = 0)}{p(Z_m = z_m | Y = 1)} \right]^{-1}$$

where Y is a Bernoulli random variable representing the biological activity and $\mathbf{X}$ a random p -dimensional vector comprising of p physico-chemical properties and/or molecular descriptors.

To obtain a set of orthogonal and standardized variables, the original p molecular descriptors x are transformed into M variables z by applying principal component analysis; M is the number of chosen significant components. Then, each probability density $p(Z_m = z_m)$ is estimated by constructing a histogram from the training set by representing each compound with a Gaussian density function with a given variance. To express the biological activity Y in a binary form ($Y = 1$ or $Y = 0$), a threshold value has to be chosen.

• Multilevel Chemical Compatibility (MLCC)

Multilevel chemical compatibility is a measure of drug likeness of a compound derived from the comparison of the compound with a drug library [Wang and Ramnarayan, 1999].

A compound is described by a vector containing the occurrence numbers of **Local Chemical Environments** that are $\rightarrow$ *substructure descriptors* defined using single atoms as well as dicentered, tricentered, and tetracentred groups of atoms. An n -centered group is defined as a molecular fragment containing n -connected non-hydrogen atoms plus their first neighboring atoms. Atoms are distinguished on the basis of their chemical elements; hybridization state is accounted only for carbon, oxygen, and nitrogen atoms.

First, the atom types within a molecule are assigned; then, n -centered groups are searched for, n varying from 1 to 4, and the occurrence frequency of each is recorded in the final descriptor vector.

A set of unique group types is derived from analysis of all the molecules within a reference drug library. For each group type, three quantities are calculated: the minimum frequency w_1 that is the occurrence number in the library molecule in which the group type appears the smallest number of times, the maximum frequency w_2 that is the occurrence number in the library molecule in which the group type appears the largest number of times, and the fraction of molecules containing the group type. Then, to estimate the drug likeness of a compound, all its n -centered groups are compared with the group types in the drug library. A group of a compound is considered compatible with the drug library if its frequency of occurrence in the compound is between the corresponding w_1 and w_2 in the drug library. If all the groups at level n are compatible with the library, level n is called *compatible level*.

Finally, the MLCC score of a compound is defined as the highest level n at which the compound is found compatible with the reference drug library, or zero if no compatibility is found at any level. This method was tested by using 11 704 drugs from CMC and MDDR databases.

- **Property and Pharmacophore Features Score (PPFS)**

Property and Pharmacophore Features fingerprints (or *PPF* fingerprints) encode information about eight molecular descriptors: molecular weight, CLOGP, number of rings, number of hydrogen-bond donors, number of hydrogen-bond acceptors, number of positively charged groups, number of negatively charged groups, and number of rotatable bonds [Oprea, 2000].

By using the same approach as the → *cell-based methods*, each molecular descriptor is mapped into a binary vector of 30 bits length, each bit corresponding to a certain value range of the molecular descriptor. Then, by concatenation of the individual property vector, the PPF fingerprint is obtained consisting of 240 bits. Each bit is treated as a separate variable in a linear model evaluated by PLS Discriminant Analysis (PLS-DA) applied on a large training set comprising of drug and nondrug compounds. The positive and negative model coefficients represent the specific contributions of feature ranges to drug likeness. This PLS-DA model is a scoring function, which was called Property and Pharmacophore Features Score (*PPFS*).

A similar scoring function calculated by PLS-DA applied to the → *Daylight fingerprints* was called **Daylight-FingerPrint drug-like Score (DFPS)** [Oprea, 2000].

- **Xu–Stevenson Drug-Like Index (DLI)**

Developed using 4836 compounds from the CMC database, DLI is defined in terms of 25 structural descriptors defining the → *chemical space* and based on the idea that drug-like compounds cluster into a certain region of this chemical space [Xu and Stevenson, 2000].

To calculate the drug-like index for a compound, a twofold procedure is used: (1) The drug-like cluster center is computed for a drug database. (2) Based on the 25 molecular descriptor values of the compound, the DLI is calculated by means of comparing its molecular descriptor values with the drug-like cluster center. The drug-like cluster is defined by the parameters of the 25 molecular descriptors given in Table S2. Then, DLI is calculated as

Table S2 Xu–Stevenson drug-like index.

ID	Molecular descriptor	Best	Min	Max
1	Non-hydrogen atoms	22	10	50
2	Number of rings	3	1	6
3	Molecular cyclized degree (%)	10%	4%	18%
4	Rotatable bonds	9	3	35
5	Polar bonds	8	1	25
6	Terminal methyl groups	0	0	7
7	Amino H-bond donors	0	0	3
8	Hydroxyl H-bond donors	0	0	4
9	H-bond donors	1	0	5
10	H-bond acceptors	3	0	10
11	Oxygen + nitrogen atoms	4	0	15
12	2-Degree acyclic atoms (vertex degree)	1	0	12
13	Unsubstituted acyclic atoms	8	1	19
14	3-Degree acyclic atoms (vertex degree)	1	0	5
15	Substituted acyclic atoms	6	1	15
16	1-Level bonding pattern	0	0	5
17	2-Level bonding pattern	0	0	5

(Continued)

Table S2 (Continued)

ID	Molecular descriptor	Best	Min	Max
18	3-Level bonding pattern	0	0	4
19	Building blocks	1	4	11
20	Aromatic systems	1	0	3
21	Cyclic building blocks	2	1	5
22	Linkers	1	0	3
23	Caps	2	0	8
24	Max ring size	6	3	13
25	Max cap size (no. of atoms)	1	0	12

A *cap* (substituent or side chain) is an acyclic substructure with one attachment connecting to other structure building blocks; a *linker* is an acyclic substructure with more than one attachment connecting to other structure building blocks; *building blocks* are functional groups and core structures; and a *core* is a cyclic substructure without linker or cap.

$$DLI = \sqrt[25]{\prod_{j=1}^{25} score(descriptor_j)}$$

where the summation runs over the structural descriptors and the score seems to be inversely proportional to a standardized distance of the compound from the cluster center.

Being DLI based on a geometric mean, even when only one score is zero, the entire DLI becomes zero.

• Pharmacophore Point Filter (PF)

This filter is based on a set of simple rules defined to classify molecules as drug-like or nondrug-like on the basis of their functional groups [Muegge, Heald *et al.*, 2001; Muegge, 2002, 2003].

The considered functional groups are those having some hydrogen-bonding capabilities that are essential for specific drug interactions with the target. These are called pharmacophore points and include amine, amide, alcohol, ketone, sulfone, sulfonamide, carboxylic acid, carbamate, guanidine, amidine, urea, and ester.

A score is calculated for each compound by counting the pharmacophore points present in the compound. A score between 2 and 7 identifies drug-like compounds.

In a modified version of the pharmacophore point filter, a score equal to 1 is allowed to qualify a compound as drug-like if the pharmacophore point present in the compound is of type carboxylic acid, amine, guanidine, or amidine. This modified filter was defined to correctly classify the small active drugs belonging to the central nervous system (CNS) class.

The pharmacophore point filter also includes the following rules:

- (1) Pharmacophore points are fused and counted as one when their heteroatoms are not separated by more than one carbon atom.
- (2) Pyrrole, indole, thiazole, isooxazole, other azoles, or diazines are not considered pharmacophore points.
- (3) Compounds with more than one carboxylic acid are dismissed.

- (4) Compounds without a ring structure are dismissed.
- (5) Intracyclic amines that occur in the same ring (e.g., piperazines) are fused and counted one pharmacophore point.

• PC-based drug-like index

It is derived from the third principal component PC_3 calculated by applying $\rightarrow$ *Principal Component Analysis* on Maybridge library compounds described by 26 molecular descriptors [Brüstle, Beck *et al.*, 2002]:

$$\Delta = PC_3 - 1.15$$

where PC_3 is the linear combination of the 26 molecular descriptors, whose coefficients are the loadings given in Table S3. This drug-like score is positive for most drugs and negative for most of the nondrugs. The third component explains 9.1% of the whole variance of the data.

Table S3 Loadings of the third principal component from [Brüstle, Beck *et al.*, 2002].

Descriptor	PC_3 loading	Descriptor	PC_3 loading
Dipole moment	−0.0130	Min MEP	−0.4189
Dipolar density	−0.0186	Mean + MEP	−0.2571
Polarizability	0.0528	Mean − MEP	−0.3062
SMEP(H)	0.2837	Variance	0.2627
SMEP(N)	−0.1263	Balance parameter	−0.4740
SMEP(O)	−0.0507	Variance $\times$ balance parameter	−0.1071
SMEP(P)	0.1507	HBA	0.0381
SMEP(S)	0.0246	HBD	0.0510
SMEP(F)	0.3908	AR	−0.0581
SMEP(Cl)	−0.1323	MW	−0.01093
SMEP(Br)	−0.0376	Volume	0.0318
SMEP(I)	0.0069	Total surface area	0.0150
Max MEP	−0.1726	Globularity factor	0.0301

MEP, molecular electronic potential; SMEP(X), sum of the electrostatic potential-derived atomic charges on atom X; Mean, average MEP values; Variance, variance of MEP values; HBA, number of hydrogen-bond acceptors; HBD, number of hydrogen-bond donors; AR, number of aromatic rings; and MW, molecular weight.

• Monge–Arrault–Marot–Morin–Allory scoring functions

These scoring functions, *MAMMA scoring functions*, were derived from the analysis of 2.6 million of compounds, collected from 32 chemical databases [Monge, Arrault *et al.*, 2006].

Before applying the scoring algorithm, a preliminary screening was performed on the atom types to exclude compounds with atoms other than C, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, or Li. Then, the following eight molecular properties were calculated: molecular weight, $\log P$, number of hydrogen-bond acceptors and donors, number of rotatable bonds, number of halogens, number of rings, and maximum ring size.

For each molecular property, a penalty was calculated depending on the actual property value; penalty varies from 0 to 1. The functions to calculate the property penalties are shown in Table S4.

Table S4 MAMMA drug-like and lead-like scores. x is the value of the considered property.

Molecular descriptor	Range	Drug-like penalty	Range	Lead-like penalty
Molecular weight (MW)	$x \leq 100$	1	$x \leq 100$	1
	(100; 150]	$-0.02 x + 3$	(100; 150]	$-0.02 x + 3$
	(150; 350]	0	(150; 322]	0
	(350; 800)	$0.0022 x - 0.7778$	(322; 588)	$0.0038 x - 1.2105$
$\log P$	$x \geq 800$	1	$x \geq 588$	1
	$x \leq -5$	1	$x \leq -5$	1
	(-5; -1.5)	$-0.2857 x - 0.4286$	(-5; -1.5)	$-0.2857 x - 0.4286$
	[-1.5; 4.5]	0	[-1.5; 2.94]	0
	(4.5; 7.5)	$0.3333 x - 1.5$	(2.94; 5.46)	$0.3968 x - 1.667$
	$x \geq 7.5$	1	$x \geq 5.46$	1
H-bond acceptors	$x \leq 7$	0	$x \leq 6.3$	0
	(7; 13)	$0.1667 x - 1.1667$	(6.3; 11.7)	$0.1852 x - 1.1667$
	$x \geq 13$	1	$x \geq 11.7$	1
H-bond donors	$x \leq 3.5$	0	$x \leq 3.5$	0
	(3.5; 6.5)	$0.3333 x - 1.1667$	(3.5; 6.5)	$0.3333 x - 1.1667$
	$x \geq 6.5$	1	$x \geq 6.5$	1
Rotatable bonds	$x \leq 10.5$	0	$x \leq 7$	0
	(10.5; 19.5)	$0.1111 x - 1.1667$	(7; 13)	$0.1667 x - 1.1667$
	$x \geq 19.5$	1	$x \geq 13$	1
No. of halogens	$x \leq 4.9$	0	$x \leq 4.9$	0
	(4.9; 9.1)	$0.2381 x - 1.1667$	(4.9; 9.1)	$0.2381 x - 1.1667$
	$x \geq 9.1$	1	$x \geq 9.1$	1
No. of rings	$x \leq 4.2$	0	$x \leq 2.8$	0
	(4.2; 7.8)	$0.2778 x - 1.1667$	(2.8; 5.2)	$0.4167 x - 1.1667$
	$x \geq 7.8$	1	$x \geq 5.2$	1
Maximum ring size	$x \leq 6$	0	$x \leq 6$	0
	(6; 9.1)	$0.3226 x - 1.9355$	(6; 9.1)	$0.3226 x - 1.9355$
	$x \geq 9.1$	1	$x \geq 9.1$	1

The score of a compound is calculated by the sum of these penalties corresponding to the actual property values x of the compound.

A low score (<1) indicates a molecule that can be considered as drug-like or lead-like, whereas a score equal to or greater than 2 means that the compound is not drug-like or lead-like. Moreover, before calculating the score, the following filters have to be passed: no reactive functions, alkyl chains $\leq -(\text{CH}_2)_6\text{CH}_3$, perfluorinated chains smaller than $-\text{CF}_2\text{CF}_2\text{CF}_3$, and at least one oxygen or nitrogen.

• Hutter likeliness score

This scoring function is derived from a statistical approach based on the distributions of $\rightarrow$ atom pairs [Hutter, 2007]. The assumption is that certain atom pair combinations occur with a different frequency in drug-like molecules compared to nondrug molecules. Before applying the scoring algorithm, atom types are assigned to molecule atoms; hydrogens are not considered. Then, the total likeliness score for atom pair combinations up to six interactions is defined as

$$L = \frac{1}{\Delta} \cdot \sum_{k=0}^5 \left[\sum_{i=1}^n \sum_{j=1}^n \begin{cases} D_{ij}^k & \text{if } i-j \text{ in molecule} \\ 0 & \text{otherwise} \end{cases} \right]$$

where Δ is the number of the considered atom pairs in the molecule and n the number of defined atom types (e.g., 47 atom types from HyperChem software). The first summation runs over atom pairs at topological distances varying from 0 to 5. The term in square brackets is the likeliness score defined for a given atom pair combination; for $k = 0$, the likeliness score refers to the distributions of single atom types, for $k = 1$, to pairs of bonded atoms, for $k = 2$, to pairs of atom types at distance equal to 2, and so forth. For each k , the term D_{ij}^k is defined for a pair of atom type i and atom type j , as

$$D_{ij} = S_{ij}^{\text{AC}} - S_{ij}^{\text{IN}}$$

where AC and IN stand for active and inactive compounds, respectively. Positive D_{ij} values indicate that the i - j atom pair is more represented in the active class than in the inactive class. For each class of compounds (active or inactive), S_{ij} is the log odds score of the i - j pair, defined as

$$S_{ij} = \ln \frac{p_{ij}}{p_i \cdot p_j}$$

where p_i and p_j are the relative frequencies of finding the i th and j th atom types in the molecules of the considered activity class, respectively; p_{ij} is the relative frequency of finding a pair of atoms of type i and j , calculated as the sum of all occurrences of the i - j pair in all the class molecules divided by the sum of the occurrences of all the possible combinations of two atom types.

• consensus binding free energy

Consensus binding free energy is a scoring function for screening of compound libraries proposed in the framework of 4D-QSAR analysis [Esposito, Hopfinger *et al.*, 2003]. The idea is to use a collection of QSAR models to create a unique consensus model for binding affinity prediction. Thus, this consensus binding free energy is evaluated for each molecule according to the following:

$$\Delta G_i^c = \sum_k w_k \cdot \Delta G_{ik}$$

where ΔG_{ik} is the binding energy of the i th compound predicted by the k th model and w_k is a weighting factor related to the relative significance of each k th model. The weighting factor can be calculated as

$$w_k = \frac{R_k^2}{\sum_m R_m^2}$$

where R^2 is the coefficient of determination of the model and the summation runs over the considered models for the $\rightarrow$ consensus analysis.

📖 [Sadowski and Kubinyi, 1998; Böhm and Stahl, 2000; Frimurer, Bywater *et al.*, 2000; Wagener and van Geerestein, 2000; Gálvez, Julian-Ortiz *et al.*, 2001; Byvatov, Fechner *et al.*, 2003; Takaoka, Endo *et al.*, 2003; Mpamhanga, Chen *et al.*, 2005; Tame, 2005; Mpamhanga, Chen *et al.*, 2006]

- **secondary mesomeric effects** $\rightarrow$ electronic substituent constants
- **second-grade structural parameters** $\rightarrow$ multiple bond descriptors

- **second-order submolecular polarity parameter** → charge descriptors (⊙ submolecular polarity parameter)
- **second path matrix** → Laplacian matrix
- **second Zagreb index** → Zagreb indices
- **SEDI** $\equiv$ *shared-electron distribution index* → quantum-chemical descriptors (⊙ electron density)
- **selective PLS** → variable selection (⊙ intermediate least squares regression)
- **self-atom polarizability** → electric polarization descriptors (⊙ atom-atom polarizability)
- **self-avoiding walk** $\equiv$ *path* → graph

■ **Self-Organizing Maps (SOM)** ($\equiv$ *Kohonen maps, Kohonen artificial neural networks*)

Kohonen maps are self-organizing systems able to face the unsupervised rather than the supervised problems [Kohonen, 1989, 1990].

In Kohonen maps (Figure S1), similar objects are linked to the topologically close neurons in the network, that is, neurons that are located close to each other have similar reactions to similar inputs, whereas neurons that are far apart have different reactions to similar inputs.

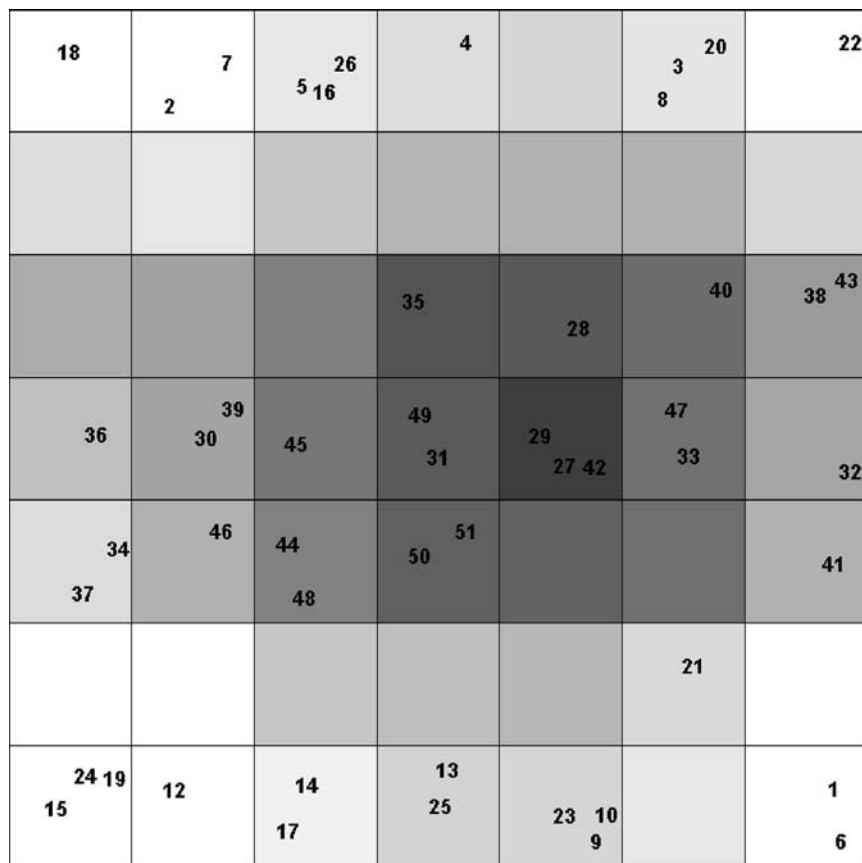


Figure S1 A Kohonen map of size (7×7) , showing the distribution of the objects; the cell colors represent the intensity of a variable (white color for minimum values and black for maximum values).

The Kohonen layer is usually characterized by being a square toroidal space and consists of a grid of $N \times N$ neurons, each containing as many elements (weights) as the number of input variables; the weights of each neuron are randomly initialized between 0 and 1 and updated on the basis of the input vectors for a certain number of times (called training epochs) (Figure S2). In each training step, for each input vector, a winning neuron (the neuron most similar to the input vector) is selected. Then, the weights of each neuron r ($\mathbf{w}_r$) are changed on the basis of the difference between their old values and the values of the input vector ($\mathbf{x}_i$); this correction ($\Delta \mathbf{w}_r$) is scaled according to the topological distance from the winning neuron (d_r):

$$\Delta \mathbf{w}_r = \eta \cdot \left(1 - \frac{d_r}{d_{\max} + 1}\right) \cdot (\mathbf{x}_i - \mathbf{w}_r^{\text{old}})$$

where η is the learning rate and $d_{\max}$ the maximum size of the considered neighborhood. The topological distance d_r is defined as the number of neurons between the neuron r and the winning neuron. The learning rate η changes during the training phase, as follows:

$$\eta = (\eta^{\text{start}} - \eta^{\text{final}}) \cdot \left(1 - \frac{t}{t_{\text{tot}}}\right) + \eta^{\text{final}}$$

where t is the number of the current training epoch, t_{tot} the total number of training epochs, and η^{start} and η^{final} are the learning rates at the beginning and at the end of the training, respectively. Values of η^{start} and η^{final} equal to 0.5 and 0.01, respectively, are commonly used.

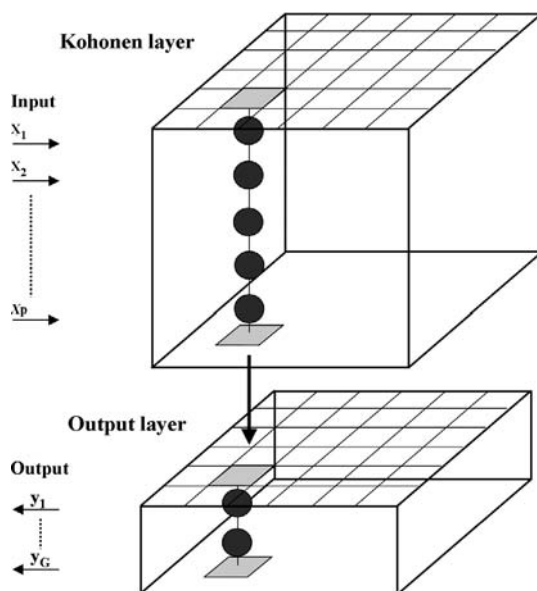


Figure S2 Kohonen network built by a 6×6 map (36 neurons) and an associated counter-propagation neural network for a two-class problem.

Several QSAR approaches are based on Kohonen maps, such as $\rightarrow$ *topological feature maps*, $\rightarrow$ *Comparative Molecular Surface Analysis*, and $\rightarrow$ *MOLMAP descriptors*.

Counter-propagation neural network is a development of Kohonen maps for classification purposes [Zupan, Novič *et al.*, 1995], which considers a set of output layers, called Grosberg

layer, in addition to the Kohonen layer. The number of the Grosberg layers is equal to the number of classes, since class belonging of objects is represented by a binary vector whose g th element is equal to 1 if the object belongs to the g th class, and zero otherwise. In other words, a class unfolding is performed during $\rightarrow$ *data set* pretreatment (Figure S2).

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **Self-Organizing Molecular Field Analysis** $\rightarrow$ grid-based QSAR techniques
- **self-returning ID number** $\rightarrow$ ID numbers ($\odot$ weighted ID number)
- **self-returning walk atomic code** $\rightarrow$ self-returning walk counts
- **self-returning walk** $\rightarrow$ graph

■ self-returning walk counts

These are a particular case of $\rightarrow$ *walk counts*. Self-returning walk counts are atomic and molecular descriptors obtained from a $\rightarrow$ *H-depleted molecular graph*, based on graph walks starting and ending at the same vertex, that is, $\rightarrow$ *self-returning walks (SRWs)* [Harary, 1969a].

The length k of a walk is the total number of edges that are traversed, repeated use of the same edge or edges being allowed.

The **atomic self-returning walk count** of k th order, denoted by $srw_i^{(k)}$, is the number of walks of length k starting and ending at the i th vertex. It is easily obtained by the k th power of the $\rightarrow$ *adjacency matrix* A ; in effect, each diagonal element of A^k can be interpreted as the sum of all self-returning walks of length k for a given vertex:

$$srw_i^{(k)} = [A^k]_{ii}$$

The **molecular self-returning walk count** of k th order is the total number of self-returning walks of length k in the graph and is simply calculated by summing up all of the atomic self-returning walk counts of the same order:

$$srw^{(k)} = \sum_{i=1}^A srw_i^{(k)} = tr(A^k)$$

where A is the number of vertices in the graph and tr the $\rightarrow$ *trace* of the k th power of the adjacency matrix, that is, the sum of the diagonal elements. From the above relation it is derived that molecular self-returning walk counts are the **spectral moments of the adjacency matrix**, which were also expressed as linear combinations of counts of certain fragments contained in the molecular graph, that is, $\rightarrow$ *embedding frequencies* [Barysz and Trinajstić, 1984; Jiang, Tang *et al.*, 1984; Kiang and Tang, 1986; Jiang and Zhang, 1989, 1990; Marković and Gutman, 1991; Jiang, Qian *et al.*, 1995; Marković and Stajkovic, 1997; Marković, 1999].

It has to be noted that the self-returning walk count of order 2 for the i th atom coincides with its $\rightarrow$ *vertex degree* δ_i , that is, the number of bonded atoms; moreover, the molecular self-returning walk counts of first and second order coincide with the number of atoms A and twice the number of bonds B in the molecule, respectively.

For the i th atom of a molecule, the sequence of atomic self-returning walk counts of increasing length up to A defines the **self-returning walk atomic code (SRWAC)**:

$$\{srw_i^{(1)}, srw_i^{(2)}, srw_i^{(3)}, \dots, srw_i^{(A)}\}$$

This vectorial descriptor characterizes each atom in the molecule [Randić, 1980c]. It can be noted that for acyclic graphs, only the counts of SRWs of even length are different from zero; for any graph, the first term in the code is equal to one.

Vertices in a molecular graph having the same SRWAC are called **endospectral vertices** and the corresponding graph **endospectral graph**. Attaching any subgraph to each endospectral vertex one at a time generates $\rightarrow$ *isospectral graphs*, that is, graphs with the same $\rightarrow$ *characteristic polynomial*.

Graphs with identical SRWACs for all the atoms are $\rightarrow$ *isocodal graphs* [Ivanciuc and Balaban, 1996b].

By summing up all the entries of the self-returning walk count atomic code, a different local invariant called **vertex structural code (SC, or structural code)** is obtained:

$$SC_i = \sum_{k=1}^A srw_i^{(k)}$$

where k is the increasing length of the walks [Barysz, Trinajstić *et al.*, 1983].

Based on these local invariants, the **ordered structural code (OSC)** is a molecular descriptor defined as the ascending ordered sequence of SC_i in the molecule [Barysz and Trinajstić, 1984]:

$$OSC = \{SC_{i(1)}, SC_{i(2)}, SC_{i(3)}, \dots, SC_{i(A)}\}$$

where A is the number of atoms in the molecule and the numbers in parenthesis represent the ordered sequence. The OSC vectors of different size molecules are transformed into $\rightarrow$ *uniform length descriptors* by selecting a dimension L equal to the maximum number of the atoms in the largest molecule of the data set and by adding zero to the empty positions.

Weighted atomic self-returning walk counts of any order k , denoted by ${}^w srw_i^{(k)}$, were proposed [Bonchev and Kier, 1992] by summing the weights of all self-returning walks SRW_i^k of a given vertex:

$${}^w srw_i^{(k)} = \sum_j w(SRW_i^k)_j$$

where w is the weight of the j th self-returning walk of length k for the i th vertex, which can be calculated as

$$w(SRW_i^k)_j = 2 \cdot \sum_{b \in j} (\delta_{b(1)} \cdot \delta_{b(2)})^{1/2} \quad \text{or} \quad w(SRW_i^k)_j = \prod_{b \in j} (\delta_{b(1)} \cdot \delta_{b(2)})$$

The first weight is obtained by summing the square root product of the vertex degrees δ of all adjacent vertices along the considered walk; the sum is multiplied by 2, each edge being traversed twice. The second self-returning walk weight is calculated by multiplying the weights of all edges along the walk, each edge weight being the product of the vertex degrees of the incident vertices. These local vertex invariants were used for the $\rightarrow$ *canonical numbering* of graph vertices. Also other $\rightarrow$ *weighting schemes* for self-returning walks have been proposed to improve QSAR modeling [Bonchev, Liu *et al.*, 1993].

Self-returning walk counts are indices derived from molecular topology that are closely related to the moments of energy derived from quantum-chemistry [Bonchev and Kier, 1992; Bonchev, Kier *et al.*, 1993; Bonchev and Gordeeva, 1995; Gutman, Bonchev *et al.*, 1995], as is explained below.

Molecular moment of energy of k th order is defined as

$$\mu^k = \sum_i E_i^k = tr(\mathcal{H}^k)$$

where E_i indicates the energy levels constituting the discrete spectrum of a molecule, $\mathcal{H}$ the Hamiltonian matrix, and tr the $\rightarrow$ trace of the matrix. Moreover, the trace of the k th power of the Hamiltonian matrix is equal to the count of the weighted molecular self-returning walks SRW^k of order k , beginning and ending at the same orbital. The weights associated with the walks are the interaction integrals $H_{ia,ib}$ involving the overlapping orbitals ia and ib . The simplest weighting results from the one-electron Hückel method in which all $H_{ia,ib} = \beta$ (the resonance integral), if ia and ib are p -orbitals located on atoms of the π -bond network. For such systems, the **atomic moments of energy** μ_i^k are defined as the k th molecular moment of the i th orbital, that is, the number of walks that start at this orbital and return to it in k steps, traversing one bond in each step. From these local vertex invariants, the corresponding molecular descriptor is derived:

$$\mu^k = \sum_{i=1}^A \mu_i^k = \beta^k \cdot \sum_{i=1}^A srw_i^{(k)} = \beta^k \cdot srw^{(k)}$$

where $srw^{(k)}$ is the total number of self-returning walks of length k in the molecule.

To avoid rapidly increasing numbers with molecule dimension and branching, **relative atomic moments (RAMs)** are defined, normalizing each atomic moment as

$$f_i^k = \frac{\mu_i^k}{\mu^k} = \frac{srw_i^{(k)}}{srw^{(k)}}$$

that is, dividing each atomic term by the corresponding k th order molecular term.

It was demonstrated that RAMs always reach a limit with the increasing power k , and this limit is numerically equal to the partial charge of the *Lowest Occupied Molecular Orbital* (LOMO), which is the most stable orbital:

$$TAC_i \equiv f_i = \lim_{k \rightarrow \infty} (f_i^k) = \lim_{k \rightarrow \infty} \left(\frac{srw_i^{(k)}}{srw^{(k)}} \right) = c_{i,LOMO}^2$$

where $c_{i,LOMO}^2$ is the square coefficient of the first molecular orbital of the i th atom. The second equality has not been demonstrated, but is believed to have general validity.

This limit is a fractional topological charge f_i , called **topological atomic charge (TAC)**, and represents the relative occurrence of the SRWs of the i th atom to all SRWs in the molecule; it can be considered as the fractional topological charge of an atom, assuming that each SRW is associated with the movement of an electron near the nucleus of the considered atom; the larger the number of SRWs of an atom, the larger the topological charge ascribable to that atom, that is, the measure of the time an electron moves near the atom.

Topological atomic charges provide a $\rightarrow$ *canonical numbering* of graph vertices, each vertex being ordered according to its branching centrality and cyclicity, that is, according to its $\rightarrow$ *molecular complexity*.

Moreover, **topological atomic valencies (TAVs, or corrected second moments)** were derived from fractional atomic charges f_i by rescaling the values with respect to the second-order molecular self-returning walk count srw^2 :

$$TAV_i = f_i \cdot srw^2$$

where srw^2 represents the total number of σ -electrons taking part in the molecular graph.

Topological valence is a real local invariant close to the chemical valence of the atom. TAVs can be interpreted as atomic valencies corrected by accounting for all higher order atom-atom connectivities; atoms with low TAV values are regarded to as possessing large free valence.

Analogous to the topological atomic charge definition is the **topological bond order**, defined as the limit of

$$p_{ij}^{(k)} = \frac{\mu_i^k + \mu_j^k}{\mu^k} = \frac{srw_i^{(k)} + srw_j^{(k)}}{srw^{(k)}}$$

that is, as

$$p_{ij} = \lim_{k \rightarrow \infty} \left(\frac{srw_i^{(k)} + srw_j^{(k)}}{srw^{(k)}} \right)$$

This definition of $\rightarrow$ *bond order* can be regarded as an extension of the occurrence of the double bonds in all the Kekulé structures to the occurrence of a bond in all self-returning walks [Bonchev and Gordeeva, 1995].

The topological bond order thus defined can be interpreted as the one-electron distribution over the bonds in a molecule, analogous to the one-electron distribution over all atoms given by the topological atomic charges:

$$\sum_i f_i = \sum_b (p_{ij})_b = 1$$

where the first summation runs over the *A* atoms and the second over the *B* bonds.

📖 [Barysz, Bonchev *et al.*, 1986; Shalabi, 1991; Bonchev and Seitz, 1995]

- **self-returning walk ordering** $\rightarrow$ canonical numbering
- **self-returning walk-sequence matrix** $\rightarrow$ sequence matrices
- **self-returning ID number** $\rightarrow$ ID numbers
- **self-return probabilities** $\rightarrow$ MARCH-INSIDE descriptors
- **SE-MFP descriptors** $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **semiempirical molecular connectivity terms** $\rightarrow$ combined descriptors
- **semiempirical topological index** $\rightarrow$ chromatographic descriptors
- **sensitivity** $\rightarrow$ classification parameters

■ sequence matrices (SM)

A sequence matrix **SM** of a $\rightarrow$ graph *G* is a rectangular unsymmetrical matrix $A \times K$, whose entry $i-k$, denoted as (sm_{ik}) , is the number of walks of increasing length k ($k = 0, 1, \dots, K$) starting from the i th vertex to all the other $A - 1$ vertices. K is the maximum length of the walk and depends on the type of considered $\rightarrow$ walk [Diudea, 1994]. The label m denotes the type of walk, d being for the shortest path (topological distance), p for path, w for random walk, and srw for self-returning walk. The maximum length K of the specified walk is the $\rightarrow$ *topological diameter D* if shortest paths between vertices are considered, the maximum $\rightarrow$ *path eccentricity* in the molecule if the paths are considered, and an arbitrary chosen number between one and infinite (however, usually limited to $A - 1$ according to the Cayley–Hamilton theorem) if walks or self-returning walks are considered.

Therefore, different sequence matrices can be obtained using different types of walks. Note that the sequence matrix based on $\rightarrow$ *topological distances d*, here called **distance-sequence matrix SD**, coincides with the $\rightarrow$ *cardinality layer matrix LC* and has also been proposed by other authors as $\rightarrow$ *λ matrix* [Skorobogatov and Dobrynin, 1988] and $\rightarrow$ *F matrix* [Diudea and Pârv, 1988].

Moreover, the **path-sequence matrix SP** based on paths p was previously proposed as τ **matrix** or **path-layer matrix** [Skorobogatov and Dobrynin, 1988; Yang, Lin *et al.*, 2002].

The entries of the distance-sequence matrix **SD** are simply the $\rightarrow$ *vertex distance counts* $^1f_i, ^2f_i, ^3f_i, \dots, ^Df_i$, that is, the frequencies of distances equal to 1, 2, 3, ..., respectively, from vertex v_i to any other vertex, while each entry in the path-sequence matrix **SP** is the $\rightarrow$ *atomic path count* ^kP_i , the $\rightarrow$ *atomic walk count* $awc_i^{(k)}$ in the **walk-sequence matrix SW** and the $\rightarrow$ *atomic self-returning walk count* $swc_i^{(k)}$ in the **self-returning walk-sequence matrix SSRW**. Note that for $k=0$, the entries sm_{i0} are all equal to one independently of the kind of walk.

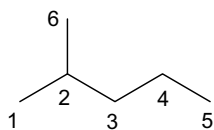
Applying the $\rightarrow$ *row sum operator* to the sequence matrices, $\rightarrow$ *local vertex invariants*, called **atomic sequence count** asc_i , are obtained as

$$asc_i = \sum_{k=1}^K sm_{ik}$$

where asc_i coincides with the $\rightarrow$ *atomic path count sum* P_i in the path-sequence matrix **SP**, the $\rightarrow$ *atomic walk count sum* awc_i of the i th atom, that is, the total number of walks of any length starting from v_i , in the walk-sequence matrix **SW**, and the $\rightarrow$ *vertex structural code* SC_i in the self-returning walk-sequence matrix **SSRW**.

Example S2

Distance-sequence matrix **SD**, path-sequence matrix **SP**, walk-sequence matrix **SW**, and self-returning walk-sequence matrix **SSRW** for 2-methylpentane. asc_i and msc^k indicate the matrix row and column sums, respectively.



$$TSC_D \equiv TSC_P = \sum_{k=0}^4 msc^k = 6 + \frac{1}{2} \cdot \sum_{i=1}^6 asc_i = 21$$

SW =

Atom/walk	0	1	2	3	4	5	asc_i
1	1	1	3	4	11	15	34
2	1	3	4	11	15	40	73
3	1	2	5	7	18	25	57
4	1	2	3	7	10	25	47
5	1	1	2	3	7	10	23
6	1	1	3	4	11	15	34
msc^k	6	5	10	18	36	65	140

SD $\equiv$ **SP** =

Atom/walk	0	1	2	3	4	asc_i
1	1	1	2	1	1	5
2	1	3	1	1	0	5
3	1	2	3	0	0	5
4	1	2	1	2	0	5
5	1	1	1	1	2	5
6	1	1	2	1	1	5
msc^k	6	5	5	3	2	21

SSRW =

Atom/walk	0	1	2	3	4	5	asc_i
1	1	0	1	0	3	0	4
2	1	0	3	0	10	0	13
3	1	0	2	0	7	0	9
4	1	0	2	0	5	0	7
5	1	0	1	0	2	0	3
6	1	0	1	0	3	0	4
msc^k	6	0	5	0	15	0	26

$$TSC_W = \sum_{k=0}^5 msc^k = 6 + \frac{1}{2} \cdot \sum_{i=1}^6 asc_i = 140 \quad TSC_{SRW} = \sum_{k=0}^5 msc^k = 6 + \frac{1}{2} \cdot \sum_{i=1}^6 asc_i = 26$$

Applying the $\rightarrow$ column sum operator CS (divided by 2) to any sequence matrix $\mathbf{SM}$, a global index for each k value different from zero, called **molecular sequence count** msc^k , is obtained:

$$msc^k = \frac{1}{2} \cdot CS_k(\mathbf{SM}) = \frac{1}{2} \cdot \sum_{i=1}^A sm_{ik}$$

where A is the number of graph vertices. In particular, msc^k is the $\rightarrow$ graph distance count ^kf for the distance-sequence matrix $\mathbf{SD}$, the $\rightarrow$ molecular path count ^kP for the path-sequence matrix $\mathbf{SP}$, the $\rightarrow$ molecular walk count $mwc^{(k)}$ for the walk-sequence matrix $\mathbf{SW}$, and the $\rightarrow$ molecular self-returning walk count $srw^{(k)}$ for the self-returning walk-sequence matrix $\mathbf{SSRW}$.

The set of molecular sequence counts constitutes a molecular $\rightarrow$ vectorial descriptor, called **molecular sequence code**:

$$\{msc^0, msc^1, msc^2, \dots, msc^K\}$$

where msc^0 is simply the number A of vertices in the graph.

The global index measuring the total number of walks of a specified type and of any length in the graph is the **total sequence count** TSC_M defined as

$$TSC_M = \sum_{k=0}^K msc^k = A + \frac{1}{2} \cdot \sum_{i=1}^A asc_i$$

where the subscript M denotes the type of sequence matrix. This index coincides with the $\rightarrow$ total path count P for the path-sequence matrix $\mathbf{SP}$ and the $\rightarrow$ total walk count TWC for the walk-sequence matrix $\mathbf{SW}$.

- **sequential search** $\rightarrow$ variable selection
- **Seri-Levy chirality coefficients** $\rightarrow$ chirality descriptors
- **SESP-Geo vectors** $\rightarrow$ distance-counting descriptors
- **SESP-Top vectors** $\rightarrow$ distance-counting descriptors
- **SE-vectors** $\equiv$ distance-counting descriptors

■ shadow indices ($\equiv$ Jurs shape indices)

A set of 3D $\rightarrow$ geometrical descriptors, similar to the $\rightarrow$ Amoore shape indices, related to the size and shape of molecules. They are calculated by projecting the molecular surface on three mutually perpendicular planes XY , XZ , and YZ , assuming van der Waals radii for atoms [Rohrbaugh and Jurs, 1987a, 1987b; Jurs, Hasan *et al.*, 1988; Bureau, Daveu *et al.*, 2002a] (Figure S3).

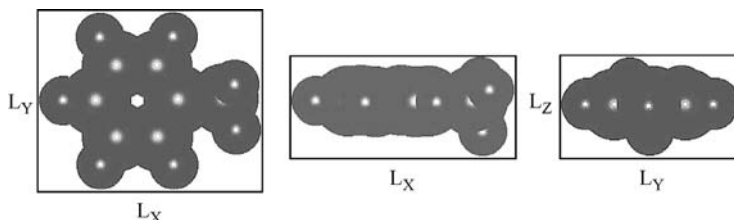


Figure S3 Projections and embedding rectangles of toluene molecule in the three principal planes.

Basically, a molecule is flattened into a plane by disregarding the third dimension; the area of the molecule, which is projected onto the remaining two dimensions, defines the shadow area of interest. To obtain invariance to rotation of the calculated projections, the X, Y, Z molecule axes are previously aligned along the three $\rightarrow$ *principal inertia axes*.

The six shadow indices are defined as the following:

SHDW1: area of the molecular shadow in the XY plane;

SHDW2: area of the molecular shadow in the XZ plane;

SHDW3: area of the molecular shadow in the YZ plane;

SHDW4: standardized molecular shadow area in the XY plane calculated as

$$\text{SHDW4} = \frac{\text{SHDW1}}{L_X \cdot L_Y}$$

where the denominator is the area of the embedding rectangle;

SHDW5: standardized molecular shadow area in the XZ plane calculated as

$$\text{SHDW5} = \frac{\text{SHDW2}}{L_X \cdot L_Z}$$

where the denominator is the area of the embedding rectangle;

SHDW6: standardized molecular shadow area in the YZ plane calculated as

$$\text{SHDW6} = \frac{\text{SHDW3}}{L_Y \cdot L_Z}$$

where the denominator is the area of the embedding rectangle.

The lengths L_X , L_Y , and L_Z are the maximum dimensions of the molecular surface projections. The ratio of the largest to the smallest dimension of the box built on each molecule shadow can be considered as a shape descriptor very similar to the $\rightarrow$ *length-to-breadth ratio*.

- **shaf(D) index** $\rightarrow$ algebraic operators ($\odot$ determinant)
- **Shannon's entropy** $\equiv$ *mean information content* $\rightarrow$ information content

■ Shannon Entropy Descriptors (SHED)

Shannon entropy descriptors are generated by using the concept of $\rightarrow$ *Shannon's entropy* to quantify the variability of a distribution of pharmacophoric type pairs [Gregori-Puigjané and Mestres, 2006]. To calculate these descriptors, atoms in a molecule are first mapped to Sybyl atom types, according to which they are assigned one or more of four different pharmacophore point types: hydrophobic (H), aromatic (R), hydrogen-bond acceptor (A), and hydrogen-bond donor (D). Then, for each of the 10 possible combinations of the four pharmacophore point types, 20 possible $\rightarrow$ *topological distances* between pairs of features are evaluated and for each distance the number of occurrence of the considered feature pair is calculated so that a final feature pair distribution is derived. Feature pairs at distance greater than 20 are counted in the last bin.

To measure the variability of each k th feature pair distribution, a projected entropy value E is calculated as follows:

$$E_k = e^{H_k} \quad H_k = - \sum_{i=1}^{20} p_i \cdot \ln p_i \quad k = 1, 2, 3, \dots, 10$$

where k is the considered feature pair, H_k is its Shannon's entropy, and p_i the probability of the i th bin of the distribution. Projected entropy values provide a measure of the expected maximum uniform occupancy from the corresponding entropy values: values of E can vary from 1, reflecting the situation of zero entropy corresponding to a feature pair population totally concentrated in a single distance bin, to 20, reflecting the situation of maximum entropy in which the feature pair is uniformly distributed among all the distance bins.

The final SHED vector is comprised of the projected entropy values E relative to the distributions of the 10 possible combinations of the considered pharmacophore point types.

■ shape descriptors

Molecular shape is related to several physico-chemical processes, such as transport phenomena as well as entropy contributions, and interaction capability between the ligand and receptor.

The degree of deviation from the spherical top is called **anisometry**.

Several shape descriptors are defined in the framework of more general approaches to → *molecular descriptors* [Woolley, 1978a; Motoc, 1983a; Kier, 1990; Arteca, 1991; Randić and Razinger, 1995b; Mezey, 1997a; Randić, 2001f; Mansfield and Covell, 2002; Zyrianov, 2005; Gramatica, 2006]. This is the case of → *Kier shape descriptors*, → *shape profiles*, → *shadow indices*, → *WHIM shape descriptors*, → *Sterimol shape parameters* L/B_1 and B_1/B_5 , molecular → *periphery codes*, → *centric indices*, and → *shape ETA indices*. Other approaches to study molecular surface and shape are → *Mezey 3D shape analysis* and Hopfinger → *molecular shape analysis*. → *Triangular descriptors* have also been used to characterize molecular shape to search for similarities among molecules. Other specific shape descriptors are listed below.

• Petitjean shape indices

A first Petitjean shape index is a topological anisometry descriptor [Petitjean, 1992], also called **graph-theoretical shape coefficient** I_2 , defined as

$$I_2 = \frac{D-R}{R} \quad 0 \leq I_2 \leq 1$$

where R and D are the → *topological radius* and the → *topological diameter*, respectively, obtained from the → *distance matrix* representing the considered → *molecular graph*. For strictly cyclic graphs, $D = R$ and $I_2 = 0$.

The **geometrical shape coefficient** I_3 is calculated in the same way but using the information of the → *geometry matrix* [Bath, Poirrette *et al.*, 1995]:

$$I_3 = \frac{{}^G D - {}^G R}{{}^G R}$$

where ${}^G R$ and ${}^G D$ are the → *geometric radius* and → *geometric diameter*, respectively.

A **radius-diameter diagram** is defined as a bivariate distribution of the → *data set* compounds in the space defined by the molecular radius and diameter; it provides a summary of the similarities among the molecule chemical shapes in the topological or geometrical space.

• Kaliszan shape parameter (η)

Defined as the ratio of the longest to the shortest side of a rectangle having the minimum area that can envelope a molecular structure drawn, assuming → *van der Waals radius* for atoms and

standard bond lengths [Kaliszan, Lamparczyk *et al.*, 1979; Radecki, Lamparczyk *et al.*, 1979; Kaliszan, 1987]. It was used to model GC retention indices. A slightly different shape parameter is the **length-to-breadth ratio** L/B that is defined as the ratio of the longest to the shortest side of the rectangle that envelopes a molecular structure and at the same time maximizes L/B ratio [Janini, Johnston *et al.*, 1975; Wise, Bonnett *et al.*, 1981].

In general, the length-to-breadth ratio is the ratio of the longest L to the shortest B side of a rectangle containing some molecular projection, once univocally defined a specific molecular orientation. For example, length-to-breadth ratio for molecular substituents is among $\rightarrow$ *STERIMOL parameters*. Moreover, length-to-breadth ratio was calculated from the dimensions of rectangles that envelope the molecule oriented along with the $\rightarrow$ *principal inertia axes* [Collantes, Tong *et al.*, 1996]. In this case, the dimensions of the rectangle enclosing the molecule are calculated using the atomic coordinates on the principal inertia axes.

The ratio between the first and second eigenvalues derived from $\rightarrow$ *WHIM descriptors* can be used as a shape descriptor related to the length-to-breadth ratio L/B. It is defined as

$$L/B_w = \frac{\lambda_{1w}}{\lambda_{2w}}$$

where w is one among the $\rightarrow$ *weighting schemes* defined in the WHIM approach. It can be noted that this shape parameter not only accounts for the distance between extreme atoms along the principal axes but also for the distribution of all atoms around the molecule center [Todeschini and Consonni, 2000].

- **Amoore shape indices**

Defined as the cross-sectional areas of the molecular surface in the inertial planes, that is, the planes obtained by the $\rightarrow$ *principal moments of inertia* of the molecule [Amoore, 1964; Meyer, 1986a]. When the three cross-sectional areas are equal, the molecule is a spherical top.

- **inertial shape factor (S_I)**

This is a shape factor based on the $\rightarrow$ *principal moments of inertia* and defined as

$$S_I = \frac{I_B}{I_A \cdot I_C}$$

where I are the principal moments of inertia [Lister, Macdonald *et al.*, 1978].

- **molecular eccentricity (ϵ)**

Among the $\rightarrow$ *spectral indices*, molecular eccentricity is a shape descriptor obtained from the eigenvalues λ_i of the $\rightarrow$ *inertia matrix* defined as [Arteca, 1991]

$$\epsilon = \frac{(\lambda_1^2 - \lambda_3^2)^{1/2}}{\lambda_1} \quad 0 \leq \epsilon \leq 1$$

where $\epsilon = 0$ corresponds to spherical top molecules and $\epsilon = 1$ to linear and planar molecules.

It is a shape descriptor defined by analogy with the eccentricity of an ellipse, which is defined as

$$\epsilon = \frac{(l_M^2 - l_m^2)^{1/2}}{l_M}$$

where l_M and l_m are the lengths of the major and minor elliptical axes, respectively.

- **asphericity** (Ω_A)

A descriptor that measures the deviation from the spherical shape [Arteca, 1991], calculated from the eigenvalues λ_i of the $\rightarrow$ *inertia matrix* as

$$\Omega_A = \frac{1}{2} \cdot \frac{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2} \quad 0 \leq \Omega_A \leq 1$$

where $\Omega_A = 0$ corresponds to spherical top molecules and $\Omega_A = 1$ to linear molecules. For prolate molecules (cigar shaped), $\lambda_1 > \lambda_2 \approx \lambda_3$ and $\Omega_A \approx 1$, whereas for oblate molecules (disk shaped) $\lambda_1 \approx \lambda_2 > \lambda_3$ and $\Omega_A \approx 0.5$.

- **spherosity index** (Ω_S)

An anisometry descriptor defined as a function of the eigenvalues, obtained by $\rightarrow$ *Principal Component Analysis* applied to the covariance matrix calculated from the $\rightarrow$ *molecular matrix M*:

$$\Omega_S = \frac{3 \cdot \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} \quad 0 \leq \Omega_S \leq 1$$

Spherosity index varies from zero for flat molecules, such as benzene, to one for totally spherical molecules [Robinson, Barlow *et al.*, 1997a].

- **linearity index** (L_i)

Based on a similar approach of the unweighted $\rightarrow$ *WHIM shape* index K_u , it is defined as [Patel and Cronin, 2001]

$$L_i = \sqrt{\frac{\lambda_1/\lambda_2/\lambda_3}{MW^2}} \times 100$$

where the inertia moments are calculated on the unweighted atom coordinates and MW is the molecular weight. High values are associated with small linear molecules (4–7), whereas small values are associated with highly branched nonlinear molecules (0–0.5).

- **ovality index** (O)

The ovality index O is an anisometry descriptor, that is, a measure of the departure of a molecule from the spherical shape, based on the property that for a fixed volume, the spherical shape presents the minimum surface [Bodor, Gabanyi *et al.*, 1989; Bodor, Buchwald *et al.*, 1998]. It is calculated from the ratio of the actual molecular surface area SA over the minimum surface area SA_0 corresponding to the actual molecule $\rightarrow$ *van der Waals volume* V^{vdw} :

$$O = \frac{SA}{SA_0} = \frac{SA}{4\pi R^2} = \frac{SA}{4\pi \cdot \left(\frac{3 \cdot V^{vdw}}{4\pi}\right)^{2/3}} \quad O \geq 1$$

where R is the molecule radius. The ovality index is equal to 1 for spherical top molecules and increases with increasing linearity of the molecule. This index was also later called **roughness** [Zyrianov, 2005]. The square of the ovality index was also proposed for modeling purposes [Bodor, Harget *et al.*, 1991].

The inverse of the ovality index is the **globularity factor** G ($0 < G \leq 1$), which is between zero and one [Meyer, 1986c]. For molecules with the same volume, the most spherical species have G values approximating one, and for molecules of nonequal volume, G reflects the relative compactness. When both the effective surface area and volume of the molecule are available, the **surface–volume ratio** $G' = SA/V$ can be used as a descriptor of molecular congestion. More specifically, it was interpreted as a measure of the capability of a compound to adapt its shape to the requirements of an approaching reagent [Meyer, 1988b].

- **Meyer visual descriptor of globularity**

A geometrical shape descriptor based on the radii of three spheres centered at the barycenter of the molecule [Meyer, 1986a]. The first sphere of radius R_1 has volume equal to the $\rightarrow$ *van der Waals volume*; the second sphere of radius R_2 has volume equal to the $\rightarrow$ *molecular volume*, and the third sphere of radius R_3 is defined as the sphere embedding the whole molecule. The shape descriptor is then defined as

$$R_M = \frac{R_3 - R_2}{R_2 - R_1}$$

Spherical top corresponds to small values of the R_M parameter.

- **Ciubotariu shape indices**

These indices are defined in terms of the $\rightarrow$ *van der Waals volume* V^{vdw} and $\rightarrow$ *van der Waals molecular surface* SA^{vdw} [Ciubotariu, Medeleanu *et al.*, 2004]. A **packing density index** was proposed as

$$R^{vdw} = \frac{V^{vdw}}{SA^{vdw}}$$

which is a steric/shape measure of a molecule.

Moreover, valid only for acyclic molecules, two globularity indices were proposed. The first index is defined as

$$G^{LOB} = \frac{R^{vdw}}{R^S}$$

where R^S represents the ratio between the volume and surface of an equivalent sphere, which surrounds the molecule, with the radius equal to the half of the longest dimension of the parallelepiped that embeds the molecule.

The second globularity index is defined as

$$G^{LEL} = \frac{V^{EL}}{V^S}$$

where V^{EL} is the volume of the ellipsoid surrounding the whole molecule and V^S the volume of a sphere with a radius equal to half of the longest ellipsoidal axis.

The two globularity measures decrease with the growth of the linear chains and increase toward unity when the molecule is highly branched or compacted.

- **geometrical eccentricity**

Geometrical eccentricity is calculated as

$$\varepsilon_G = \frac{L \times W \times T}{R^3}$$

where L , W , and T are the length, width, and thickness of the smallest box that can be formed around a rigid molecule, respectively [Johnson and Yalkowsky, 2005]. The term R denotes the radius of the equivalent sphere containing the same van der Waals volume as the molecule. Eccentricity, like flexibility and $\rightarrow$ *spirality*, increases the entropy of melting.

- **characteristic ratio (C_∞)**

A descriptor of average shape features of macromolecules, polymers, and proteins, that is, it can be considered as a measure of the degree of folding, defined as [Arteca, 1991]

$$C_\infty = \lim_{B \rightarrow \infty} C = \frac{\langle R_G^2 \rangle}{B \cdot l^2}$$

where $\langle R_G^2 \rangle$ is the mean square $\rightarrow$ *radius of gyration* averaged on all of the conformations (or configurations), B the number of bonds, and l the $\rightarrow$ *Kuhn length*.

The characteristic ratio is also defined for the $\rightarrow$ *end-to-end distance* d_{ee} as

$$C'_\infty = \lim_{B \rightarrow \infty} C' = \frac{\langle d_{ee}^2 \rangle}{B \cdot l^2}$$

- **path/walk shape indices**

Similar to the invariants derived from the $\rightarrow$ *distance/distance matrix* D/D , **atomic path/walk indices** are defined [Randić, 2001f] for each atom as the ratio of $\rightarrow$ *atomic path count* ${}^m P_i$ over $\rightarrow$ *atomic walk count* $awc_i^{(m)}$ of the same length m , that is,

$$(p/w)_i^m = {}^m P_i / awc_i^{(m)}$$

Whereas the number of paths in a molecule is bounded and determined by the molecule diameter, the number of walks is unbounded. However, being interested only in quotients, the walk count is terminated when the walk exceeds the length of the corresponding path.

Molecular path/walk indices are defined as the average sum of atomic path/walk indices of equal length:

$$(p/w)^m = \frac{1}{A} \cdot \sum_{i=1}^A (p/w)_i^m$$

Alternatively, they are obtained by separately summing all the paths and walks of the same length, and then calculating the ratio between their counts:

$$(P/W)^m = \frac{1}{A} \cdot \frac{\sum_{i=1}^A {}^m P_i}{\sum_{i=1}^A awc_i^{(m)}} = \frac{1}{A} \cdot \frac{{}^m P}{mwc^{(m)}}$$

where ${}^m P$ and $mwc^{(m)}$ are the $\rightarrow$ *molecular path count* and $\rightarrow$ *molecular walk count* of the m th order, respectively.

It should be noted that the counts of the paths and walks of length one are equal and, therefore, the corresponding molecular indices always equal one for all molecules.

As the path/walk count ratio is independent of molecular size, these descriptors can be considered as shape descriptors. Both the proposed indices can be transformed into $\rightarrow$ *uniform-length descriptors* by fixing the maximum length (e.g., $m = 5$) (Table S5):

$$\{(p/w)^2, (p/w)^3, (p/w)^4, (p/w)^5\} \quad \{(P/W)^2, (P/W)^3, (P/W)^4, (P/W)^5\}$$

Table S5 Molecular path/walk indices obtained by summing atomic path/walk indices for octane isomers (C8, Appendix C – Set 1).

Molecule	$(p/w)^2$	$(p/w)^3$	$(p/w)^4$	$(p/w)^5$	Molecule	$(p/w)^2$	$(p/w)^3$	$(p/w)^4$	$(p/w)^5$
n-Octane	0.458	0.223	0.101	0.048	33MM	0.554	0.283	0.091	0.016
2M	0.502	0.212	0.087	0.043	34MM	0.542	0.332	0.096	0.015
3M	0.500	0.262	0.094	0.047	2M3E	0.538	0.328	0.122	0
4M	0.492	0.255	0.126	0.032	3M3E	0.560	0.362	0.083	0
3E	0.492	0.302	0.134	0.036	223MMM	0.603	0.307	0.061	0
22MM	0.556	0.200	0.078	0.047	224MMM	0.593	0.180	0.097	0
23MM	0.542	0.288	0.092	0.032	233MMM	0.608	0.345	0.047	0
24MM	0.540	0.247	0.101	0.029	234MMM	0.588	0.310	0.077	0
25MM	0.546	0.203	0.075	0.059	2233MMMM	0.670	0.321	0	0

• shape factor (E_T)

This index is defined as

$$E_T = \frac{\sum_k n_k \cdot (k+1)}{L}$$

where n_k is the number of vertices at a topological distance equal to k from any vertex belonging to the longest path in the $\rightarrow$ *H-depleted molecular graph*; L is the length, that is, the number of edges, of the longest path [Gálvez, García-Domenech *et al.*, 1995; Gozalbes, Gálvez *et al.*, 1999]. The lower the E_T value, the more elongated the graph; E_T represents the eccentricity of the graph if it is compared to an ellipse.

The **surface factor** S is calculated for cyclic compounds as

$$S = \sum_f (E_T \cdot L^2)_f$$

where the summation runs over all the rings and aliphatic fragments in the molecule.

Note that the double bonds are considered simple with respect to their contribution to the surface factor.

• dBx descriptors

These descriptors count the number of atoms in consecutive spherical layers of specified depth centered at the centroid of the molecule [Bayada, Hemersma *et al.*, 1999].

$$\text{dBx} = \frac{n_x}{\text{Vol}_x}$$

where n_x is the number of atoms in volume Vol_x centered at molecule centroid, Vol_x being the volume of the sphere with radius d_x minus volume of the sphere with radius d_{x-1} . The considered radii are $d_0 = 0$ and $d_{1-5} = 2, 4, 6, 9, 12 \text{ \AA}$; the corresponding molecular descriptor symbols are dB1–dB5.

Table S6 Shape indices for some selected compounds.

Compound	I_2	I_3	$(p/w)^2$	$(p/w)^3$	$(p/w)^4$	$(p/w)^5$	Ω_S	Ω_A	$^1\kappa_\alpha$	$^2\kappa_\alpha$	Ku	Km
Ethane	0.000	0.424	0.000	0.000	0.000	0.000	0.341	0.435	2.000	1.000	0.384	0.659
Propane	1.000	0.989	0.333	0.000	0.000	0.000	0.202	0.411	3.000	2.000	0.426	0.634
n-Butane	0.500	0.610	0.417	0.167	0.000	0.000	0.127	0.617	4.000	3.000	0.589	0.784
n-Pentane	1.000	0.968	0.433	0.200	0.067	0.000	0.087	0.702	5.000	4.000	0.670	0.837
n-Hexane	0.667	0.731	0.444	0.214	0.089	0.033	0.063	0.785	6.000	5.000	0.747	0.886
Isobutane	1.000	0.987	0.500	0.000	0.000	0.000	0.230	0.148	4.000	1.333	0.280	0.385
Neopentane	1.000	0.985	0.600	0.000	0.000	0.000	1.000	0.000	5.000	1.000	0.000	0.000
Cyclopropane	0.000	0.402	0.500	0.000	0.000	0.000	0.348	0.106	1.333	1.000	0.115	0.326
Cyclobutane	0.000	0.310	0.500	0.250	0.000	0.000	0.226	0.150	2.250	0.750	0.215	0.387
Cyclopentane	0.000	0.269	0.500	0.250	0.125	0.000	0.227	0.149	3.200	1.440	0.249	0.387
Cyclohexane	0.000	0.266	0.500	0.250	0.125	0.063	0.207	0.157	4.167	2.222	0.274	0.396
Benzene	0.000	0.283	0.500	0.250	0.125	0.063	0.000	0.250	3.412	1.606	0.500	0.500
Toluene	0.333	0.530	0.538	0.274	0.126	0.062	0.017	0.353	4.382	1.784	0.483	0.535
Phenol	0.333	0.497	0.538	0.274	0.126	0.062	0.000	0.361	4.344	1.757	0.500	0.538
Benzoic acid	0.667	0.785	0.556	0.303	0.139	0.066	0.000	0.463	5.984	2.421	0.531	0.650
Naphthalene	0.667	0.847	0.554	0.326	0.182	0.095	0.000	0.392	5.483	2.144	0.500	0.576
Anthracene	0.750	0.798	0.573	0.340	0.195	0.114	0.000	0.571	7.573	2.846	0.646	0.741

I_2 and I_3 , Petitejean shape indices; $(p/w)^m$, molecular path/walk indices; Ω_S , sphericity index; Ω_A , asphericity; $^1\kappa_\alpha$ and $^2\kappa_\alpha$, Kier alpha-modified shape descriptors; and Ku and Km , WHIM shape indices.

• PathFinder fingerprints

PathFinder fingerprints are $\rightarrow$ *vectorial descriptors* encoding information about the molecular shape starting from a surface representation based on $\rightarrow$ *molecular interaction fields* (MIFs) [McLay, Hann *et al.*, 2006].

A subset of points, uniformly distributed on the molecular surface, is selected (e.g., 100 points). Then, the molecular surface is encoded into a weighted graph; its vertices are connected to each other when adjacent, through surface-lying paths. The minimum path is computed for each pair of vertices as the walk of minimum weight, the weight being defined as the sum of the weights of edges composing the walk. Only walks on the molecular surface are considered and the weights associated to edges are the geometric distances between two vertices, calculated from the Cartesian coordinates. Therefore, the minimum path is defined as the shortest connection between two points when “walking on the molecular surface.”

A path-distance matrix is then computed by combining information about paths on the surface and geometric distances for all the point pairs. Elements of this matrix count the frequency of two points on the molecular surface at a specific distance and path. Their values are related to molecular elongation, size, and variegation of the surface.

The final PathFinder fingerprint is obtained by unfolding of this path-distance matrix into a single vector.

📖 [Pitzer, Lippmann *et al.*, 1955; Pitzer, 1955; Woolley, 1978a; Motoc, 1983b; Arteca, 1991; Mezey, 1993c, 1997a; Kuz'min, Trigub *et al.*, 1995; Randić and Krilov, 1997b, 1999; Mössner, Lopez de Alda *et al.*, 1999]

- **shape ETA indices** → ETA indices
- **shape factor** → shape descriptors
- **shape group method** → Mezey 3D shape analysis
- **shape profiles** → molecular profiles
- **shape similarity coefficient** → chirality descriptors (⊙ Seri-Levy chirality coefficients)
- **shared-electron distribution index** → quantum-chemical descriptors (⊙ electron density)
- **SHED** ≡ *Shannon Entropy Descriptors*
- **shell matrices** ≡ *layer matrices*

■ Sh indices

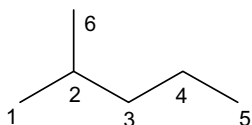
These are a set of topological indices, based on the same approach of the → *Schultz molecular topological index* and the → *Xu index*, calculated by different combinations of → *valence vertex degree* δ^v and → *vertex distance degree* σ as [Shamsipur, Hemmateenejad *et al.*, 2004; Shamsipur, Ghavami *et al.*, 2004]

$$\begin{aligned}
 Sh_1 &= \log \left(\sum_b \left(\frac{\sigma_i \cdot \sigma_j}{\delta_i^v \cdot \delta_j^v} \right)_b \right) & Sh_2 &= \log \left(\sum_b \left(\frac{\delta_i^v \cdot \delta_j^v}{\sigma_i \cdot \sigma_j} \right)_b \right) \\
 Sh_3 &= \log \left(\sum_b \left(\sigma_i \cdot \sigma_j \cdot \delta_i^v \cdot \delta_j^v \right)_b^{-1/2} \right) & Sh_4 &= \log \left(\sum_b \left(\frac{\delta_i^v \cdot \delta_j^v}{\sigma_i \cdot \sigma_j} \right)_b^{-1/2} \right) \\
 Sh_5 &= \sum_b \left(\sigma_i \cdot \sigma_j + \delta_i^v \cdot \delta_j^v \right)_b^{-1/2} & Sh_6 &= \log \sum_b \left(\sigma_i \cdot \sigma_j + \delta_i^v \cdot \delta_j^v \right)_b \\
 Sh_7 &= \sum_b \left(\delta_i^v \cdot \delta_j^v + \log(\sigma_i \cdot \sigma_j) \right)_b & Sh_8 &= \log(\boldsymbol{\sigma}^T \cdot \mathbf{v}) \\
 Sh_9 &= \log \left(\sum_{i=1}^A \sum_{j=1}^A [\mathbf{Sd}]_{ij} \right) & Sh_{10} &= \log(\text{MaxSp}(\mathbf{Sd})) \\
 Sh &= N_C + \sqrt{N_C} \cdot Sh_1
 \end{aligned}$$

where $\boldsymbol{\sigma}$ is the column vector collecting the distance degrees, $\mathbf{v}$ the column vector collecting the valence vertex degrees, and $\mathbf{Sd}$ the square $A \times A$ matrix obtained by the inner product of the two vectors $\boldsymbol{\sigma}$ and $\mathbf{v}$, that is, $\mathbf{Sd} = \boldsymbol{\sigma} \cdot \mathbf{v}^T$. In the Sh indices 1–7, the summations run over all the adjacent vertices. Sh_9 index is the sum over all the entries of the $\mathbf{Sd}$ matrix, whereas Sh_{10} index is the logarithm of its highest eigenvalue; Sh index is derived from the index Sh_1 , including the number of carbon atoms N_C to account for molecular size. This last index was proposed because it is highly correlated with boiling points of alkanes.

Example S3

Calculation of the indices Sh_1 , Sh_6 , Sh_{10} , and Sh for the H-depleted molecular graph of 2-methylpentane. D is the distance matrix and σ and δ^v the distance sums and the valence vertex degrees, respectively.



$D =$

Atom	1	2	3	4	5	6	σ_i	δ_i^v
1	0	1	2	3	4	2	12	1
2	1	0	1	2	3	1	8	3
3	2	1	0	1	2	2	8	2
4	3	2	1	0	1	3	10	2
5	4	3	2	1	0	4	14	1
6	2	1	2	3	4	0	12	1

$$Sh_1 = \log\left(\frac{12 \cdot 8}{1 \cdot 3} + \frac{8 \cdot 8}{3 \cdot 2} + \frac{8 \cdot 10}{2 \cdot 2} + \frac{10 \cdot 14}{2 \cdot 1} + \frac{8 \cdot 12}{1 \cdot 3}\right) = 2.22$$

$$Sh_6 = \log(12 \cdot 8 + 1 \cdot 2) + (8 \cdot 8 + 3 \cdot 2) + (8 \cdot 0 + 2 \cdot 2) + (10 \cdot 14 + 2 \cdot 1) + (8 \cdot 12 + 1 \cdot 3) = 2.68$$

$$\lambda_1(Sd) = 119.33 \quad Sh_{10} = \log(119.33) = 2.08 \quad Sh = 11.43$$

- **short hafnian** → algebraic operators (⊙ determinant)
- **SIBIS model** → minimal topological difference
- **side chain topological index** → biodescriptors (⊙ peptide sequences)
- **Siegel–Kormany inductive constant** → electronic substituent constants (⊙ inductive electronic constants)
- **signature descriptors** → substructure descriptors (⊙ fingerprints)
- **signed graph** → graph
- **similarity** → similarity/diversity

■ similarity/diversity

The concept of **similarity** and its dual concept of **diversity** play a fundamental role in several QSAR strategies and chemometric methods [Willett, 1987; Martin, Kofron *et al.*, 2002; Farnum, DesJarlais *et al.*, 2003; Willett, 2003a; Maldonado, Doucet *et al.*, 2006]. By definition, similarity is a binary relationship, that is, a relationship between two objects.

Similarity searching is a standard tool for → *drug design*, based on the idea that given a target structure with interesting properties, similar compounds chosen in large databases should have similar properties. Often also called Quantitative Molecular Similarity Analysis (QMSA) [Basak, Gute *et al.*, 2003], similarity searching involves the specification of the target structure and its characterization by one or more structural descriptors; then, this set of reference structural descriptors is compared with the corresponding sets of descriptors for each of the molecules in the database. A measure of similarity between the target structure and each of the database structures allows a ranking of decreasing similarity with the target for all the molecules. The numerical value of a similarity/diversity measure depends on three main components: (a) the description of the objects (e.g., molecular descriptors), (b) the weighting scheme of the

description elements, and (c) the selected similarity index or distance [Rouvray, 1990b; Bath, Poirrette *et al.*, 1994; Klein, 1995; Willett, Barnard *et al.*, 1998; Downs and Willett, 1999].

→ *Cluster analysis* methods, → *Principal Component Analysis* and related techniques, and different → *artificial neural networks* (such as → *Self-Organizing Maps*) are usually used to search for clusters of similar compounds, a cluster being comprised of distinct objects that are more similar to each other than to any other object outside the group.

Different methods have been developed for similarity searching, some based on topological features of the molecule and others on the three-dimensional structures; the latter are also able to account for different conformations of each molecule [Johnson and Maggiora, 1990]. Moreover, two distinct similarity approaches have been recognized. The direct-comparison methods [Dean and Chau, 1987; Dean, Callow *et al.*, 1988] search for molecule similarity on the basis of counts of substructures common to a pair of molecules or, in general, on the best superimposition between two molecules obtained by some → *alignment rules*, such as → *molecular shape similarity descriptors* or the → *maximum common substructure*. The descriptor-based similarity methods search for similarity among molecules using any kind of → *molecular descriptors* calculated independently for each molecule, being often → *substructure descriptors*, → *molecular profiles*, → *BCUT descriptors*, → *SWM signals*, → *EVA descriptors*, → *MoRSE descriptors*, → *autocorrelation descriptors*, and any set of → *vectorial descriptors*; as these methods are very fast, they are particularly suitable for searches in large databases [Sheridan, Miller *et al.*, 1996; Brown, 1997; Lewis, Mason *et al.*, 1997].

Distance measures d_{st} between the molecules s and t are scalar values representing the basic measurements of diversity, that is, $d_{st} = 0$ for identical molecules [Sneath and Sokal, 1973; Cuadras, 1989; Frank and Todeschini, 1994; Willett, Barnard *et al.*, 1998]. Note that the variables considered in the distance measures have to be in a comparable scale, otherwise preliminary scaling procedures must be performed.

For a distance measure to be defined as a metric (or Euclidean), it must have the following properties:

- (1) $d_{st} \geq 0$, positivity
- (2) $d_{ss} = 0$, identity
- (3) $d_{st} = d_{ts}$, symmetry
- (4) $d_{st} \leq d_{sz} + d_{zt}$, triangular inequality

Semimetric distances (or pseudometrics) do not satisfy the fourth requirement, that is, the triangle inequality axiom, whereas nonmetric distances may take negative values, thus violating the first assumption, that is, the property of positiveness of metrics.

Although diversity is measured by distances, similarity is dually related to diversity and quantitatively evaluated by **similarity indices**.

In the case of chemical compounds, similarity indices quantify the degree of structural resemblance between their structural representations.

A similarity index s_{st} calculated for molecules s and t should have the following properties:

- (1) $0 \leq s_{st} \leq 1$ closure
- (2) $s_{ss} = 1$ identity
- (3) $s_{st} = s_{ts}$ symmetry

where $s_{st} = 1$ represents identical molecules and $s_{st} = 0$ represents the maximum dissimilarity. In spite of the common assumption of the closure condition, some similarity indices have an

upper bound greater than one. Moreover, in this framework, correlation measures can be considered as a special case of similarity measure.

The most important relationships between distance and similarity measures are given below.

For similarity measures varying between 0 and 1, as it is generally the case, the corresponding distance may be calculated as

$$d_{st} = 1 - s_{st} \quad d_{st} = \sqrt{1 - s_{st}} \quad d_{st} = \sqrt{1 - s_{st}^2}$$

Distances not bounded by some fixed upper value may be normalized using one of the following two equations:

$$d_{st}^N = \frac{d_{st}}{d_{\max}} \quad d_{st}^N = \frac{d_{st} - d_{\min}}{d_{\max} - d_{\min}}$$

where $d_{\min}$ and $d_{\max}$ are the minimum and maximum value taken by the distance measure, respectively, within a data set. Measures of similarity can be obtained from normalized distances d^N as follows:

$$s_{st} = 1 - d_{st}^N \quad s_{st} = 1 - (d_{st}^N)^2 \quad s_{st} = \sqrt{1 - (d_{st}^N)^2}$$

If normalization is obtained by using the minimum and maximum values of the data set, these equations give a relative similarity with respect to the pair of objects having the maximum (and minimum) distance, that is, at least exists a pair of objects with a similarity equal to zero.

For distance measures with no upper bound, the following equation can be used:

$$s_{st} = \frac{1}{1 + d_{st}}$$

This similarity measure is independent of the objects of the data set other than s and t .

Depending of the kind of variables (continuous, binary, ranks, angles, etc.), several different measures of distance and similarity were defined.

The most important distance measures for continuous variables are the Euclidean distance and the average Euclidean distance. However, depending on the considered problem, other distance measures can be legitimately used. Some are listed below (Table S7), where p is the number of real variables and x_{sj} and x_{tj} are the values of the j th element (variable, attribute, or descriptor) representing s and t objects, respectively; $\mathbf{x}_s$ and $\mathbf{x}_t$ are the descriptor p -dimensional vectors of the two objects. If the objects are chemical compounds, x_{ij} are the values of the molecular descriptors chosen for their representation, such as $\rightarrow$ *topological indices*, $\rightarrow$ *physico-chemical properties*, and vectors of substructural descriptors.

Table S7 Distance measures between the objects s and t , described by p quantitative variables. In the last column the average distance is also given.

Distance	Function	Range	Average
Euclidean distance	$d_{st} = \sqrt{\sum_{j=1}^p (x_{sj} - x_{tj})^2}$	$0 \leq d_{st} < \infty$	$\bar{d}_{st} = \frac{d_{st}}{\sqrt{p}}$

(Continued)

Table S7 (Continued)

Distance	Function	Range	Average
Canberra distance	$d_{st} = \sum_{j=1}^p \frac{ x_{sj} - x_{tj} }{ x_{sj} + x_{tj} }$	$0 \leq d_{st} \leq p$	$\bar{d}_{st} = \frac{d_{st}}{p}$
Lance–Williams distance/ Bray–Curtis distance	$d_{st} = \frac{\sum_{j=1}^p x_{sj} - x_{tj} }{\sum_{j=1}^p (x_{sj} + x_{tj})}$	$0 \leq d_{st} \leq 1$	$\bar{d}_{st} = \frac{d_{st}}{p}$
Manhattan distance/ city–block distance	$d_{st} = \sum_{j=1}^p x_{sj} - x_{tj} $	$0 \leq d_{st} < \infty$	$\bar{d}_{st} = \frac{d_{st}}{p}$
Lagrange distance/ Chebyshev distance	$d_{st} = \max_j x_{sj} - x_{tj} $	$0 \leq d_{st} < \infty$	—
Minkowski distance	$d_{st} = \left(\sum_{j=1}^p x_{sj} - x_{tj} ^r \right)^{1/r}$	$r > 0; 0 \leq d_{st} < \infty$	$\bar{d}_{st} = \frac{d_{st}}{p^{1/r}}$
Clark distance/ coefficient of divergence	$d_{st} = \sqrt{\sum_{j=1}^p \left(\frac{ x_{sj} - x_{tj} }{ x_{sj} + x_{tj} } \right)^2}$	$0 \leq d_{st} \leq p$	$\bar{d}_{st} = \frac{d_{st}}{\sqrt{p}}$
Soergel distance	$d_{st} = \frac{\sum_{j=1}^p x_{sj} - x_{tj} }{\sum_{j=1}^p \max\{x_{sj}, x_{tj}\}}$	$0 \leq d_{st} \leq 1$	$\bar{d}_{st} = \frac{d_{st}}{p}$
Bhattacharyya distance	$d_{st} = \sqrt{\sum_{j=1}^p (\sqrt{x_{sj}} - \sqrt{x_{tj}})^2}$	$0 \leq d_{st} \leq \infty$	$\bar{d}_{st} = \frac{d_{st}}{\sqrt{p}}$
Wave–Edges distance	$d_{st} = \sum_{j=1}^p \left(1 - \frac{\min\{x_{sj}, x_{tj}\}}{\max\{x_{sj}, x_{tj}\}} \right)$	$0 \leq d_{st} \leq p$	$\bar{d}_{st} = \frac{d_{st}}{p}$
Correlation distance	$d_{st} = 1 - \frac{\sum_{j=1}^p (x_{sj} - \bar{x}_s) \cdot (x_{tk} - \bar{x}_t)}{\sqrt{\sum_{j=1}^p (x_{sj} - \bar{x}_s)^2 \cdot \sum_{j=1}^p (x_{sk} - \bar{x}_t)^2}}$	$0 \leq d_{st} \leq 1$	—
Mahalanobis distance	$d_{st} = \sqrt{(\mathbf{x}_s - \mathbf{x}_t)^T \mathbf{S}^{-1} (\mathbf{x}_s - \mathbf{x}_t)}$	$\mathbf{S}$: covariance matrix	—
Chord distance ^a	$d_{st} = \sqrt{2 \cdot (1 - \cos \vartheta)}$	$0 \leq d_{st} \leq \sqrt{2}$	—
Angular separation	$d_{st} = 1 - \frac{\sum_{j=1}^p x_{sj} \cdot x_{tj}}{\sqrt{\sum_{j=1}^p x_{sj}^2 \cdot \sum_{j=1}^p x_{tj}^2}}$	$0 \leq d_{st} \leq 1$	—

^a ϑ is the angle between the two points on a circle.

Note that the Minkowski distance represents a family of distance measures, for which the higher the value of r , the greater the importance given to large differences. For $r = 1$, the Minkowski distance is the Manhattan distance, for $r = 2$ is the Euclidean distance, and for $r \rightarrow \infty$ is the Lagrange distance.

Weighted distances are obtained from the previously defined distances by weighting each j th variable by a user-defined weight w_j , usually under the constraint:

$$\sum_{j=1}^p w_j = 1$$

For example, the weighted Euclidean distance is defined as

$$d_{st} = \sqrt{\sum_{j=1}^p w_j \cdot (x_{sj} - x_{tj})^2}$$

Besides the transformations outlined above to derive similarity measures from distance measures, other common similarity indices for continuous variables are the following:

$$\begin{aligned} \text{Jaccard/Tanimoto coefficient : } s_{st} &= \frac{\sum_{j=1}^p x_{sj} \cdot x_{tj}}{\sum_{j=1}^p x_{sj}^2 + \sum_{j=1}^p x_{tj}^2 - \sum_{j=1}^p x_{sj} \cdot x_{tj}} \\ \text{cosine coefficient : } s_{st} &= \frac{\sum_{j=1}^p x_{sj} \cdot x_{tj}}{\sqrt{\sum_{j=1}^p x_{sj}^2 \cdot \sum_{j=1}^p x_{tj}^2}} \end{aligned}$$

Note that the complement of the cosine coefficient is the angular separation defined in Table S7.

When variables are represented by $\rightarrow$ *binary descriptors*, that is, variables whose values are either zero or one, different appropriate distance and similarity measures must be used.

Given two objects s and t , represented by p binary values 0/1, the **binary distance measures** and **binary similarity measures** are based on the frequencies arising from Table S8:

Table S8 Frequency table of the four combinations of two binary variables.

	$t = 1$	$t = 0$	
$s = 1$	a	b	$a + b$
$s = 0$	c	d	$c + d$
	$a + c$	$b + d$	p

where a , b , c , and d are the frequencies of the events $s = 1$ and $t = 1$, $s = 1$ and $t = 0$, $s = 0$ and $t = 1$, and $s = 0$ and $t = 0$, respectively, in the pair of binary vectors describing the objects s and t ; p is the total number of variables, equal to $a + b + c + d$, which is the length of the binary vector. In other words, a is the number of bits equal to one in both objects (common “presences”) and d the number of bits equal to zero in both objects (common “absences”), $a + b$ the number of bits equal to one in the s th object, and $a + c$ the number of bits equal to one in the t th object.

Therefore, the diagonal entries a and d give information about the similarity between the two vectors, whereas the entries b and c give information about their dissimilarity. Symmetrical

measures of similarity use both a and d , that is, the state double zero (d) for two objects is treated in exactly the same way as any other pair of values and should be used when the state zero is a valid basis for comparing two objects; asymmetrical measures skip double-zero state in the similarity calculation.

Similarity measures on binary variables provide **similarity coefficients** s_{st} , also called **association coefficients** a_{st} , and $\rightarrow$ **correlation measures** r_{st} [Frank and Todeschini, 1994; Legendre and Legendre, 1998; Salim, Holliday *et al.*, 2003].

A list of similarity coefficients (s) and correlation measures (r) is given in Table S9, where s and t represent the two binary vectors and a , b , c , and d the occurrence number defined in Table S8.

Table S9 Similarity coefficients between the objects s and t , described by p binary variables.

Distance	Function	Range	Distance	Function	Range
Hamming	$s_{st} = a + d$	$[0, p]$	Baroni–Urbani/ Buser	$s_{st} = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$	$[0, 1]$
Simple matching	$s_{st} = \frac{a + d}{p}$	$[0, 1]$	Kulczynski (1)	$s_{st} = \frac{a}{b + c}$	$[0, \infty]$
Rogers–Tanimoto	$s_{st} = \frac{a + d}{p + b + c}$	$[0, 1]$	Kulczynski (2)	$s_{st} = \frac{1}{2} \cdot \left[\frac{a}{a + b} + \frac{a}{a + c} \right]$	$[0, 1]$
Jaccard–Tanimoto	$s_{st} = \frac{a}{a + b + c} = \frac{a}{p - d}$	$[0, 1]$	Sokal–Sneath (1)	$s_{st} = \frac{a}{a + 2b + 2c}$	$[0, 1]$
Hamann	$s_{st} = \frac{a + d - b - c}{p}$	$[0, 1]$	Sokal–Sneath (2)	$s_{st} = \frac{2a + 2d}{p + a + d}$	$[0, 1]$
Dice/Sørensen/ Czekanowski	$s_{st} = \frac{2a}{2a + b + c}$	$[0, 1]$	Sokal–Sneath (3)	$s_{st} = \frac{a + d}{b + c}$	$[0, \infty]$
Russel–Rao	$s_{st} = \frac{a}{p}$	$[0, 1]$	Fossum	$s_{st} = \frac{p(a - 1/2)^2}{(a + b)(a + c)}$	$\left[\frac{1}{p}, \approx p \right]$
Forbes–Mozley	$s_{st} = \frac{pa}{(a + b)(a + c)}$	$[0, 1]$	Pearson ^a	$r_{st} \equiv \Phi = \frac{ad - bc}{Q}$	$[\pm 1]$
Simpson	$s_{st} = \frac{a}{\min\{(a + b), (a + c)\}}$	$[0, 1]$	Yule	$r_{st} = \frac{ad - bc}{ad + bc}$	$[\pm 1]$
Braun–Blanquet	$s_{st} = \frac{a}{\max\{(a + b), (a + c)\}}$	$[0, 1]$	McConnaughey	$r_{st} = \frac{a^2 - bc}{(a + b)(a + c)}$	$[\pm 1]$
Cosine/Ochiai	$s_{st} = \frac{a}{\sqrt{(a + b)(a + c)}}$	$[0, 1]$	Dennis	$r_{st} = \frac{ad - bc}{\sqrt{p(a + b)(a + c)}}$	$\left[-1, \approx \frac{\sqrt{p}}{2} \right]$

$$^a Q = \sqrt{(a + b) \cdot (a + c) \cdot (b + d) \cdot (c + d)}.$$

Similarity coefficients for binary variables are also used as $\rightarrow$ **classification parameters** for two-class problems; among these, the most used is the **Pearson coefficient** Φ (Table S9), which is also known as Matthews correlation index.

A weighted version of the Jaccard–Tanimoto association measure is the **Tversky similarity coefficient**, given as [Tversky, 1977]

$$s_{st} = \frac{a}{\alpha \cdot b + \beta \cdot c + a} \quad 0 \leq s_{st} \leq 1$$

where α and β are user-defined parameters. In particular, equal values of α and β provide a symmetrical contribution to the two dissimilarity parameters b and c , such as in the Jaccard coefficient when $\alpha = \beta = 1$ and in the Dice/Sørensen coefficient when $\alpha = \beta = 1/2$; different values of α and β provide asymmetrical contribution, as, for example, when $\alpha = 1$ and $\beta = 0$, the corresponding similarity is $s_{st} = a/(a + b)$ and can be interpreted as the fraction of object s , which is in common with object t [Wang, Eckert *et al.*, 2007].

Another weighted variant of the Jaccard/Tanimoto distance was suggested considering a trade-off between the presence and absence of common features. In effect, the Jaccard/Tanimoto similarity emphasizes the presence of common features a , neglecting the absence of common features d and it is written as

$$T_1 = \frac{a}{a + b + c} = \frac{a}{p - d}$$

If the reversed case is defined, that is, emphasizing the absence of common features, the following similarity definition derives

$$T_2 = \frac{d}{a + b + d} = \frac{d}{p - c}$$

A **modified weighted Tanimoto coefficient** was also proposed as [Fligner, Verducci *et al.*, 2002]

$$MT_\alpha = \alpha \cdot (T_1) + (1 - \alpha) \cdot T_2$$

where the parameter α should be determined to make the similarity measure independent of total number p of variables (under the assumption of the Bernoulli distribution). In this case, an α value equal to $(2 - p)/3$ has been suggested.

Another weighted measure for binary variables is the **azzoo similarity coefficient**, which is a weighted Hamming similarity coefficient and aims at weighing the contribution of the term d as

$$s_{st} = a + \sigma \cdot d \quad 0 \leq \sigma < \infty \quad [0, < \infty]$$

where σ is a user-defined parameter equal to or greater than zero. Of course, for $\sigma = 1$, the azzoo similarity coefficient coincides with the Hamming coefficient, whereas for $\sigma = 0$, only the positive matches (a) are considered.

If the terms b and c are replaced in such a way that n_s is the number of bits “1” of the object s , that is, $a + b$, and n_t the number of bits “1” of the object t , that is, $a + c$, the previous binary measures can take a different form. For example, the Tanimoto index can be written as

$$s_{st} = \frac{a}{b + c + a} = \frac{a}{(a + b) + (a + c) - a} = \frac{a}{n_s + n_t - a}$$

This is a general formula allowing the comparison of molecules represented by vectorial descriptors of different lengths (e.g., vectors of structural features).

A modified Tanimoto index was further proposed to deal with occurrence frequencies [Carhart, Smith *et al.*, 1985]:

$$s_{st} = \frac{2 \cdot \sum_{j=1}^p \min\{f_{sj}, f_{tj}\}}{\sum_{j=1}^p f_{sj} + \sum_{j=1}^p f_{tj}}$$

or, similarly [Filimonov, Poroikov *et al.*, 1999]

$$s_{st} = \frac{\sum_{j=1}^p \min\{f_{sj}, f_{tj}\}}{\sum_{j=1}^p f_{sj} + \sum_{j=1}^p f_{tj} - \sum_{j=1}^p \min\{f_{sj}, f_{tj}\}}$$

where f are the frequencies of the j th descriptor (e.g., fragment) in each object (e.g., molecule).

The most popular distance binary measures are Hamming and Tanimoto distances that are listed below (Table S10), together with other distance measures on binary vectors.

Table S10 Some distance measures derived from binary variables.

Distance	Function	Range
Hamming distance	$d_{st} = b + c$	$[0, p]$
Square root Hamming distance	$d_{st} = \sqrt{b + c}$	$[0, \sqrt{p}]$
Tanimoto distance	$d_{st} = \frac{b + c}{p}$	$[0, 1]$
Square root Tanimoto distance	$d_{st} = \sqrt{\frac{b + c}{p}}$	$[0, 1]$
Watson nonmetric distance	$d_{st} = \frac{b + c}{2a + b + c}$	$[0, 1]$
Soergel binary distance	$d_{st} = \frac{b + c}{a + b + c}$	$[0, 1]$

It must be noted that comparing distances for binary and continuous variables, the Hamming distance coincides with the Manhattan distance, square root Hamming distance is the Euclidean distance, Tanimoto distance coincides with average Manhattan distance and squared Tanimoto with the average Euclidean distance. Moreover, the Watson nonmetric distance corresponds to the Lance–Williams distance and is the complement of the Sørensen coefficient; the Soergel binary distance corresponds to the Soergel distance and is the complement of the Jaccard/Tanimoto coefficient.

When the objects of the data set are ranked, measures of distance can also be applied on the ranks r_{sj} and r_{tj} , representing the ranks of the objects s and t , respectively, for the j th variable. The most important distance measures on ranked data are listed below:

$$\text{Mahalanobis-like distance : } d_{st} = 2 \cdot \sum_{j=1}^p \frac{(r_{sj} - r_{tj})^2}{(r_{sj} + r_{tj})}$$

$$\text{Rank distance : } d_{st} = \sum_{j=1}^p \frac{(r_{sj} - r_{tj})^2}{s_j^2}$$

where s_j^2 is the variance of the variable j .

When the considered variables are probabilities (p), appropriate similarity or distance measures should take into account the probability constraints. The **Dice correlation coefficient**

is defined for this case as

$$r_{st} = \frac{2 \cdot p(x, y)}{p(x)^2 + p(y)^2}$$

Here $p(x)$ and $p(y)$ are the probability for event x or y that appear separately; $p(x, y)$ is the probability for event x and y appearing together.

Another distance measure between probabilities is the following:

$$d_{st} = [p(x) - p(y)] \cdot \log_2 \left(\frac{p(x)}{p(y)} \right) \quad p(x) > 0 \quad p(y) > 0$$

When the variables are represented by angles (0° – 360°), the **angular distance** is the absolute value of the shortest distance between the two points in a circle. For example, the angular distance between 90° and 120° is 30° , and between 350° and 10° is 20° .

Similarity/distance measures can be also thought of as similarity/distance measures between sets.

Examples of distance measures between the two sets K_s and K_t can be defined by the following expressions [Skvortsova, Baskin *et al.*, 1998]:

$$d_{st} = |K_s| + |K_t| - 2 \cdot |K_s \cap K_t| \quad \text{or} \quad d_{st} = \left(1 - \frac{|K_s \cap K_t|^2}{|K_s| \cdot |K_t|} \right) / (|K_s| + |K_t|)$$

where $|K|$ denotes the total number of elements in the set K , that is, its cardinality, and $|K_s \cap K_t|$ is the cardinality of the intersection of the two sets, that is, the number of common elements. For example, if the sets K contain structural fragments of molecular graph, then their intersection would be their $\rightarrow$ *maximum common substructure* MCS. Note that the first measure corresponds to the Hamming distance for binary variables.

Examples of similarity measures between the two sets K_s and K_t can be defined by the following expressions:

$$(1) s_{st} = \frac{|K_s \cap K_t|}{|K_s| + |K_t|} \quad (2) s_{st} = \frac{2 \cdot |K_s \cap K_t|}{|K_s| + |K_t|} \quad (3) s_{st} = \frac{|K_s \cap K_t|^2}{|K_s| \cdot |K_t|}$$

where $|K|$ denotes the total number of elements in the set K , that is, its cardinality, and $|K_s \cap K_t|$ is the cardinality of the intersection of the two sets, that is, the number of common elements [Basak, Bertelsen *et al.*, 1994; Skvortsova, Baskin *et al.*, 1998]. Note that the first measure corresponds to the Jaccard/Tanimoto coefficient, the second one to the Dice coefficient, and the third one to the square of the cosine coefficient.

Distance measures between two sets of variables are important to avoid, for example, the selection of models, which are only seemingly diverse due to the presence of different descriptors, but closely correlated among themselves. Distance between the sets of variables can be measured by the Hamming distance where the total distance is the sum of the variables that differ in the two sets. However, the Hamming distance usually overestimates the distance between the two sets of variables, neglecting the variable correlations.

A distance between models was proposed and called *Model Distance*; this measure is able to take into account the correlation between the two sets of variables [Todeschini, Consonni *et al.*, 2004c, 2004d]. An improved (and simplified) version of the original proposed distance between the sets of variables is here proposed. The new quantity is called **CMD index** and

is defined as [Authors, this book]

$$CMD_{AB} = p_A + p_B - 2 \cdot \sum_{j=1}^M \sqrt{\lambda_j} \quad 0 \leq CMD_{AB} \leq (p_A + p_B)$$

where A and B are the two compared sets of variables, p_A and p_B the number of variables present in sets A and B, respectively; M is the minimum rank between the two $\mathbf{Q}_A$ and $\mathbf{Q}_B$ symmetrized cross-correlation matrices, and λ their eigenvalues. It must be noted that this distance measure refers to pairs of the variable sets, independently of the specific model they are used for, such as regression, classification, or $\rightarrow$ *Principal Component Analysis*. This distance is zero only if there exists a pairwise correlation equal to one between all the pairs of variables.

The correlation between the variables of the two sets is encoded into the unsymmetrical cross-correlation matrix $\mathbf{C}_{AB}$ of size $(p_A \times p_B)$ and defined as

		Model B			
		$x_1(B)$	$x_2(B)$	$\dots$	$x_c(B)$
Model A	$x_1(A)$	$r_{1A,1B}$	$r_{1A,2B}$	$\dots$	$r_{1A,cB}$
	$x_2(A)$	$r_{2A,1B}$	$r_{2A,2B}$	$\dots$	$r_{2A,cB}$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$x_b(A)$	$r_{bA,1B}$	$r_{bA,2B}$	$\dots$	$r_{bA,cB}$

where r are the pairwise correlation coefficients between each variable of the set A and each variable of set B. Of course, the counterpart of $\mathbf{C}_{AB}$ is the cross-correlation matrix $\mathbf{C}_{BA}$ (size $p_B \times p_A$).

Both cross-correlation matrices can be transformed into a symmetric matrix as the following:

$$\mathbf{Q}_A = \mathbf{C}_{AB} \mathbf{C}_{BA} \quad (p_A \times p_A) \quad \mathbf{Q}_B = \mathbf{C}_{BA} \mathbf{C}_{AB} \quad (p_B \times p_B)$$

The M nonzero eigenvalues of both matrices $\mathbf{Q}_A$ and $\mathbf{Q}_B$ coincide (M being the minimum rank between the two $\mathbf{Q}$ matrices) and twice the sum of the square root of these eigenvalues λ is the interset common variance v_{AB} :

$$v_{AB} = \sum_{j=1}^M \sqrt{\lambda_j}$$

Being the maximum theoretical value of the common variance,

$$v_{AB}^{MAX} = \sqrt{p_A \cdot p_B}$$

a quantity that measures the degree of linear association between the two sets of variables is called **CMC index** and defined as

$$CMC_{AB} = \frac{\sum_{j=1}^M \sqrt{\lambda_j}}{\sqrt{p_A \cdot p_B}} \quad 0 \leq CMC_{AB} \leq 1$$

If no correlation exists between the two variable blocks, that is, $CMC_{AB} = 0$, the variable set distance coincides with the Hamming distance, being all the variables different and uncorrelated.

Example S4

Four variables x_1 – x_4 and a number of sets obtained by combining these variables in different ways are considered. In Table S11, the pairwise correlations of the four variables are collected, whereas in Table S12, the Hamming distance d_H , the *CMD* index, and the *CMC* index are given for different pairs of sets of variables.

Table S11 Pairwise correlations between the variables x_1 – x_4 .

	x_1	x_2	x_3	x_4
x_1	1	0.979	0.061	0.475
x_2	0.979	1	0.194	0.593
x_3	0.061	0.194	1	0.240
x_4	0.475	0.593	0.240	1

Table S12 Hamming distance, *CMD*, and *CMC* indices for pairs of variable sets; variable pairwise correlations are collected in Table S11.

Set A	Set B	p_A	p_B	b	c	d_H	<i>CMD</i>	<i>CMC</i>
x_1, x_2, x_3, x_4	x_1, x_2, x_3, x_4	4	4	0	0	0	0	1
x_1, x_2, x_3, x_4	x_1, x_2, x_3	4	3	1	0	1	0.591	0.925
x_1, x_2, x_3, x_4	x_1, x_2, x_4	4	3	1	0	1	0.819	0.892
x_1, x_2, x_3	x_1, x_2, x_4	3	3	1	1	2	1.295	0.784
x_3	x_4	1	1	1	1	2	1.520	0.240
x_1, x_3, x_4	x_2, x_3, x_4	3	3	1	1	2	0.028	0.995
x_1, x_2, x_4	x_3, x_4	2	2	2	1	3	2.272	0.432
x_1, x_2	x_3, x_4	2	2	2	2	4	2.291	0.427
x_1	x_2, x_3, x_4	1	3	1	3	4	1.821	0.629

CMD and *CMC* indices can be useful in the evaluation of the diversity of $\rightarrow$ *chemical spaces* in molecule library design, in selecting the optimal number of significant principal components in $\rightarrow$ *PCA*, and in selecting the most diverse QSAR models for $\rightarrow$ *consensus analysis*.

The pairwise distances calculated on the $\rightarrow$ *data matrix* $X(n \times p)$ can be arranged into a square symmetric matrix, called **data distance matrix** or **diversity matrix** ($n \times n$), in which both rows and columns correspond to objects. In analogous way, the pairwise similarity indices calculated from the data matrix $X(n \times p)$ can be arranged into a matrix, called **similarity matrix** ($n \times n$), in which rows and columns correspond to objects.

The $\rightarrow$ *correlation matrix* can be considered a special case of similarity matrix.

The most common data distance matrix is the **Euclidean distance matrix**, that is, the matrix obtained by using the Euclidean distance measure. A very important Euclidean distance matrix is the $\rightarrow$ *geometry matrix*, where rows and columns represent the molecule atoms, and matrix elements are interatomic distances calculated from the (x , y , and z) spatial atomic coordinates.

Analogous to the Euclidean distance matrix are the data distance matrices obtained using different distance measures, such as Manhattan distance, Canberra distance, Lagrange distance, and so on. Moreover, an $\rightarrow$ *Euclidean-distance map matrix* was defined, which encodes information about graphs used to describe $\rightarrow$ *proteomics maps*.

Note. The data distance matrix is usually simply called *distance matrix*. In this book, it is called *data distance matrix* to avoid confusion with the topological vertex $\rightarrow$ *distance matrix*, which contains the topological distances d_{ij} between the vertices of a molecular graph. The distances between the pairs of objects s – t of the data distance matrix are indicated by d_{st} , whereas d_{ij} are the topological distances between the pairs of vertices i – j .

Similarity matrices can be used to describe a set of molecules to search for QSAR models, correlating the biological activity of the molecules with their similarity to each other [Rum and Herndon, 1991; Good, Peterson *et al.*, 1993; Kubinyi, 1997]; these matrices are usually known as **molecular similarity matrices** and the corresponding QSAR method as **Quantitative Similarity–Activity Relationships** (QSiAR) [Kubinyi, Hamprecht *et al.*, 1998]. In other words, similarity and diversity measures can be used as input variables for modeling molecular properties or biological activities [Benigni, Cotta Ramusino *et al.*, 1995; Horwell, Howson *et al.*, 1995; So and Karplus, 1997a, 1997b].

The procedure consists in transforming the initial data matrix $\mathbf{X}$, with n compounds and p molecular descriptors, into a similarity or diversity matrix obtaining a $n \times n$ square symmetric matrix, after the selection of the distance (similarity) measure and the appropriate scaling of the original variables. A regression model is then performed using as the molecular descriptors the columns $\mathbf{d}_j$ of the distance matrix (*diversity descriptors*), where the column elements d_{ij} represent the distances between each i th molecule and the j th molecule. Analogously, molecular descriptors can be defined as the columns $\mathbf{s}_j$ of the similarity matrix (*similarity descriptors*).

As the number of variables equals the number of molecules, regression models should be obtained by using variable selection techniques or methods such as partial least squares or principal component regression.

Regression models obtained by this approach are formally written in the following form:

$$\hat{y}_i = b_0 + \sum_{j=1}^M b_j \cdot d_{ij} \quad \text{or} \quad \hat{y}_i = b_0 + \sum_{j=1}^M b_j \cdot s_{ij}$$

where b are the regression coefficients and M the number of selected matrix columns.

Once the M relevant distance descriptors are obtained, the model can be interpreted easily by considering the sign of each model descriptor: for a j th variable with a positive regression coefficient, the response increases when a compound is more dissimilar to the j th compound, whereas for a j th variable with a negative coefficient, the response increases when a compound is more similar to the j th compound. The opposite interpretation holds if similarity descriptors are used instead. In practice, such models give easily interpretable information in terms of similarity (or diversity) of each compound with respect to a set of reference compounds corresponding to the variables that are selected as relevant.

Differential descriptors such as $\rightarrow$ *steric misfit* can also be considered as similarity/diversity descriptors.

 Additional references are collected in the thematic bibliography (see Introduction).

- **similarity indices** → similarity/diversity
- **similarity matrix** → similarity/diversity
- **similarity score** → quantum-similarity
- **similarity searching** → similarity/diversity
- **Similog keys** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **simple graph** → graph
- **simple matching similarity coefficient** → similarity/diversity (⊙ Table S9)
- **simple mean difference** $\equiv$ *mean difference* → statistical indices (⊙ indices of dispersion)
- **simple random walks** → walk counts
- **Simplest Topological Integers from Molecular Structures** $\equiv$ *STIMS indices* → count descriptors
- **simple topological index** → vertex degree
- **Simpson similarity coefficient** → similarity/diversity (⊙ Table S9)
- **S index** → Schultz molecular topological index
- **single evaluation set technique** → validation techniques (⊙ training/evaluation set splitting)
- **singly occupied molecular orbital** → quantum-chemical descriptors
- **site-property analysis** → Hansch analysis
- **six position number** → steric descriptors (⊙ number of atoms in substituent specific positions)

■ size descriptors

These are → *molecular descriptors* related to the dimension of the molecule and often calculated from the → *molecular geometry*. Combined with molecular shape information, they are closely related to → *steric descriptors*. The simplest size descriptors are → *atom count*, → *bond count*, → *molecular weight*, and some among the → *volume descriptors* such as → *van der Waals volume*. Other size descriptors are → *Sterimol parameters* and → *WHIM size descriptors*.

Moreover, several → *topological indices* are explicitly or implicitly related to the molecular size, for example, → *Wiener index*, → *Zagreb indices*, → *Lovasz–Pelikan index*, and → *connectivity indices*.

Other size descriptors are listed below. Several of them are often used to represent geometrical characteristics of long chain molecules such as polymers and macromolecules [Volkenstein, 1963; Flory, 1969; Arteca, 1991].

• span (R)

A size descriptor defined as the radius of the smallest sphere, centered on the → *center of mass*, completely enclosing all atoms of a molecule [Volkenstein, 1963]:

$$R = \max_i(r_i)$$

where r_i is the distance of the i th atom from the center of mass.

The **average span** descriptor, calculated as the average value of conformational changes and denoted by $\bar{R}$, is used to describe long chain molecules, such as macromolecules, polymers, and proteins, and is related to the Kuhn length (see below).

- **Kuhn length (l)**

For long chain molecules, the Kuhn length is the mean of the $\rightarrow$ bond distances, that is,

$$l = \frac{\sum_{b=1}^B r_b}{B}$$

where B is the number of bonds and r_b is the b th bond distance [Flory, 1969].

It is a size descriptor used for macromolecules, polymers, and proteins. In this case, a useful parameter is also the **contour length** L_C defined as

$$L_C \equiv \text{SBL} = B \cdot l = \sum_{b=1}^B r_b$$

This descriptor was later called **sum bond length** (SBL) and proposed as a simple size descriptor for all molecules [Liu, Liang *et al.*, 2006].

Moreover, for large B values, Kuhn length is related to the average span $\bar{R}$ by the following relationships:

$$\bar{R} \cong B \cdot l \quad \bar{R} = B^{1/2} \cdot l \quad \bar{R} = B^{1/3} \cdot l$$

where the first relationship holds when almost all linear conformations are retained, the second when accessible conformations correspond to randomly folded chains, and the third when the most compact (folded) conformations are retained.

- **end-to-end distance (r_{ee})**

A simple size descriptor for long chain molecules defined as [Flory, 1969]

$$r_{ee} = r_{1A} = ||\mathbf{r}_1 - \mathbf{r}_A||$$

where r_{1A} is the interatomic distance between the first and the last atoms of the chain and $\mathbf{r}$ is the vector of the atom coordinates with respect to the center of mass.

- **persistence length (L_P)**

A size descriptor adopted for long chain molecules such as polymers, determined by both geometrical and topological information [Flory, 1969]. Let a linear chain be defined as a sequence of straight line segments (the bonds): the i th bond is a vector with direction $\mathbf{r}_{i+1} - \mathbf{r}_i$ and the positions of the successive bonds relative to the i th bond are projected along the $\mathbf{r}_{i+1} - \mathbf{r}_i$ direction. Then, the persistence length L_P is defined as the conformational (or configurational) average of the sum of these projections, for any i th bond.

Under the assumption that $L_C/L_P \gg 1$, persistence length is related to the $\rightarrow$ characteristic ratio C'_∞ of the end-to-end distance r_{ee} and to contour length L_C by the following relationships:

$$L_P = \frac{l \cdot C'_\infty}{2} = \frac{\langle r_{ee}^2 \rangle}{2 \cdot B \cdot l} = \frac{\langle r_{ee}^2 \rangle}{2 \cdot L_C}$$

where B and l are the number of bonds and the Kuhn length, respectively.

- **gravitational indices**

→ *Geometrical descriptors* reflecting the mass distribution in a molecule, defined as [Katritzky, Mu *et al.*, 1996b]

$$G_1 = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{m_i \cdot m_j}{r_{ij}^2} \quad G_2 = \sum_{b=1}^B \left(\frac{m_i \cdot m_j}{r_{ij}^2} \right)_b$$

where m_i and m_j are the atomic masses of the considered atoms, r_{ij} the corresponding → *interatomic distances*, and A and B the number of atoms and bonds of the molecule, respectively. The G_1 index takes into account all atom pairs in the molecule whereas the G_2 index is restricted to the pairs of bonded atoms. These indices are related to the bulk cohesiveness of the molecules accounting, simultaneously, for both atomic masses (volumes) and their distribution within the molecular space. For modeling purposes, the square root and cube root of the gravitational indices were also proposed [Wessel, Jurs *et al.*, 1998].

Both indices can be extended to any atomic property other than the atomic mass, such as → *atomic polarizability*, atomic → *van der Waals volume*, and so on.

- **radius of gyration (R_G)**

A size descriptor for the distribution of atomic masses in a molecule [Tanford, 1961; Volkstein, 1963], defined as

$$R_G = \sqrt{\frac{\sum_{i=1}^A m_i \cdot r_i^2}{MW}}$$

where r_i is the distance of the i th atom from the center of mass of the molecule, m_i the corresponding atomic mass, A the atom number, and MW the → *molecular weight*.

The radius of gyration can also be calculated from the → *principal moments of inertia* I ; for planar molecules ($I_C = 0$), it is defined as

$$R_G = \sqrt{\frac{(I_A \cdot I_B)^{1/2}}{MW}}$$

and for nonplanar molecules as

$$R_G = \sqrt{\frac{2\pi \cdot (I_A \cdot I_B \cdot I_C)^{1/3}}{MW}}$$

The radius of gyration is a measure of molecular compactness for long chain molecules such as polymers, that is, small values are obtained when most of the atoms are close to the center of mass. It is also related to the → *characteristic ratio*.

A **size–shape geometrical constant** α_G is derived from the radius of gyration as [Wilding and Rowley, 1986]

$$\alpha_G = -7.706 \cdot 10^{-4} + 0.033 \cdot R_G + 0.01506 \cdot R_G^2 - 9.997 \cdot 10^{-4} \cdot R_G^3$$

• **Meyer anchor sphere volume**

A substituent size descriptor defined as the volume V^a of the portion of the substituent within a sphere centered at the link atom [Meyer, 1986b]. The radius of the sphere was chosen equal to 0.3 nm to comprise the substituent portion responsible for the steric effect of the substituent. It was used, together with the $\rightarrow$ *ovality index* calculated on the substituent, to estimate substituent steric effects; for substituents with equal volume V^a , much larger steric effects are observed for globular substituents.

- **size–shape geometrical constant** $\rightarrow$ size descriptors ($\odot$ radius of gyration)
- **Slater determinant** $\rightarrow$ quantum-chemical descriptors
- **SLOGP** $\rightarrow$ lipophilicity descriptors
- **smallest binary label** $\rightarrow$ canonical numbering
- **Smallest Set of Smallest Rings** $\rightarrow$ ring descriptors
- **SMARTS** $\rightarrow$ molecular descriptors
- **SMF descriptors** $\equiv$ *ISIDA descriptors* $\rightarrow$ substructure descriptors ($\odot$ fingerprints)
- **SMILES** $\rightarrow$ molecular descriptors
- **SMILOGP** $\rightarrow$ lipophilicity descriptors ($\odot$ Broto–Moreau–Vanduyke hydrophobic atomic constants)
- **Snyder descriptors** $\rightarrow$ Linear Solvation Energy Relationships
- **Soergel binary distance** $\rightarrow$ similarity/diversity ($\odot$ Table S10)
- **Soergel distance** $\rightarrow$ similarity/diversity ($\odot$ Table S7)
- **softness density** $\equiv$ *local softness* $\rightarrow$ quantum-chemical descriptors ($\odot$ softness indices)
- **softness indices** $\rightarrow$ quantum-chemical descriptors
- **soil sorption partition coefficient** $\rightarrow$ physico-chemical properties ($\odot$ partition coefficients)
- **soil–water partition coefficient** $\equiv$ *soil sorption partition coefficient* $\rightarrow$ physico-chemical descriptors ($\odot$ partition coefficients)
- **Sokal–Sneath similarity coefficients** $\rightarrow$ similarity/diversity ($\odot$ Table S9)
- **solubility** $\rightarrow$ physico-chemical properties
- **solute aqueous dissolution and solvation descriptors** $\rightarrow$ Membrane Interaction QSAR analysis
- **solute HBA basicity** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- **solute HBD acidity** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- **solute polarity parameter** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ dipolarity/polarizability term)
- **solvation connectivity indices** $\rightarrow$ connectivity indices
- **solvatochromic equation** $\rightarrow$ Linear Solvation Energy Relationships
- **solvatochromic parameters** $\rightarrow$ Linear Solvation Energy Relationships
- **solvent-accessible molecular surface** $\rightarrow$ molecular surface
- **solvent-accessible surface area** $\rightarrow$ molecular surface ($\odot$ solvent-accessible molecular surface)
- **solvent cohesive energy density** $\equiv$ *Hildebrand solubility parameter*
- **solvent-excluded volume** $\rightarrow$ molecular surface ($\odot$ solvent-accessible molecular surface)
- **solvent HBA basicity** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- **solvent HBD acidity** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)

- **solvent polarity scales** → Linear Solvation Energy Relationships (⊖ dipolarity/polarizability term)
- **SOMFA** ≡ *Self-Organizing Molecular Field Analysis* → grid-based QSAR techniques
- **Somoyai function** → quantum-chemical descriptors
- **Sørensen similarity coefficient** ≡ *Dice similarity coefficient* → similarity/diversity (Table S9)

■ **SP indices** (≡ *Subgraph Property indices*)

These are bond additive molecular descriptors derived from the → *H-depleted molecular graph* by an approach similar to that of → *graphical bond order*; they are defined as [Diudea, Minailiuc *et al.*, 1996, 1997a]

$$SP = \sum_{b=1}^B SP_b$$

where the summation runs over all the B graph edges.

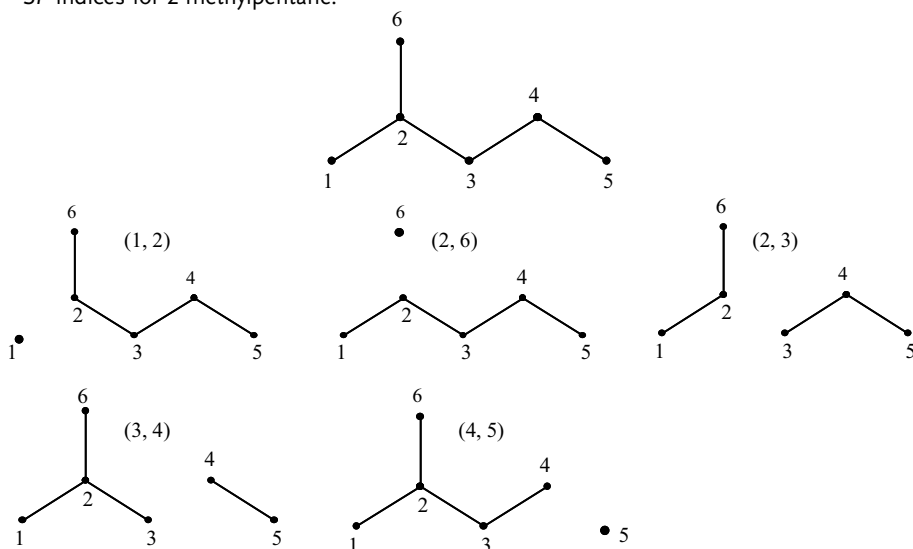
The bond descriptor SP_b is obtained by erasing the b edge from the molecular graph and evaluating the properties of the two remaining subgraphs G_1 and G_2 as the following:

$$SP_b = P^*(G_1) \cdot P^*(G_2) = \frac{\sum_{v \in V_{1,b}} P_v}{\sum_v P_v} \times \frac{\sum_{v \in V_{2,b}} P_v}{\sum_v P_v} = P^*(G_1) \cdot [1 - P^*(G_1)]$$

where P_v is the contribution to the molecular property P due to the v th vertex. P^* of the two subgraphs are the normalized properties obtained by dividing the sum of subgraph vertex contributions by the molecular property calculated on the whole graph. V_1 and V_2 are the subsets of vertices relative to G_1 and G_2 subgraphs, respectively. Possible vertex properties are any → *local vertex invariants*.

Example S5

SP indices for 2-methylpentane.



Atom	${}^1w_i$	δ_i	χ_i^{1w}	
1	1	1	0.5774	
2	1	3	1.7321	
3	1	2	0.9082	w_i : unitary weight
4	1	2	1.2071	δ_i : vertex degree
5	1	1	0.7071	χ_i^{1w} : Randic-Razinger index of first order
6	1	1	0.5774	
$\sum_i P_i$	6	10	5.7093	

Edge	Sl_b	$S\delta_b$	$S\chi_b^w$
(1, 2)	$\frac{1}{6} \cdot \frac{5}{6} = 0.1389$	$\frac{1}{10} \cdot \frac{9}{10} = 0.09$	$\frac{0.5774}{5.7093} \cdot \frac{5.1319}{5.7093} = 0.0909$
(2, 6)	$\frac{1}{6} \cdot \frac{5}{6} = 0.1389$	$\frac{1}{10} \cdot \frac{9}{10} = 0.09$	$\frac{0.5774}{5.7093} \cdot \frac{5.1319}{5.7093} = 0.0909$
(2, 3)	$\frac{3}{6} \cdot \frac{3}{6} = 0.2500$	$\frac{5}{10} \cdot \frac{5}{10} = 0.25$	$\frac{2.8869}{5.7093} \cdot \frac{2.8224}{5.7093} = 0.2499$
(3, 4)	$\frac{4}{6} \cdot \frac{2}{6} = 0.2222$	$\frac{7}{10} \cdot \frac{3}{10} = 0.21$	$\frac{3.7951}{5.7093} \cdot \frac{1.9142}{5.7093} = 0.2229$
(4, 5)	$\frac{1}{6} \cdot \frac{5}{6} = 0.1389$	$\frac{1}{10} \cdot \frac{9}{10} = 0.09$	$\frac{5.0022}{5.7093} \cdot \frac{0.7071}{5.7093} = 0.1085$
SP	0.8889	0.73	0.7631

- **span** → size descriptors
- **spanning tree** → graph
- **spanning-tree density** → Laplacian matrix
- **spanning tree number** → Laplacian matrix
- **sparse matrices** → algebraic operators
- **sparse Wiener matrix** → Wiener matrix
- **Spearman rank correlation coefficient** → statistical indices (⊙ correlation measures)
- **specificity** → classification parameters

■ spectra descriptors

Several spectroscopic techniques are systematically used in chemistry for characterization and recognition of chemicals. This suggests that experimentally recorded spectra contain much information related to the molecular structure, which can be converted into molecular descriptors by proper algorithms. The signals obtained by nuclear magnetic resonance (NMR), namely, ^{13}C - and ^1H -NMR, and mass spectrometry (MS) are sharp enough to be easily transformed into → *vectorial descriptors*.

Intensities of infrared spectra signals (IR spectra), for example, sampled at 10/cm in the fingerprint region (1500–600/cm) were used as molecular descriptors, each spectrum being scaled in the range 0–100 [Benigni, Passerini *et al.*, 1999a].

In QSAR/QSPR studies, the representation of a molecule by using the whole spectrum, without any compression (spectrum compression can be achieved by selecting spectral regions,

by principal component analysis, Fourier analysis, etc.), brings both useless and redundant information resulting into not reliable models.

Fourier analysis transforms spectral data in the wavelength domain into the frequency domain, by approximation of the original spectrum to a desired degree of accuracy by sums of periodic sine and cosine functions of increasing frequency [McClure, Hamid *et al.*, 1984; Pasti, Jouan-Rimbaud *et al.*, 1998; Jetter, Depczynski *et al.*, 2000]. The Fourier approximation of the spectra is characterized by Fourier coefficients that are linear transformations of the original spectral data. Then, a spectral model can be established between these coefficients and the independent variable by multiple regression analysis. The most important feature of the Fourier analysis is reduction of multicollinearity and dimension of the original spectrum. However, the Fourier coefficients bear no simple relationship to individual features of the spectrum so that it will not be clear what information is being used in calibration.

Given a matrix $\mathbf{X}(n, p)$, where n is the number of molecules and p is the number of vector descriptors (e.g., spectrum signals), each row of the matrix is transformed into cosine and sine terms according to the following equation:

$$F_{ix}(k) = \sum_{j=1}^p [\mathbf{X}]_{ij} \cdot \cos(\phi_{kj}) \quad F_{iy}(k) = \sum_{j=1}^p [\mathbf{X}]_{ij} \cdot \sin(\phi_{kj})$$

where F_{ix} and F_{iy} are the two Fourier coefficients describing the i th molecule and ϕ is defined as

$$\phi_{kj} = \frac{2 \cdot \pi \cdot k \cdot (j-1)}{p} \quad k = 1, 2, \dots, p$$

Depending on the size p of the matrix $\mathbf{X}(n, p)$, a different number of linearly independent Fourier coefficients are obtained: if p is odd, then this number is $(p-1)/2$, if p is even, then there are $(p-2)/2$ independent Fourier coefficients.

Wavelet analysis is the representation of a function by wavelets that are mathematical functions used to divide a given function into different frequency components and study each component with a resolution that matches its scale [Depczynski, Jetter *et al.*, 1997, 1999b; Alsberg, 2000; Jetter, Depczynski *et al.*, 2000; Walczak, 2000].

The wavelets are scaled and translated copies (known as “daughter wavelets”) of a finite-length or fast-decaying oscillating waveform (known as the “mother wavelet”). Wavelet transforms have advantages over traditional Fourier transforms for representing functions that have discontinuities and sharp peaks.

Wavelet transforms (WT) are classified into *continuous wavelet transforms* (CWTs) and *discrete wavelet transforms* (DWTs). Wavelet is defined as the dilation and translation of the basis function $\psi(t)$, and the **continuous wavelet transforms** is defined as [Shao, Leung *et al.*, 2003]

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \cdot \psi\left(\frac{t-b}{a}\right) \quad a, b \in R, a \neq 0$$

where a and b are the scale (dilation) and the position (translation) parameters, respectively. If a and b are discretized in such a way that $a = a_0^j$ and $b = k \cdot b_0 \cdot a_0^j$, the **discrete wavelet transforms** (DWTs) are obtained:

$$\Psi_{j,k}(t) = \frac{1}{a_0^{j/2}} \cdot \psi\left(\frac{t}{a_0^j} - k \cdot b_0\right)$$

CWTs operate over every possible scale and translation whereas DWTs use a specific subset of scale and translation values.

Generally, the DWT is used for data compression and the CWT for signal analysis. Wavelet transforms are applied in several fields, such as molecular dynamics, quantum-chemical calculations, optics, image analysis, DNA and protein sequence analysis, and seismic geophysics.

Wavelet analysis has been studied in NIR spectra data processing [Chen and Wang, 2001]. It proves to be more powerful than Fourier transform in capturing the local features of a spectrum because it can decompose a signal into components that are well localized in both the time and frequency domains, whereas Fourier transform can only reflect upon the frequency information. Performing a wavelet analysis of a signal yields a vector of wavelet coefficients that are assigned to different frequency bands. Each band expands over the complete wavelength domain and reflects upon a certain frequency range of the signal. By selecting appropriate wavelet coefficients, similarly a spectral model can be established by regression of the coefficients against the independent variable.

Some applications of wavelet analysis are discussed in Refs. [Jouan-Rimbaud, Walczak *et al.*, 1997; Wold and Sjöström, 1998; Cocchi, Seeber *et al.*, 2001; Wold, Trygg *et al.*, 2001; Cocchi, Seeber *et al.*, 2003; Liu and Brown 2004; Cocchi, Corbellini *et al.*, 2005; Esteban-Diez, González-Sáiz *et al.*, 2006a; Tabaraki, Khayamian *et al.*, 2006; Yiyu, Minjun *et al.*, 2003].

A mass spectrum of a compound can be transformed into a vectorial molecular descriptor with components x_j , with $j = 1, 2, 3, \dots, p$, each component x_j being the peak intensities at selected masses (in percent of the base peak), or weighted peak intensities, or some spectral features derived from the peak intensities [Demuth, Karlovits *et al.*, 2004]. Different weighting schemes for peak intensities have been proposed; for instance, the vector component x_j is calculated from peak intensity I_m at mass m as

$$x_j = m^\alpha \cdot I_m^\beta$$

where I_m is the peak intensity, normalized to the base peak with an intensity of 100%, exponent α has been varied between 0 and 2, and exponent β between 0.01 and 2.

A **spectral feature** is a characteristic number that can be automatically computed from a spectrum by using some data transformation. Then, the spectrum is represented by a set of variables (i.e., spectral features) that are more closely related to chemical structure properties than the original spectral data [Werther, Demuth *et al.*, 2002]. **Mass spectral features** are the most significant variables derived from a mass spectrum; they are divided into nine groups given in Table S13. Group 1 contains peak intensities at selected masses; whereas group 2 contains the same peak intensities but normalized to the local ion current, which is the sum of peak intensities in a mass interval $\pm \Delta m$ around a considered mass m . Group 3 contains averaged peak intensities of mass intervals. Group 4 contains features calculated as $\ln(I_m/I_{m+\Delta m})$; peak intensities below 1% are set to 1 to avoid divisions by zero. Group 5 contains modulo-14 descriptors, which belong to the class of **modulo- L descriptors**; these were proposed as *uniform-length descriptors* to condense the chemical information of a molecule characterized by mass spectra [Scsibrany and Varmuza, 1992b]. For each mass spectrum, L descriptors H_k ($k = 1, L$) are calculated by summing the peak intensities I_m at masses m with a difference of L :

$$H_k = \left(\sum_m I_m \right)_k \quad k = m \bmod L$$

These descriptors are successively normalized to a constant sum equal to one for each spectrum. Modulo-14 descriptors were proposed as optimal features. In this case, the first descriptor H_1 is obtained by summing the peak intensities at masses 1, 15, 29, ...; the fourteenth descriptor H_{14} by summing the peak heights at masses 14, 28, 42, ... These descriptors were used, after Principal Component Analysis, to find maximum common substructures in large mass spectrometric databases.

Group 6 contains $\rightarrow$ *autocorrelation descriptors* derived from peak intensities. Features of group 7 are the relative peak intensities in the low mass range, the base peak intensity in percentage of the total intensity sum, and the proportion of peak intensities summed at even mass numbers. Group 8 contains the features that indicate the presence of a defined target peak pattern (e.g., isotope peak pattern); each of these features is calculated as

$$x_j = 100 \cdot \max(r_m^3)$$

where r is the correlation coefficient calculated from the peak intensities of the selected peak pattern (shifted across the spectrum) and the actual peak intensities in the spectrum, starting at mass m .

Finally, group 9 contains features that indicate the joint presence of peaks at defined mass numbers; these features are calculated as

$$x_j = \frac{1}{n} \cdot \sum_m I'_m \quad I'_m = \begin{cases} \frac{100 \cdot (I_m - I_0)^\lambda}{(100 - I_0)^\lambda} & \text{if } I_m > I_0 \\ 0 & \text{if } I_m \leq I_0 \end{cases}$$

where I'_m is the scaled peak intensity at mass m of the target peak pattern, I_0 an intensity threshold (e.g., 5% of the base peak intensity), n the number of peaks in the target peak pattern and the summation goes over all target peaks, and λ is a variable parameter (e.g., 0.2).

Table S13 Mass spectral features.

Group	Description	No.
1	Intensities at masses 12,13, 15, 17, 19–27, 29–31, 33–200	184
2	Normalized intensities at masses 12,13, 15, 17, 19–27, 29–31, 33–200 ($\Delta m = \pm 3$)	184
3	Averaged intensities of mass intervals 33–50, 51–70, 71–100, 101–150	4
4	Log intensity ratios for mass differences of 1 and 2, and lower masses 39–150	224
5	Modulo-14 descriptors for mass intervals 31–120, 121–800, 31–800	42
6	Autocorrelation for mass differences 1, 2, 14–60, and mass intervals 31–120, 100–800, 31–800	147
7	Spectra type	3
8	Isotope peak patterns for $\text{Cl}_1\text{--Cl}_5$, $\text{Br}_1\text{--Br}_5$, and Cl_xBr_y ($x + y = 2, 3, 4, 5$) up to mass 800	20
9	Characteristic peak groups	54
Sum		862

In the **Comparative Spectra Analysis (CoSA)**, digitalized spectra were used as molecular descriptors for QSAR/QSPR modeling [Bursi, Dao *et al.*, 1999; Bursi, Verwer *et al.*, 2001; Beger

and Wilkes, 2001b]. More specifically, experimental mass, IR, and ^1H -NMR spectra and simulated IR and ^{13}C -NMR spectra were used as 3D molecular descriptors.

For each molecule, experimental and simulated spectra are digitalized by recording into separate arrays signal intensities at given intervals of the spectrum.

For simulated IR and ^{13}C -NMR spectra, no intensities were provided and all intensities were set to 1. The sampling interval (L) determines the level of detail that is preserved in the digitized spectrum. Finally, the spectrum is normalized to obtain the same total intensity for each spectrum, independent of the size of the molecule.

In the CoSA model, there are 256 bins, each of which is populated or does not depend on the pattern of simulated chemical shifts. This approach does not require an identification of the shift with the carbon that produced it.

The **Comparative Structurally Assigned Spectral Analysis (CoSASA)** is another QSAR approach that is based on ^{13}C -NMR spectra where chemical shifts at selected positions in the molecule backbone template are previously identified [Beger, Buzatu *et al.*, 2001, 2002; Beger and Wilkes, 2001a]. This approach combines structural information from molecules with the assigned simulated ^{13}C -NMR chemical shifts.

In both CoSA and CoSASA approaches, models are obtained by using Partial Least Squares (PLS) regression on the original chemical shifts or on the PCs obtained by Principal Component Analysis.

Chemical Shift Sum (CSS) was proposed as single molecular descriptor derived from spectra as the sum of the chemical shifts of the carbon atoms from ^{13}C -NMR spectroscopy [Randić, 1980a]. It was shown that at least for alkanes, the chemical shift sum varies regularly with some physico-chemical properties of alkanes. QSAR models for CSS, based on artificial neural networks, were proposed by using the first four path counts [Ivanciuc, Rabine *et al.*, 1997].

NMR spectra were also characterized for $\rightarrow$ *similarity/diversity* analysis [Župerl, Pristovšek *et al.*, 2007] by $\rightarrow$ *graph invariants* obtained using the sequential nearest neighbor method proposed to characterize $\rightarrow$ *proteomics maps* [Randić, Novič *et al.*, 2005]. Moreover, selected chemical shifts were directly used as molecular descriptors for modeling lipophilicity of alcohols [Khadikar, Sharma *et al.*, 2005b].

📖 [Klawun and Wilkins, 1996a, 1996b; Selzer, Schuur *et al.*, 1996; Meiler, Sanli *et al.*, 2002; Thomsen, Dobel *et al.*, 2002; Steinbeck, 2003; Khadikar, Sharma *et al.*, 2005a]

- **spectral diameter** $\rightarrow$ spectral indices
- **spectral feature** $\rightarrow$ spectra descriptors

■ **spectral indices** ($\equiv$ *eigenvalue-based descriptors*)

These are molecular descriptors defined in terms of the eigenvalues of a square $\rightarrow$ *graph-theoretical matrix* $\mathbf{M}$ of size $(n \times n)$ [Hall, 1981, 1986, 1992, 1993; Lee and Yeh, 1993; Cvetković, Doob *et al.*, 1995; Randić, Vračko *et al.*, 2001]. The eigenvalues are the roots of the $\rightarrow$ *characteristic polynomial* of the matrix $\mathbf{M}$ and the set of the eigenvalues is the matrix spectrum $\Lambda(\mathbf{M}) = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$.

The most common eigenvalue functions used to derive spectral indices are given below in a general form, which can be applied to any molecular matrix $\mathbf{M}(w)$, calculated with the $\rightarrow$ *weighting scheme* w [Ivanciuc and Balaban, 1999c; Ivanciuc, 1999c, 2001c, 2002b, 2003d;

Ivanciuc, Ivanciuc *et al.*, 1999a; Consonni and Todeschini, 2008]:

$$\begin{aligned}
 SpSum^k(\mathbf{M}, w) &= \sum_{i=1}^n |\lambda_i|^k & SpSum_+^k(\mathbf{M}, w) &= \sum_{i=1}^{n^+} (\lambda_i^+)^k & SpSum_-^k(\mathbf{M}, w) &= \sum_{i=1}^{n^-} |\lambda_i^-|^k \\
 SpAD(\mathbf{M}, w) &= \sum_{i=1}^n |\lambda_i - \bar{\lambda}| & SpMAD(\mathbf{M}, w) &= \sum_{i=1}^n |\lambda_i - \bar{\lambda}| / n \\
 MinSp(\mathbf{M}, w) &= \min_i \{\lambda_i\} & MaxSp(\mathbf{M}, w) &= \max_i \{\lambda_i\} \\
 MaxSpA(\mathbf{M}, w) &= \max_i \{|\lambda_i|\} & SpDiam(\mathbf{M}, w) &= MaxSp - MinSp
 \end{aligned}$$

where k is a parameter, usually taken equal to one; for negative values of k , eigenvalues equal to zero must not be considered. For $k = 1$, $SpSum$ is the sum of the n absolute values of the spectrum eigenvalues; this quantity calculated on the $\rightarrow$ *adjacency matrix* of simple graphs $SpSum(\mathbf{A})$ was called **graph energy** and denoted by E [Gutman, 1978b, 2001, 2005; Gutman, Vidović *et al.*, 2003a; Zhou, 2004b; Yu, Lu *et al.*, 2005; Zhou, Gutman *et al.*, 2007]; the same quantity derived from the $\rightarrow$ *Laplacian matrix* was called **Laplacian graph energy** [Gutman and Zhou, 2006; Zhou and Gutman, 2007]. $SpSum_+$ is the sum of the n^+ positive eigenvalues, $SpSum_-$ is the sum of the absolute values of the n^- negative eigenvalues. $SpAD$ is the sum of the absolute deviations of the eigenvalues from their mean and is called **generalized graph energy** [Consonni and Todeschini, 2008]; $SpMAD$ is the mean absolute deviation and is called **generalized average graph energy** [Consonni and Todeschini, 2008]. $MinSp$ is the minimum eigenvalue, $MaxSp$ is the maximum eigenvalue, called **leading eigenvalue** or **spectral radius**, $MaxSpA$ is the maximum absolute value of the spectrum, and $SpDiam$ is the **spectral diameter** of the molecular matrix, defined as the difference between $MaxSp$ and $MinSp$. These kinds of functions were called by Ivanciuc **matrix spectrum operators** [Ivanciuc, 2003d].

It has been demonstrated that the leading eigenvalue of a symmetric matrix $\mathbf{M}$ is bounded from above and from below by its largest and smallest row sum:

$$\min_i [VS_i(\mathbf{M})] \leq MaxSp(\mathbf{M}) \equiv \lambda_1(\mathbf{M}) \leq \max_i [VS_i(\mathbf{M})]$$

where VS is the $\rightarrow$ *row sum operator*.

$\rightarrow$ *Spectral moments* of the matrix $\mathbf{M}(w)$ are other molecular descriptors defined in terms of the k th power of the eigenvalues:

$$\mu^k(\mathbf{M}; w) = \sum_{i=1}^n \lambda_i^k$$

where $k = 1, \dots, n$ is the order of the spectral moment. It is noteworthy that for even k values, spectral moments μ^k coincide with spectral indices $SpSum^k$.

Spectral indices defined above and spectral moments were tested in QSAR/QSPR, calculated from a number of graph-theoretical matrices such as $\rightarrow$ *adjacency matrix* [Ivanciuc and Balaban, 1999c], $\rightarrow$ *distance matrix* [Ivanciuc and Balaban, 1999c], $\rightarrow$ *reciprocal distance matrix* [Ivanciuc and Balaban, 1999c], $\rightarrow$ *Laplacian matrix* [Ivanciuc and Balaban, 1999, 1999c; Ivanciuc, 2001g], $\rightarrow$ *edge adjacency matrix* [Ivanciuc and Balaban, 1999c], $\rightarrow$ *distance-path matrix* [Ivanciuc and Ivanciuc, 1999; Ivanciuc and Balaban, 1999c; Ivanciuc, 2001e], $\rightarrow$ *distance-delta matrix* [Ivanciuc and Ivanciuc, 1999], $\rightarrow$ *Szeged matrix* [Ivanciuc, 2002b], $\rightarrow$ *reverse*

Wiener matrix [Ivanciuc, Ivanciuc *et al.*, 2002b], $\rightarrow$ distance–valency matrices [Ivanciuc, 1999c], $\rightarrow$ geometry matrix [Ivanciuc and Balaban, 1999c], $\rightarrow$ Burden matrix [Ivanciuc and Balaban, 1999c; Ivanciuc, 2001f], $\rightarrow$ complementary distance matrix [Ivanciuc, Ivanciuc *et al.*, 2000a], $\rightarrow$ resistance distance matrix, and $\rightarrow$ conductance matrix [Ivanciuc, 2000h].

Examples of spectral indices are $\rightarrow$ WHIM descriptors and $\rightarrow$ G-WHIM descriptors of size and shape, $\rightarrow$ EA indices, $\rightarrow$ quasi-Wiener index, $\rightarrow$ Mohar indices, several $\rightarrow$ shape descriptors, $\rightarrow$ EVA descriptors, $\rightarrow$ EEVA descriptors, and $\rightarrow$ 4D-MS descriptors.

Other popular spectral indices are given below.

• eigenvalues of the adjacency matrix

The eigenvalues of the $\rightarrow$ adjacency matrix **A** can be used as molecular descriptors. These eigenvalues take both positive and negative values, their sum being equal to zero.

Level pattern indices (LPI) of molecules are defined in Hückel theory as the numbers of bonding n^+ , nonbonding n^0 , and antibonding n^- molecular orbitals and correspond to the numbers of positive, zero, and negative eigenvalues of the adjacency matrix of the molecule [Jiang and Zhu, 1994]. The level pattern indices relate to the stability and valence state of the molecule: if $n^0 > 0$, the molecule will be either a free radical or an unstable species with a low $\rightarrow$ HOMO–LUMO energy gap; if $n^0 = 0$ and $n^+ \neq n^-$, the molecule in the ground state will be an anion or cation. The presence of nonbonding orbitals n^0 can be easily detected by the determinant of the adjacency matrix:

$$\det(\mathbf{A}) = \prod_{i=1}^A \lambda_i = 0$$

The sum of the level pattern indices corresponds to the number *A* of atoms in the molecule: $A = n^+ + n^0 + n^-$. Moreover, the eigenvalues may be closely related to the π -electron energy of the molecule:

$$E_\pi = \sum_{i=1}^A g_i \cdot \lambda_i$$

where *A* is the number of atoms, λ_i the eigenvalues of the adjacency matrix, and g_i the occupation number on the *i*th molecular orbital, which can take values 0, 1, or 2.

The largest eigenvalue of adjacency matrix **A** is among the most popular graph invariants and is known as the **Lovasz–Pelikan index** λ_1^{LP} [Lovasz and Pelikan, 1973] (Table S14):

$$\lambda_1^{\text{LP}} \equiv \text{MaxSp}(\mathbf{A})$$

This eigenvalue has been suggested as an index of $\rightarrow$ molecular branching, the smallest values corresponding to chain graphs and the highest to the most branched graphs. It is not a very discriminant index because in many cases the same value is obtained for two or more nonisomorphic graphs.

Moreover, the coefficients of the eigenvector associated with the largest eigenvalue were used by Randić for $\rightarrow$ canonical numbering of graph vertices.

Balaban proposed [Balaban, Ciubotariu *et al.*, 1991] the use of the coefficients of the eigenvector associated with the lowest (largest negative) eigenvalue as $\rightarrow$ local vertex invariants able to provide discrimination among graph vertices; lower values correspond to vertices of lower degree, farther from the center or from a vertex of high degree. Some molecular

descriptors based on eigenvalues and corresponding eigenvectors of the adjacency matrix have also been proposed. The **VAA indices** were defined as

$$\text{VAA1} = \text{SpSum}^+(\mathbf{A}) = \sum_{i=1}^{n^+} \lambda_i^+ \quad \text{VAA2} = \frac{\text{VAA1}}{A} \quad \text{VAA3} = \frac{A}{10} \cdot \log(\text{VAA1})$$

where λ^+ are the positive eigenvalues and n^+ the number of positive eigenvalues; A is the number of molecular graph vertices.

The **VEA indices** are defined by the coefficients (i.e., loadings) ℓ_{iA} of the eigenvector associated with the largest negative eigenvalue (i.e., the A th eigenvalue of the decreasing eigenvalue sequence):

$$\text{VEA1} = \sum_{i=1}^A \ell_{iA} \quad \text{VEA2} = \frac{\text{VEA1}}{A} \quad \text{VEA3} = \frac{A}{10} \cdot \log(\text{VEA1})$$

The **VRA indices** are $\rightarrow$ *Randić-like indices* defined in terms of the coefficients ℓ_{iA} of the eigenvector associated with the largest negative eigenvalue:

$$\text{VRA1} = \sum_b (\ell_{iA} \cdot \ell_{jA})_b^{-1/2} \quad \text{VRA2} = \frac{\text{VRA1}}{A} \quad \text{VRA3} = \frac{A}{10} \cdot \log(\text{VRA1})$$

where the summation runs over all of the edges in the molecular graph; ℓ_{iA} and ℓ_{jA} are the LOVIs of the two vertices incident to the considered edge.

$\rightarrow$ *Schultz-type indices* that are based on the eigenvector associated with the lowest (largest negative) eigenvalue of the adjacency matrix have also been proposed.

📖 [Balaban, 1992; Hall, 1992; Gineityte, 1998]

Table S14 Some spectral indices derived from the adjacency matrix, distance matrix, and Barysz distance matrix for the data set of phenethylamines (Appendix C – Set 2).

Molecules	X	Y	λ_1^{LP}	VEA1	VEA2	VRA1	VED1	VED2	VRD1	MaxSp(^2D)	SpSum $^1(^2\text{D})$
1	H	H	2.236	3.187	0.266	11.489	3.416	0.285	11.575	28.733	57.023
2	H	F	2.266	3.297	0.254	12.478	3.549	0.273	12.579	33.146	65.516
3	H	Cl	2.266	3.297	0.254	12.478	3.549	0.273	12.579	32.500	63.911
4	H	Br	2.266	3.297	0.254	12.478	3.549	0.273	12.579	32.141	63.600
5	H	I	2.266	3.297	0.254	12.478	3.549	0.273	12.579	32.028	63.556
6	H	Me	2.266	3.297	0.254	12.478	3.549	0.273	12.579	33.862	67.282
7	F	H	2.272	3.287	0.253	12.477	3.554	0.273	12.580	32.427	64.078
8	Cl	H	2.272	3.287	0.253	12.477	3.554	0.273	12.580	31.821	62.554
9	Br	H	2.272	3.287	0.253	12.477	3.554	0.273	12.580	31.487	62.302
10	I	H	2.272	3.287	0.253	12.477	3.554	0.273	12.580	31.383	62.275
11	Me	H	2.272	3.287	0.253	12.477	3.554	0.273	12.580	33.105	65.768
12	Cl	F	2.312	3.370	0.241	13.465	3.685	0.263	13.583	35.968	70.514
13	Br	F	2.312	3.370	0.241	13.465	3.685	0.263	13.583	35.650	70.304
14	Me	F	2.312	3.370	0.241	13.465	3.685	0.263	13.583	37.203	73.631
15	Cl	Cl	2.312	3.370	0.241	13.465	3.685	0.263	13.583	35.332	68.928
16	Br	Cl	2.312	3.370	0.241	13.465	3.685	0.263	13.583	35.011	68.721

(Continued)

Table S14 (Continued)

Molecules	X	Y	λ_1^{LP}	VEA1	VEA2	VRA1	VED1	VED2	VRD1	$\text{MaxSp}^1(\mathbf{ZD})$	$\text{SpSum}^1(\mathbf{ZD})$
17	Me	Cl	2.312	3.370	0.241	13.465	3.685	0.263	13.583	36.579	72.070
18	Cl	Br	2.312	3.370	0.241	13.465	3.685	0.263	13.583	34.980	68.652
19	Br	Br	2.312	3.370	0.241	13.465	3.685	0.263	13.583	34.657	68.438
20	Me	Br	2.312	3.370	0.241	13.465	3.685	0.263	13.583	36.234	71.797
21	Me	Me	2.312	3.370	0.241	13.465	3.685	0.263	13.583	37.900	75.358
22	Br	Me	2.312	3.370	0.241	13.465	3.685	0.263	13.583	36.363	72.062

λ_1^{LP} is the Lovasz–Pelikan index; VEA1, VEA2, and VRA1 are the VEA indices and VRA indices from the adjacency matrix, respectively; VED1, VED2, and VRD1 are the VED indices and VRD indices from the distance matrix, respectively; $\text{MaxSp}^1(\mathbf{ZD})$ and $\text{SpSum}^1(\mathbf{ZD})$ are the leading eigenvalue and the sum of the absolute eigenvalues of the Barysz distance matrix, respectively.

• subgraph centrality

The subgraph centrality of the vertex v_i is a modification of the $\rightarrow$ vertex structural code SC_i that is defined as the sum of self-returning walks of different lengths starting and ending at vertex v_i [Estrada and Rodríguez-Velásquez, 2005a, 2005b, 2006b; Estrada, 2006b]:

$$C_S(i) = \sum_{k=0}^{\infty} \frac{\text{srw}_i^{(k)}}{k!}$$

where $\text{srw}_i^{(k)}$ is the self-returning walk count of k th order for the i th vertex and each count $\text{srw}_i^{(k)}$ is weighted so that shorter self-returning walks have more influence on the centrality of the vertex than longer self-returning walks. The subgraph centrality of a vertex is a variant of the $\rightarrow$ eigenvector centrality that is a measure of vertex centrality; in effect, it was demonstrated that the subgraph centrality of a vertex can also be calculated from the eigenvalues λ and coefficients ℓ of the eigenvectors of the $\rightarrow$ adjacency matrix $\mathbf{A}$ of the molecular graph being comprised of A vertices as follows:

$$C_S(i) = \sum_{j=1}^A (\ell_{ij})^2 \cdot e^{\lambda_j}$$

where ℓ_{ij} is the i th component of the eigenvector associated to the j th eigenvalue λ_j .

The subgraph centrality of the molecular graph, called **Estrada index**, is defined as average subgraph centrality as [Estrada and Hatano, 2007; Gutman, Estrada *et al.*, 2007; Gutman, Radenković *et al.*, 2007]

$$\text{EE} \equiv C_S(\mathbf{G}) = \sum_{i=1}^A C_S(i) = \sum_{i=1}^A e^{\lambda_i}$$

The advantage of the exponential function is the possibility to take contemporarily into account both positive and negative eigenvalues, without compensation effects.

The Estrada index is a molecular descriptor encoding information on complexity of molecular graphs and is also used to describe characteristic features of complex systems of physico-chemical interest, such as reaction, metabolic, and protein–protein interaction networks.

The $\rightarrow$ protein folding degree index is an application of the Estrada index to the description of protein and long chain biopolymers folding. It is calculated from the eigenvalues of a matrix obtained by summing the adjacency matrix of third order $\rightarrow$ line graph and a diagonal matrix whose diagonal elements are the cosines of the dihedral angles, associated to the vertices in the

considered line graph [Estrada, 2000, 2002a, 2004a, 2004b, 2007; Estrada and Uriarte, 2005; Estrada and Rodríguez-Velásquez, 2006a; Estrada, Uriarte *et al.*, 2006].

• Burden eigenvalues

These are molecular descriptors defined as eigenvalues of a modified connectivity matrix, which is the $\rightarrow$ *Burden matrix* **B** [Burden, 1989]. The matrix **B** representing a $\rightarrow$ *H-depleted molecular graph* is defined as the following: the diagonal elements B_{ii} are the atomic numbers Z_i of the atoms; the off-diagonal elements B_{ij} representing two bonded atoms i and j are equal to π^*10^{-1} , where π^* is the $\rightarrow$ *conventional bond order*, that is, 0.1, 0.2, 0.3, and 0.15 for a single, double, triple, and aromatic bond, respectively; off-diagonal elements B_{ij} corresponding to terminal bonds are augmented by 0.01; and all other matrix elements are set at 0.001.

The ordered sequence of the n smallest eigenvalues of the **B** matrix was proposed as a molecular descriptor with high discrimination power to be used in the recognition and ordering of molecular structures. The basic assumption was that the lowest eigenvalues contain contributions from all atoms and thus reflect the topology of the whole molecule.

The diagonal elements of the **B** matrix can be set in different ways to account for different features of the molecule. It was proposed that a matrix **B** could represent a $\rightarrow$ *hydrogen-filled molecular graph*: the diagonal elements are roughly proportional to the electronegativity of the atoms, based on an electronegativity scale where the carbon atom value is assumed equal to zero, whereas the off-diagonal terms corresponding to pairs of bonded atoms are the square roots of conventional bond order, that is, $B_{ij} = \sqrt{\pi_{ij}^*}$.

Including information on the electronic environment of the atoms in the matrix should relate the matrix eigenvalues to the electronic distribution of the whole molecule. This led to the proposing of the **Chemically Intuitive Molecular index (CIM index)**: this is an ordered sequence of the n largest absolute eigenvalues of the matrix **B**, based on atomic electronegativities as defined above (the same electronegativity value, $\chi = 2.3$, was used for all the halogen atoms) [Burden, 1997].

This vector descriptor was later improved [Benigni, Passerini *et al.*, 1999b] by a more extended electronegativity scale, which distinguished halogens and took sulfur and phosphorus atoms into consideration (Table S15).

Table S15 Atom electronegativity values for the Burden matrix.

Atom	χ	Atom	χ	Atom	χ
C	0.00	F	2.30	S	0.50
H	0.15	Cl	0.90	P	0.50
N	0.90	Br	0.80		
O	0.90	I	0.50		

The **Burden modified eigenvalues** are the largest absolute eigenvalues of the above defined matrix:

$$\text{BME} \equiv \{\lambda_1, \lambda_2, \dots, \lambda_L\}$$

where L is a user-defined maximum length (e.g., $L = 15$) resulting in $\rightarrow$ *uniform-length descriptors* (Table S16).

Proposed to address searching for chemical $\rightarrow$ *similarity/diversity* on large databases [Pearlman and Smith, 1998; Pearlman, 1999], **BCUT descriptors** (Burden – CAS – University

of Texas eigenvalues) are based on a significant extension of the Burden approach, considering three classes of matrices whose diagonal elements correspond to (1) atomic charge-related values, (2) atomic polarizability-related values, and (3) atomic H-bond abilities. In addition, a variety of definitions were considered for the off-diagonal terms, including functions of interatomic distances, overlaps, computed bond-orders, and so on. Moreover, for the off-diagonal terms, not only was a 2D approach used, but also a 3D approach, to account for geometric $\rightarrow$ interatomic distances.

Further types of atomic features could also be considered, together with other proximity measures and weighting schemes. Among the eigenvalues obtained from each of these matrices, the highest and the lowest have been demonstrated to reflect relevant aspects of molecular structure, and are therefore useful for similarity searching.

A modified Burden matrix $\mathbf{Q}$ was also defined as follows [Sheridan, 2002]:

$$[\mathbf{Q}]_{ii} = Z_i + 0.1 \cdot \delta_i + 0.01 \cdot n_i^\pi \quad \text{and} \quad [\mathbf{Q}]_{ij} = 0.4/d_{ij}$$

where Z_i is the atomic number of the i th atom, δ_i the number of non-hydrogen neighbors of the i th atom (i.e., the $\rightarrow$ vertex degree), n_i^π the number of π electrons, and d_{ij} the topological distance between the i th and j th atoms.

Then, a hash string, describing the molecule, is obtained by concatenation of the highest and lowest eigenvalues of the matrix $\mathbf{Q}$ expressed to six decimal places. This string representation of compounds is suitable for similarity/diversity analysis.

Table S16 Some Burden eigenvalues derived from the Burden matrix weighted by relative atomic masses for the data set of phenethylamines (Appendix C – Set 2).

Molecules	X	Y	$\lambda_1^+(\mathbf{B}, \mathbf{m})$	$\lambda_2^+(\mathbf{B}, \mathbf{m})$	$\lambda_3^+(\mathbf{B}, \mathbf{m})$	$\lambda_4^+(\mathbf{B}, \mathbf{m})$	$\lambda_5^+(\mathbf{B}, \mathbf{m})$	$\lambda_1^-(\mathbf{B}, \mathbf{m})$	$\lambda_2^-(\mathbf{B}, \mathbf{m})$	$\lambda_3^-(\mathbf{B}, \mathbf{m})$	$\lambda_4^-(\mathbf{B}, \mathbf{m})$	$\lambda_5^-(\mathbf{B}, \mathbf{m})$
1	H	H	6.841	3.763	3.349	2.776	2.617	1.958	1.699	1.323	1.249	1.082
2	H	F	6.841	3.790	3.355	2.862	2.617	1.936	1.694	1.304	1.249	1.082
3	H	Cl	6.841	3.915	3.406	3.219	2.617	1.928	1.692	1.300	1.249	1.082
4	H	Br	6.844	6.839	3.699	3.338	2.677	1.921	1.691	1.295	1.249	1.082
5	H	I	10.673	6.841	3.716	3.341	2.693	1.918	1.690	1.294	1.249	1.082
6	H	Me	6.841	3.799	3.357	2.939	2.617	1.979	1.708	1.437	1.249	1.227
7	F	H	6.841	3.791	3.353	2.827	2.698	1.935	1.695	1.315	1.249	1.018
8	Cl	H	6.841	3.917	3.392	3.223	2.744	1.928	1.694	1.314	1.249	1.007
9	Br	H	6.845	6.838	3.697	3.341	2.756	1.920	1.692	1.312	1.249	0.999
10	I	H	10.673	6.841	3.714	3.343	2.759	1.917	1.692	1.312	1.249	0.997
11	Me	H	6.841	3.800	3.355	2.909	2.726	1.980	1.706	1.428	1.287	1.249
12	Cl	F	6.841	3.938	3.393	3.226	2.830	1.907	1.688	1.294	1.249	1.004
13	Br	F	6.845	6.838	3.722	3.348	2.847	1.899	1.686	1.293	1.249	0.994
14	Me	F	6.841	3.826	3.360	2.931	2.782	1.958	1.702	1.428	1.270	1.249
15	Cl	Cl	6.841	4.026	3.407	3.332	3.187	1.900	1.686	1.290	1.249	1.003
16	Br	Cl	6.845	6.838	3.853	3.406	3.215	1.893	1.684	1.288	1.249	0.993
17	Me	Cl	6.841	3.943	3.406	3.224	2.875	1.951	1.701	1.428	1.267	1.249
18	Cl	Br	6.844	6.839	3.856	3.391	3.218	1.893	1.683	1.285	1.249	1.003
19	Br	Br	6.883	6.841	6.800	3.642	3.329	1.886	1.682	1.284	1.249	0.993
20	Me	Br	6.844	6.839	3.734	3.346	2.894	1.943	1.699	1.428	1.263	1.249
21	Me	Me	6.841	3.835	3.362	2.965	2.826	1.999	1.713	1.442	1.412	1.249
22	Br	Me	6.845	6.838	3.731	3.351	2.931	1.942	1.704	1.437	1.249	1.221

☞ [Menard, Lewis *et al.*, 1998; Menard, Mason *et al.*, 1998; Burden and Winkler, 1999b; Pearlman and Smith, 1999; Stanton, 1999; Stanton, Morris *et al.*, 1999; Pirard and Pickett, 2000; Benigni, Giuliani *et al.*, 2001; Burden, 2001; Gao, 2001; Ivanciuc, 2001f; Feng, Lurati *et al.*, 2003; Jensen, Sørensen *et al.*, 2003; Manallack, Tehan *et al.*, 2003; Pino, Giuliani *et al.*, 2003; Randić, 2003b; Young, 2003; Drakulić, Juranić *et al.*, 2005; Gupta and Prabhakar, 2005; Saiz-Urra, Pérez González *et al.*, 2006; Ijjaali, Petit *et al.*, 2007]

• eigenvalues of the distance matrix

The largest eigenvalue of the topological $\rightarrow$ distance matrix **D** representing a $\rightarrow$ *H-depleted molecular graph* was proposed as a molecular descriptor, and was called **leading eigenvalue of the distance matrix** [Schultz, Schultz *et al.*, 1990], defined as $\lambda_1^D \equiv \text{MaxSp}(\mathbf{D})$.

It was found to be a good discriminant in a series of compounds of increasing size [Gutman and Medeleanu, 1998].

Balaban proposed [Balaban, Ciubotariu *et al.*, 1991] the unique negative eigenvalue of the distance matrix $\text{MinSp}(\mathbf{D})$ as a molecular descriptor, together with two of its transformations; these descriptors were called **VAD indices**:

$$\text{VAD1} = \text{MinSp}(\mathbf{D}) \quad \text{VAD2} = \frac{\text{VAD1}}{A} \quad \text{VAD3} = \frac{A}{10} \cdot \log(\text{VAD1})$$

The coefficients of the eigenvector associated with the unique negative eigenvalue of the distance matrix were used as $\rightarrow$ local vertex invariants (LOVIs), able to provide discrimination among graph vertices; higher values correspond to vertices of lower degree, those farther from the center or from a vertex of high degree. Based on the sum of these LOVIs, the **VED indices** were proposed as molecular descriptors:

$$\text{VED1} = \sum_{i=1}^A \ell_{iA} \quad \text{VED2} = \frac{\text{VED1}}{A} \quad \text{VED3} = \frac{A}{10} \cdot \log(\text{VED1})$$

where A is the number of molecular graph vertices and ℓ_{iA} are the coefficients (i.e., loadings) of the eigenvector associated with the largest negative eigenvalue (i.e., the A th eigenvalue of the decreasing eigenvalue sequence).

The **VRD indices** were defined as Randić-like indices based on the coefficients ℓ_{iA} of the eigenvector associated with the largest negative eigenvalue as

$$\text{VRD1} = \sum_b (\ell_{iA} \cdot \ell_{jA})_b^{-1/2} \quad \text{VRD2} = \frac{\text{VRD1}}{A} \quad \text{VRD3} = \frac{A}{10} \cdot \log(\text{VRD1})$$

where the sum runs over all the edges in the molecular graph; ℓ_{iA} and ℓ_{jA} are the LOVIs of the two vertices incident to the considered edge.

$\rightarrow$ *Schultz-type indices* were also proposed based on the eigenvector associated with the lowest (largest negative) eigenvalue of the adjacency matrix.

• leading eigenvalue of $(\mathbf{A} + \mathbf{D})$ (λ_1^{AD})

It is the largest eigenvalue of the $\rightarrow$ adjacency-plus-distance matrix obtained as the sum of the adjacency matrix **A** and the topological distance matrix **D** representing a molecular graph [Schultz, Schultz *et al.*, 1990]: $\lambda_1^{AD} = \text{MaxSp}(\mathbf{A} + \mathbf{D})$. Its logarithm was used to model physico-chemical properties [Cash, 1995c].

- **$\lambda\lambda_1$ branching index ($\lambda\lambda_1$)**

One of the $\rightarrow$ *double invariants*, it is the largest eigenvalue of the $\rightarrow$ *path matrix* that is among the $\rightarrow$ *graphical matrices* [Randić, 1997b, 1998b]. The path matrix contains the paths between each pair of vertices of the molecular graph and each path is characterized by the $\rightarrow$ *Lovasz–Pelikan index* λ_1^{LP} calculated from the adjacency matrix representing the path itself. Therefore, it is a generalization of the Lovasz–Pelikan index to characterize molecular branching of acyclic molecules; unlike λ_1^{LP} , it assumes the highest value for chain graphs and the lowest for the most branched graphs.

- **characteristic root index (CRI)**

This is the sum of the positive eigenvalues λ^+ of the $\rightarrow$ *path- χ matrix* χ_p , based on the path connectivities calculated by the $\rightarrow$ *valence vertex degree* δ^v of the atoms in the path [Saçan and Inel, 1993, 1995; Saçan and Balcioglu, 1996]:

$$CRI = SpSum^+(\chi_p) = \sum_{i=1}^{n^+} \lambda_i^+$$

CRI descriptor encodes information about all connectivities in the H-depleted molecular graph and is sensitive to the presence of heteroatoms in the molecule.

- **PPP eigenvalues**

These are the eigenvalues of the Hamiltonian matrix obtained by the *Pariser–Parr–Pople* (PPP) *method*.

A molecular descriptor λ^{PPP} is defined as [Balasubramanian, 1991, 1994]

$$\lambda^{\text{PPP}} = \sum_{i=1}^A \left| \frac{\lambda_i^{\text{PPP}} - \alpha - \gamma/2}{\beta} \right|$$

where the summation runs over all the eigenvalues; α and β are the Hückel parameters and γ the $\gamma_{\mu\mu}$ PPP integral [Pariser and Parr, 1953a, 1953b; Pople, 1953; Randić, 1991a]. The set of PPP eigenvalues was also proposed to measure the similarity between the pairs of molecules.

- **Wiener matrix eigenvalues (λ^{W})**

They are the eigenvalues obtained from the $\rightarrow$ *Wiener matrix* $\mathbf{W}$. In particular, by analogy with the $\rightarrow$ *Lovasz–Pelikan index*, the **Wiener matrix leading eigenvalue** was used as an alternative descriptor for the molecular branching [Randić, Guo *et al.*, 1994]: $\lambda_1^{\text{W}} = \text{MaxSp}(\mathbf{W})$.

- **Extended Adjacency matrix indices ($\equiv$ EA indices)**

They are two molecular descriptors calculated from $\rightarrow$ *extended adjacency matrices* $\mathbf{EA}$ [Yang, Xu *et al.*, 1994]. The first is the sum of the absolute eigenvalues of the matrix $\mathbf{EA}$, called $EA\Sigma$ index:

$$EA\Sigma \equiv SpSum(\mathbf{EA}) = \sum_{i=1}^A |\lambda_i^{\text{EA}}|$$

The second molecular descriptor is the maximum absolute eigenvalue of the matrix **EA**, called **EAmix index**:

$$EAmix \equiv MaxSpA(\mathbf{EA}) = \max_i |\lambda_i^{EA}|$$

These descriptors account for heteroatoms and multiple bonds, possess high discriminating power, and correlate well with a number of physico-chemical properties and the biological activities of organic compounds.

- **folding degree index (ϕ)**

This is the largest eigenvalue λ_1^{DD} obtained by the diagonalization of the $\rightarrow$ *geometric distance/topological distance quotient matrix* **G/D**, then normalized dividing it by the number of atoms *A* [Randić, Kleiner *et al.*, 1994; Randić and Krilov, 1999]:

$$\phi = \frac{\lambda_1^{DD}}{A} \quad 0 < \phi < 1$$

This quantity tends to one for linear molecules (of infinite length) and decreases in correspondence with the folding of the molecule. For example, ϕ values for transoid molecules are always greater than the values for the corresponding cisoid molecules. Thus, ϕ can be thought of as a measure of the folding degree of the molecule because it indicates the degree of departure of a molecule from strict linearity. This index allows a quantitative measure of similarity between chains of the same length but with different geometries; it is sensitive to conformational changes and can also be considered among the $\rightarrow$ *shape descriptors*.

The **folding profile** of a molecule was proposed as

$$\{^1\phi, ^2\phi, ^3\phi, \dots, ^k\phi, \dots\}$$

where $^k\phi$ is the normalized first eigenvalue of the k th order distance/distance quotient matrix whose elements are derived raising to the k th power the elements of the matrix **G/D**. Obviously, $^1\phi$ is the folding degree index. These vectorial descriptors were used to study the folding of $\rightarrow$ *peptide sequences* [Randić, 1997d].

The folding degree index is a measure of the conformational variability of the molecule, that is, the capability of a flexible molecule (often macromolecules, proteins) to assume conformations close to itself. Other descriptors related to the folding degree are the $\rightarrow$ *average span* and $\rightarrow$ *characteristic ratio*.

- **A_{xi} eigenvalue indices**

They are three molecular descriptors defined as:

$$A_{x_1} = MaxSp(\mathbf{M}_1)/2 \quad A_{x_2} = MaxSp(\mathbf{M}_2)/2 \quad A_{x_3} = MaxSp(\mathbf{M}_3)/2$$

where *MaxSp* refers to the largest eigenvalue of three symmetrized augmented path sparse matrices ($^+\mathbf{P}$) [Yao, Xu *et al.*, 1993a, 1993b; Xu, Yao *et al.*, 1995; Li, Xu *et al.*, 1996; Jiang, Li *et al.*, 2003]:

$$\mathbf{M}_1 = ^1\mathbf{P} \cdot ^1\mathbf{P}^T \quad \mathbf{M}_2 = ^2\mathbf{P} \cdot ^2\mathbf{P}^T \quad \mathbf{M}_3 = ^3\mathbf{P} \cdot ^3\mathbf{P}^T$$

where the superscript “T” indicates the transpose matrix.

The path matrices ${}^1\mathbf{P}$, ${}^2\mathbf{P}$, and ${}^3\mathbf{P}$ of dimension $A \times A$ are defined as the following:

$$[{}^1\mathbf{P}]_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

$$[{}^2\mathbf{P}]_{ij} = \begin{cases} 2 & \text{if there is a 2nd order path between } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

$$[{}^3\mathbf{P}]_{ij} = \begin{cases} 3 & \text{if there is a 3rd order path between } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

It can be observed that ${}^1\mathbf{P}$ is the $\rightarrow$ *adjacency matrix* $\mathbf{A}$ and that for acyclic graphs each of the above defined path matrices is coincident with the corresponding sparse matrix of the distance matrix.

The augmented matrices ${}^{1+}\mathbf{P}$, ${}^{2+}\mathbf{P}$, and ${}^{3+}\mathbf{P}$ are obtained by adding two columns to each matrix ${}^1\mathbf{P}$, ${}^2\mathbf{P}$, and ${}^3\mathbf{P}$: in the first column, there is the addition of the square roots of $\rightarrow$ *vertex degrees* δ and in the second column, the square roots of the van der Waals radii of the atoms.

To account for chirality, the same approach was used to calculate $\rightarrow$ *chirality descriptors*, called **chiral $\mathbf{A}_{xi}$ indices**, denoted as eA_{x1} , eA_{x2} , and eA_{x3} [Xu, Zhang *et al.*, 2006].

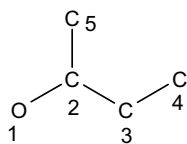
For chiral compounds, in the first column of the three augmented matrices ${}^{1+}\mathbf{P}$, ${}^{2+}\mathbf{P}$, and ${}^{3+}\mathbf{P}$, the elements corresponding to the chiral atoms in *S*-configuration are substituted with negative values, whereas the values are not changed for chiral atoms in *R*-configuration. Moreover, the diagonal elements in matrices ${}^1\mathbf{P}$, ${}^2\mathbf{P}$, and ${}^3\mathbf{P}$ are replaced by the electronegativities for the corresponding chiral atoms. Then, the chirality eigenvalues are defined as:

$$eA_{x1} = \text{MaxSp}(\mathbf{M}_1^*)/2 - A_{x1} \quad eA_{x2} = \text{MaxSp}(\mathbf{M}_2^*)/2 - A_{x2} \quad eA_{x3} = \text{MaxSp}(\mathbf{M}_3^*)/2 - A_{x3}$$

In the example presented below, the unique element that should be changed to calculate the chiral indices is the element (2,1) in all the three matrices, that is, ± 1.73 .

Example S6

The augmented matrices ${}^1\mathbf{P}$, ${}^2\mathbf{P}$, and ${}^3\mathbf{P}$ for 2-butanol in *R*-configuration. Van der Waals radii used in the example are 1.80 for carbon and 1.40 for oxygen atoms, respectively.



${}^{1+}\mathbf{P} =$		Atom	<i>a</i>	<i>b</i>	1	2	3	4	5
		1	1	1.18	0	1	0	0	0
		2	1.73	1.34	1	2.25	1	0	1
		3	1.41	1.34	0	1	0	1	0
		4	1	1.34	0	0	1	0	0
		5	1	1.34	0	1	0	0	0

${}^{2+}\mathbf{P} =$		Atom	<i>a</i>	<i>b</i>	1	2	3	4	5
		1	1	1.18	0	0	2	0	2
		2	1.73	1.34	0	2.25	0	2	0
		3	1.41	1.34	2	0	0	0	2
		4	1	1.34	0	2	0	0	0
		5	1	1.34	2	0	2	0	0

${}^{3+}\mathbf{P} =$		Atom	<i>a</i>	<i>b</i>	1	2	3	4	5
		1	1	1.18	0	0	0	3	0
		2	1.73	1.34	0	2.25	0	0	0
		3	1.41	1.34	0	0	0	0	0
		4	1	1.34	3	0	0	0	3
		5	1	1.34	0	0	0	3	0

ND indices are molecular descriptors calculated from augmented matrices ${}^1\mathbf{P}$, ${}^2\mathbf{P}$, and ${}^3\mathbf{P}$ where the first two columns contain (1) the square root of the bond vertex degree δ^b and (2) the $\rightarrow$ equilibrium electronegativity of atoms [Nie, Dai *et al.*, 2005]. ND indices are the leading eigenvalues of the symmetrized augmented path matrices $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$:

$$ND_1 = \text{MaxSp}(\mathbf{M}_1) \quad ND_2 = \text{MaxSp}(\mathbf{M}_2) \quad ND_3 = \text{MaxSp}(\mathbf{M}_3)$$

AEI indices are atomic descriptors calculated from augmented matrices ${}^1+\mathbf{P}$ and ${}^2+\mathbf{P}$ derived from a topological graph representing the structure of valence electrons of atoms, where vertices are the valence electrons and edges indicate interactions between valence electrons [Li, Dai *et al.*, 2005].

Then, four columns are added to the matrices ${}^1\mathbf{P}$ and ${}^2\mathbf{P}$ derived from the valence electron graph, which are defined as

$$\begin{aligned} [{}^1+\mathbf{P}]_{i1} &= [{}^2+\mathbf{P}]_{i1} = L^{-1/2} & [{}^1+\mathbf{P}]_{i2} &= [{}^2+\mathbf{P}]_{i2} = (l+1)^{-1/2} \\ [{}^1+\mathbf{P}]_{i3} &= [{}^2+\mathbf{P}]_{i3} = \pm |m_s|^{-1/2} & [{}^1+\mathbf{P}]_{i4} &= [{}^2+\mathbf{P}]_{i4} = E^{-1/2} \end{aligned}$$

where L is the principal quantum-number, l the angular quantum-number, m_s the spin quantum-number, and E the valence electron energy.

Then, AEI indices are calculated as

$$AEI_1 = \ln \sum_{i=1}^{Z^v} \lambda_i(\mathbf{M}_1) \quad AEI_2 = \ln \sum_{i=1}^{Z^v} \lambda_i(\mathbf{M}_2)$$

where Z^v is the number of valence electrons and the summation is on the eigenvalues of the symmetrized augmented path matrices $\mathbf{M}_1$ and $\mathbf{M}_2$. These atomic descriptors give structural information on size of atoms, shape of orbitals, spin, and interaction between the valence electrons.

• Nt index

A molecular descriptor calculated as [Zhou, Nie *et al.*, 2007]

$$Nt = \log A \cdot \log[\text{MaxSp}({}^a\mathbf{D}(r))]$$

where A is the number of atoms, $\text{MaxSp}({}^a\mathbf{D}(r))$ the leading eigenvalue of an augmented $\rightarrow$ bond length-weighted distance matrix, whose off-diagonal elements are the sum of the relative bond lengths to the carbon atom (1.54 Å) along the shortest path between the two vertices and the diagonal elements are the $\rightarrow$ equilibrium electronegativities of the atoms.

• Topological Electronegativity Index (TEI)

This is a spectral topochemical index proposed to describe alkyl groups and derived from an $\rightarrow$ augmented adjacency matrix ${}^a\mathbf{A}(\chi^{\text{PA}})$ weighted by the $\rightarrow$ Pauling electronegativity χ^{PA} as [Cao and Luo, 2007]

$$[{}^a\mathbf{A}(\chi^{\text{PA}})]_{ij} = \begin{cases} \chi_i^{\text{PA}} & \text{if } i=j \\ 1 & \text{if } (i,j) \in E(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases}$$

where $E(\mathcal{G})$ is the set of edges of the graph $\mathcal{G}$.

TEI is calculated as the geometric mean of the eigenvalues of the ${}^aA(\chi^{PA})$ matrix as

$$TEI = \left(\prod_{i=1}^n \lambda_i \right)^{1/n}$$

where n is the number of atoms of the alkyl group, which corresponds to the number of eigenvalues. It must be noted that for a single atom, TEI is equal to the Pauling electronegativity of that atom.

Example S7

Calculation of the TEI index for the methyl group $-CH_3$.

$${}^aA(\chi^{PA}) = \begin{vmatrix} \chi_C^{PA} & 1 & 1 & 1 \\ 1 & \chi_H^{PA} & 0 & 0 \\ 1 & 0 & \chi_H^{PA} & 0 \\ 1 & 0 & 0 & \chi_H^{PA} \end{vmatrix} = \begin{vmatrix} 2.55 & 1 & 1 & 1 \\ 1 & 2.20 & 0 & 0 \\ 1 & 0 & 2.20 & 0 \\ 1 & 0 & 0 & 2.20 \end{vmatrix}$$

Eigenvalues : 4.1159, 2.2000, 2.2000, 0.6341

$$TEI = [4.1159 \times 2.2000 \times 2.2000 \times 0.6341]^{1/4} = 1.8853$$

[Herndon and Ellzey Jr, 1975; McClelland, 1982; Mohar, 1991a; Hall, 1993; Cvetković and Fowler, 1999; Dias, 1999; Randić, 2000a; Rücker *et al.*, 2002; Rücker, Rücker *et al.*, 2002; Gutman, Indulal *et al.*, 2008]

- **spectral moments** → characteristic polynomial-based descriptors
- **spectral moments of iterated line graph sequence** → iterated line graph sequence
- **spectral moments of the adjacency matrix** → self-returning walk counts
- **spectral moments of the bond distance-weighted edge adjacency matrix** → edge adjacency matrix
- **spectral moments of the edge adjacency matrix** → edge adjacency matrix
- **spectral radius** $\equiv$ leading eigenvalue → spectral indices
- **spectral weighted invariant molecular descriptors** → *SWM signals*
- **spectral weighted molecular signals** $\equiv$ *SWM signals*

■ spectrum-like descriptors (S)

Proposed by the Zupan group [Novič and Zupan, 1996], these are → *3D-descriptors* for a spectrum-like representation of the molecular structure, defined by its → *molecular geometry*. They essentially describe the local geometry of molecules, the structure representation being obtained by the projection of the A atoms of the molecule to $2N$ points on two mutually perpendicular circles of a sphere with arbitrary radius.

To calculate spectrum-like descriptors, the molecules are aligned and located in a common center within the sphere (Figure S4). Each circle is spanned with a fixed angle ϕ that defines the resolution of the structure representation, that is, $N = 360/\phi$ values are considered. For example, by fixing ϕ value at 36° , 18° , or 10° , a number of descriptors equal to 10, 20, and 36, respectively, are obtained for each circle.

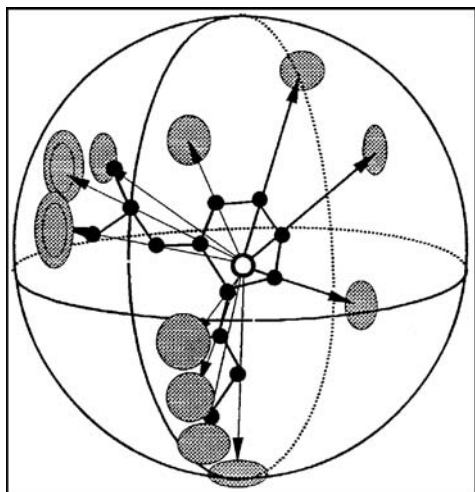


Figure S4 Spectrum-like projections.

At each j th projection point, the a th atom contribution to the j th descriptor is obtained by a Lorentzian function L (but can be any “bell-shaped” function):

$$L_{ja} \equiv L_j(\rho_a, \phi_a, \phi_j, \sigma_a) = \frac{\rho_a}{(\phi_j - \phi_a)^2 + \sigma_a^2}$$

where ρ_a is the distance of the a th atom from the center of the projection plane, ϕ_a the angle of the atom, ϕ_j the j th spanned angle. The parameter σ_a can be used as an atom-type dependent parameter. If only atom positions are considered, σ_a is set to 1, otherwise the Mulliken charge of the atom can be used. The suggested σ values are 0.2, 1.0, 1.2, and 1.4 for hydrogen, carbon, nitrogen, and oxygen atoms, respectively.

As can be observed, the function takes the maximum value when the a th atom is exactly along the spanned direction angle ϕ_j .

The $2N$ molecular descriptors S_j are obtained by summing, over all the A atoms, the Lorentzian function values of the two orthogonal circles (N for each circle):

$$S_j = \sum_{a=1}^A L_j(\rho_a, \phi_a, \phi_j, \sigma_a) = \sum_{a=1}^A \frac{\rho_a}{(\phi_j - \phi_a)^2 + \sigma_a^2} \quad j = 1, \dots, 2N$$

A modification of the Lorentzian function was proposed by Forina and coworkers [Forina, Boggia *et al.*, 1997] as

$$L_{ja} \equiv L_j(\rho_a, \phi_a, \phi_j, h_a, k) = \frac{\rho_a \cdot h_a}{k \cdot (\phi_j - \phi_a)^2 + 1}$$

where k is a constant, whose value controls the width of the function, that is, the function has a width, measured in degrees on the projection circle, which increases with decreasing k (the optimal k value is chosen by cross-validation between 1–10). The atom-type dependent parameter h_a substitutes the original σ_a and controls the height of the contribution of each atom to the Lorentzian signal.

Therefore, the **modified spectrum-like descriptors** are calculated as

$$S_j = \sum_{a=1}^A L_j(\rho_a, \phi_a, \phi_j, h_a, k) = \sum_{a=1}^A \frac{\rho_a \cdot h_a}{k \cdot (\phi_j - \phi_a)^2 + 1} \quad j = 1, \dots, 2N$$

Spectrum-like descriptors give information about molecular size and depend on molecule alignment, that is, on the two perpendicular circles onto which the atom projections fall. From $\rightarrow$ *variable selection*, partial $\rightarrow$ *reversible decoding* can also be performed.

📖 [Vračko, 1997; Zupan and Novič, 1997; Baumann, 1999; Zupan, Vračko *et al.*, 2000]

- **spectrum of a graph** $\rightarrow$ characteristic polynomial-based descriptors
- **spectrum of a matrix** $\rightarrow$ algebraic operators ($\odot$ characteristic polynomial)
- **spherosity index** $\rightarrow$ shape descriptors

■ spirality

Spirality (μ) is calculated as the number of “arcs,” such as in benzo(c)phenanthrene (Figure S5), present in the molecule [Johnson and Yalkowsky, 2005].

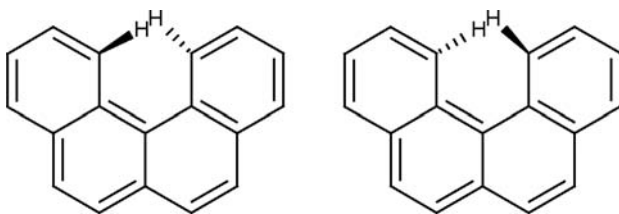


Figure S5 The two configuration of benzo(c)phenanthrene.

Each circular area causes deviation from planarity where the proximity of the hydrogens (or other substituents) results in marked steric interaction and, consequently, a somewhat spiraled molecule. For example, the spirality value of benzo(c)phenanthrene is 2 because each of such area has two possible absolute configurations.

Spirality, like flexibility and $\rightarrow$ *geometrical eccentricity*, increases the entropy of melting.

- **S_p statistics** $\rightarrow$ regression parameters
- **SQRT index** $\rightarrow$ molecular transforms
- **square root Hamming distance** $\rightarrow$ similarity/diversity ($\odot$ Table S10)
- **square root molecular weight** $\rightarrow$ physico-chemical descriptors ($\odot$ molecular weight)
- **square root Tanimoto distance** $\rightarrow$ similarity/diversity ($\odot$ Table S10)
- **SSAA index** $\rightarrow$ charged partial surface area descriptors
- **SSAH index** $\rightarrow$ charged partial surface area descriptors
- **SSIA descriptors** $\rightarrow$ biodescriptors ($\odot$ amino acid descriptors)
- **SSKey-type descriptors** $\rightarrow$ substructure descriptors ($\odot$ structural keys)
- **SSSR** $\equiv$ *Smallest Set of Smallest Rings* $\rightarrow$ ring descriptors
- **stability index** $\rightarrow$ characteristic polynomial-based descriptors
- **standard deviation** $\rightarrow$ statistical indices ($\odot$ indices of dispersion)

- **standard deviation error in calculation** $\equiv$ *root mean square error* $\rightarrow$ regression parameters
- **standard deviation error in prediction** $\equiv$ *root mean square error in prediction* $\rightarrow$ regression parameters
- **standard error in calculation** $\equiv$ *root mean square error* $\rightarrow$ regression parameters
- **standard error in prediction** $\equiv$ *root mean square error in prediction* $\rightarrow$ regression parameters
- **standard error of estimate** $\equiv$ *residual standard deviation* $\rightarrow$ regression parameters
- **standardized complementary distance matrix** $\rightarrow$ distance matrix
- **standardized information content** $\equiv$ *standardized Shannon's entropy* $\rightarrow$ information content
- **standardized information content on the leverage equality** $\rightarrow$ GETAWAY descriptors
- **standardized regression coefficients sum** $\rightarrow$ model complexity
- **standardized Shannon's entropy** $\rightarrow$ information content
- **standard reaction enthalpy** $\rightarrow$ physico-chemical properties ($\odot$ enthalpies)
- **star graph** $\rightarrow$ graph
- **start-end vectors** $\equiv$ *distance-counting descriptors*
- **static inductive effect** $\equiv$ *inductive effect* $\rightarrow$ electronic substituent constants
- **static polarizability** $\equiv$ *polarizability* $\rightarrow$ electric polarization descriptors
- **static reactivity indices** $\rightarrow$ reactivity indices

■ statistical indices

Statistical indices are fundamental numerical quantities measuring some statistical property of one or more variables. They are applied in any statistical analysis of data and hence in most of QSAR methods as well as in some algorithms for the calculation of $\rightarrow$ *molecular descriptors*. The most important univariate statistical indices are indices of central tendency and indices of dispersion, the former measuring the center of a distribution, the latter the dispersion of data in a distribution. Among the bivariate statistical indices, the correlation measures play a fundamental role in all the sciences. Other important statistical indices are the diversity indices, which are related to the $\rightarrow$ *information content* of a variable, the $\rightarrow$ *regression parameters*, used for regression model analysis, and the $\rightarrow$ *classification parameters*, used for classification model analysis.

A short list of definitions of the most common statistical indices is given below [Zar, 1984; Arnold, 1990; Johnson and Wichern, 1992]; in all the definitions, x represents the variable, i the i th sample, and n the total number of samples.

For weighted indices, w_i is the **statistical weight** assigned to the i th sample, which accounts for the importance of the sample in the statistics; statistical weight should usually satisfy two conditions:

$$w_i \geq 0 \quad \sum_{i=1}^n w_i = 1$$

Statistical weights for the molecule atoms ($\rightarrow$ *weighting schemes*) are commonly used in the calculation of several molecular descriptors.

Note that in most of the cases, statistical indices can be written both for data represented by n objects of a set $\{x_1, x_2, x_3, \dots, x_n\}$ or, equivalently, for data collected into K intervals (bins), each containing f_k values $\{f_1, f_2, f_3, \dots, f_K\}$.

• quantiles

A quantile is a value within the range of the variable, which divides the data into two groups such that the fraction of the observations specified by the quantile falls below and the complement

fraction falls above. For example, a quantile $Q(0.6)$ indicates a value of the variable for which 60% of the observations falls below and 40% falls above.

The most common quantiles are *quartiles* ($Q(0.25)$, $Q(0.50)$, $Q(0.75)$), *deciles* ($Q(0.1)$, $Q(0.2)$, ..., $Q(0.9)$), and *percentiles* ($Q(0.01)$, $Q(0.02)$, ..., $Q(0.98)$, $Q(0.99)$).

- **indices of central tendency**

Indices of central tendency (or *location*) provide a single value of a variable that is the most central and typical for representing the set of observations or a distribution.

Arithmetic mean. The most important measure of the center of a x variable, defined as

$$\bar{x} = \sum_{i=1}^n x_i$$

The generalization of the arithmetic mean is the *weighted mean*, defined as

$$\bar{x} = \frac{\sum_{i=1}^n w_i \cdot x_i}{\sum_{i=1}^n w_i}$$

where w_i is the weight associated with each i th object.

Geometric mean. A measure of the center, appropriate when all the data are positive, defined as the n th root of the product of the n data:

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^n x_i}$$

It is noteworthy that unlike the arithmetic mean, the geometric mean is equal to zero even though only one value of the variable is equal to zero.

Harmonic mean. A measure of the center, appropriate only when all the data are positive, defined as

$$\bar{x}_H = \frac{\sum_{i=1}^n (1/x_i)}{n}$$

Weighted mean of k th power. A general measure of center, defined as

$$M^{(k)} = \frac{\sum_{i=1}^n w_i \cdot x_i^k}{\sum_{i=1}^n w_i}$$

where k can take user-defined values. For each value of k , a different measure of center is obtained; for instance, for $k=1$ and $k=-1$, with unitary weights, the power mean is the arithmetic mean and the harmonic mean, respectively.

Median. A measure of center defined as the value of a variable, which divides the total frequency in two halves.

$$\bar{x}_m = \begin{cases} x_{(n+1)/2} & \text{odd samples} \\ \frac{x_{n/2} + x_{n/2+1}}{2} & \text{even samples} \end{cases}$$

that is, for an odd number of samples, the median is the value x of the sample located at position $(n + 1)/2$ in the ordered sample sequence, whereas for even samples, the median is usually taken as the arithmetic mean of the values of the two central samples.

Mode. It is usually defined as the most frequently occurring value in a variable. Distributions having one most frequent value are called unimodal, with two equally most frequent values bimodal, and with several equally most frequent values multimodal; however, a distribution having more than one most frequent value is assumed to have no mode.

Root mean square. Also called **quadratic mean** and abbreviated as **RMS** or *rms*, it is the mean of a quadratic variable, mainly useful when the variable values are both positive or negative:

$$rms = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

When the square variable is a difference between values predicted by a model or an estimator and the values actually observed, this quantity take the name of *root mean square deviation* (**RMSD** or *root mean square error*, **RMSE**) and is among the $\rightarrow$ *regression parameters*.

• indices of dispersion

Indices of dispersion (or *variation*) provide a single value of a variable, which describes the spread of a set of observations or the spread of a distribution around its center.

Variance. The most important measure of the x variable dispersion with respect to its center, defined as

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

where $\bar{x}$ is the arithmetic mean of the x variable and $n - 1$ are the degrees of freedom. The square root of the variance is called **standard deviation** (s).

Weighted variance. A weighted measure of variance, defined as

$$s^2 = \frac{\sum_{i=1}^n w_i \cdot (x_i - \bar{x})^2}{\sum_{i=1}^n w_i}$$

where w_i is the statistical weight assigned to the i th object.

Mean absolute deviation. A measure of the x variable dispersion with respect to its center, defined in terms of absolute deviations:

$$MAD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

This measure has the same units as the data.

Mean difference ($\equiv$ *simple mean difference*) A dispersion measure of the x variable based on the average of the absolute differences between all possible pairs of variable values; it is defined as

$$\Delta = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{n \cdot (n-1)}$$

The same quantity with repetitions is defined as

$$\Delta_R = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{n^2} = \frac{n-1}{n} \cdot \Delta$$

These two indices were proposed by Gini [Gini, 1909] to derive the $\rightarrow$ *Gini concentration ratio*.

Range. A simple dispersion measure of the x variable, defined as

$$range = x_{\max} - x_{\min}$$

where $x_{\max}$ and $x_{\min}$ are the largest and smallest value, respectively, assumed by the x variable. The range measure is strongly influenced by outliers.

Coefficient of variation. A dispersion measure that is independent of the magnitude of the observations defined as

$$CV = \frac{s}{\bar{x}}$$

where s is the standard deviation of the samples and $\bar{x}$ their arithmetic mean.

Interquartile range. A robust dispersion measure defined as the difference between the upper (Q_3) and lower (Q_1) quartiles:

$$IQR = Q_3 - Q_1$$

The **quartile deviation** is defined as the half interquartile range, that is, $IQR/2$.

• moment statistical functions

Moments about the mean. Statisticians refer to as k th moment about the mean (or k th **central moment**) the expression

$$m^k = \frac{\sum_{i=1}^n (x_i - \bar{x})^k}{n}$$

where $\bar{x}$ is the mean and n the number of samples. The second moment about the mean is the already defined $\rightarrow$ *variance*.

Pearson's first index. Based on the third-order central moment, it is the most general measure of asymmetry of the distribution of the x variable; it is defined as

$$\kappa_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3 \cdot n}$$

where $\bar{x}$ is the arithmetic mean of the x variable and s^3 the third power of its standard deviation.

Negative values derive from right-tailed distributions, whereas positive values from left-tailed distributions. Other measures of distribution asymmetry are *Pearson's second index*, *skewness*, and *Bonferroni index*.

Kurtosis. Based on the fourth-order central moment, it is a measure of the degree of bimodality of the distribution of the x variable, defined as

$$\kappa_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{s^4 \cdot n}$$

where $\bar{x}$ is the arithmetic mean of the x variable and s^4 the fourth power of its standard deviation. For a peak distribution, $\kappa = \infty$, whereas for a complete bimodal distribution, $\kappa = 1$. Uniform and normal distributions have $\kappa = 1.8$ and $\kappa = 3$, respectively.

• concentration indices

Whereas the indices of dispersion measure the variability of samples around a center, concentration indices measure the mutual variability of samples [Bonckaert and Egghe, 1991]. Concentration indices are used, for example, in $\rightarrow$ *cell-based methods* for the $\rightarrow$ *similarity/diversity* analysis of chemical libraries.

Some concentration indices are directly derived from dispersion indices, such as the **Yule characteristic K** and the **CON index**, both defined in terms of the $\rightarrow$ *coefficient of variation CV*:

$$K = \frac{CV^2}{n} \quad \text{CON} = \frac{CV}{\sqrt{n-1}}$$

where n is the number of objects.

Other specific concentration indices are listed below.

Gini concentration ratio. It is derived from the simple mean difference Δ and defined as [Gini, 1909]

$$G = \frac{\Delta}{2 \cdot \bar{x}} = \frac{1}{2 \cdot \bar{x}} \cdot \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{n \cdot (n-1)}$$

where n is the number of objects and $\bar{x}$ the arithmetic mean of the variable.

It is also equivalently defined as twice the area between the nonconcentration line and the Lorenz curve, defined in a diagram where on the horizontal axis, there are the points $\{0, 1, 2, \dots, n\}$ and on the vertical axis, values y_i^{NC} of the nonconcentration line and values y_i^{LC} of the Lorenz curve, calculated as

$$x_i = i \quad y_i^{NC} = \frac{i}{n} \quad y_i^{LC} = \sum_{j=1}^i a_i \quad i = 0, 1, 2, \dots, n$$

where a_i is defined as

$$a_i = \frac{x_i}{\sum_{i=1}^n x_i}$$

Therefore, the area can be calculated as

$$G = \frac{\sum_{i=1}^{n-1} (y_i^{NC} - y_i^{LC})}{\sum_{i=1}^{n-1} y_i^{NC}}$$

The Gini concentration index is closely related to the Pratt measure C (see below) by the following expression [Carpenter, 1979]:

$$G = \frac{n-1}{n} \cdot C$$

Chi-square statistics. It is the ratio of the square difference between the observed and theoretical expected values over the expected values:

$$\chi^2 = \sum_{k=1}^K \frac{(\hat{f}_k - f_k)^2}{\hat{f}_k}$$

where $\hat{f}_k$ is the expected number of objects in the k th bin, given by n/K , and f_k is the observed number of objects in the k th bin; K is the total number of bins. Derived from chi-square statistics are the **contingency coefficient**, denoted as cc , and **Cramer coefficient**, denoted as ϕ_1 ; they are, respectively, defined as

$$cc = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \phi_1 = \sqrt{\frac{\chi^2}{n}}$$

Theil's index. It is a concentration index simply defined as

$$Th = \frac{1}{n} \cdot \sum_{i=1}^n \left(\frac{x_i}{\bar{x}} \right) \cdot \ln \left(\frac{x_i}{\bar{x}} \right)$$

where n is the number of objects and $\bar{x}$ the arithmetic mean of the variable.

Pratt measure. It is a concentration measure defined as [Pratt, 1977; Egghe, 1987]

$$C = \frac{1}{2 \cdot n \cdot (n-1)} \cdot \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{\bar{x}}$$

where n is the number of objects and $\bar{x}$ the arithmetic mean of the variable.

• correlation measures

Correlation is one of the most important concepts in statistics [Pearson, 1920], being a quantity that indicates the strength and direction of a linear relationship between two random variables

(Figure S6). It is often assumed that correlation measures the departure of two variables from independence.

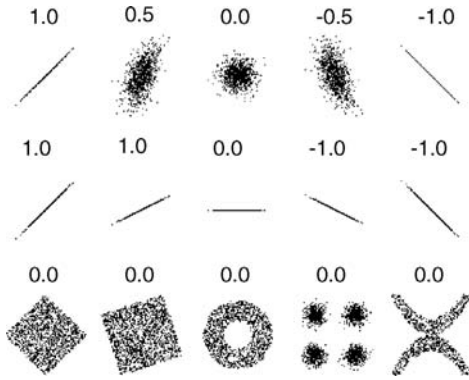


Figure S6 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom).

Several different correlation measures were proposed, depending on the nature of data.

Pearson's correlation coefficient. It is the most known bivariate correlation measure estimating the degree of association between the two variables j and k , as follows:

$$r_{jk} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) \cdot (x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \cdot \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}} = \frac{s_{jk}^2}{s_j \cdot s_k} \quad -1 \leq r_{jk} \leq +1$$

where $\bar{x}$ is the arithmetic mean of a variable, s_{jk}^2 the covariance between the two variables, and s the standard deviation.

Given p variables, the matrix whose elements are the variable pairwise correlations is called **correlation matrix** and denoted by $\mathbf{R}(p, p)$.

The **covariance** s_{jk}^2 between variables j and k is defined as

$$s_{jk}^2 = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) \cdot (x_{ik} - \bar{x}_k)}{n-1} \quad -\infty < s_{jk}^2 < +\infty$$

where $n-1$ are the degrees of freedom.

Given p variables, the matrix whose elements are the variable pairwise covariances is called **covariance matrix** and denoted by $\mathbf{S}(p, p)$. Weighted covariance matrices are the $\rightarrow$ *WHIM weighted covariance matrices*, which constitute the core for the calculation of the $\rightarrow$ *WHIM descriptors*.

The **Spearman rank correlation coefficient**, denoted as ρ_s , is a widely used correlation measure between two ranked variables defined as

$$\rho_{jk} = 1 - \frac{6 \cdot \sum_{i=1}^n (r_{ij} - r_{ik})^2}{n^3 - n} = \frac{\sum_{i=1}^n (r_{ij} - \bar{r}_j) \cdot (r_{ik} - \bar{r}_k)}{\sqrt{\sum_{i=1}^n (r_{ij} - \bar{r}_j)^2 \cdot \sum_{i=1}^n (r_{ik} - \bar{r}_k)^2}}$$

where r_{ij} and r_{ik} are the ranks of the i th object for the variables j and k . The Pearson's correlation coefficient applied to the ranks r can be used instead of the Spearman rank correlation, as shown in the right second term of the equation above.

The Spearman rank correlation coefficient is also used as a similarity measure for 3D distributed properties [Dean, 1990; Manaut, Sanz *et al.*, 1991]; it is defined as

$$R_{st} = 1 - \frac{6 \cdot \sum_{k=1}^N d_{st,k}^2}{N^3 - N}$$

where $d_{st,k}$ is a distance measure between the objects s and t at the k th point in the 3D space.

The **Kendall rank correlation coefficient**, denoted as τ , is another correlation measure between two ranked variables defined as

$$\tau = \frac{4 \cdot P}{n \cdot (n-1)} - 1 \quad -1 \leq \tau \leq +1$$

where n is the number of objects and P the number of concordant pairs calculated as the sum, over all the objects, of the number of objects ranked after the given object by both rankings.

Spearman ρ coefficient does assume that subsequent ranks indicate equidistant positions on the variable measured. Otherwise, when equidistance cannot be justified, correlation between the ordinal-level variables should be calculated by the Kendall coefficient.

Correlation coefficients are largely used to evaluate the $\rightarrow$ *similarity/diversity* between objects.

The pairwise correlation r_{jk} between the variables j and k can be viewed as a special case of the **multivariate K correlation index** (or *K correlation index*), which represents the measure of the correlation among p variables. This index is defined in terms of the distribution of the eigenvalues obtained by diagonalization of the correlation matrix $\mathbf{R}(p, p)$ [Todeschini, 1997; Todeschini, Consonni *et al.*, 1998].

Let $\{\lambda_1, \lambda_2, \dots, \lambda_p\}$ be the set of the $p \rightarrow$ *eigenvalues* of the correlation matrix (p, p) ; for this set the following properties hold:

$$\sum_{j=1}^p \lambda_j = p \quad \bar{\lambda} = 1$$

where $\bar{\lambda}$ is the average eigenvalue.

The K index is defined as the following:

$$K = \frac{\sum_{j=1}^p \left| \frac{\lambda_j}{p} - \frac{1}{p} \right|}{\frac{2 \cdot (p-1)}{p}} = \frac{\sum_{j=1}^p |\lambda_j - 1|}{2 \cdot (p-1)} \quad 0 \leq K \leq 1$$

where the denominator corresponds to the maximum value reached by the numerator and is used to scale K values between 0 and 1. It must also be observed that all of the zero eigenvalues are considered in the summation, each giving a contribution of $1/p$.

The K correlation index is a $\rightarrow$ *redundancy index*, taking the value of 1 when all of the variables are correlated and 0 when they are uncorrelated. From a geometrical point of view, $K=0$ corresponds to a spherical p -dimensional space and $K=1$ corresponds to a straight line in a p -dimensional space.

When p , the number of variables, is greater than n , the number of objects, the rank of the correlation matrix is n , this rank being less than or equal to $\min(n, p)$. In this case, at least $p - n$ eigenvalues are zero; however, the previous formula holds true, that is, the summation runs on the total number of variables p , but the K index can be viewed as the sum of the contributions:

$$K = \frac{\sum_{j=1}^n \left| \frac{\lambda_j}{p} - \frac{1}{p} \right| + (p-n) \frac{1}{p}}{\frac{2 \cdot (p-1)}{p}}$$

The summation runs from the first to the n th eigenvalue, while the second part of numerator takes into account the contribution to the correlation due to $(p - n)$ zero eigenvalues. Assuming the first part to be uncorrelated (i.e., the first n eigenvalues are p/n), an estimate of the minimum correlation within the data can be obtained from the following expression:

$$K_{IMB} = \min K = \left[n \cdot \left(\frac{1}{n} - \frac{1}{p} \right) + (p-n) \frac{1}{p} \right] \cdot \frac{p}{2 \cdot (p-1)} = \frac{p-n}{p-1}$$

This quantity is called **embedded correlation** and represents the minimum correlation of data when $p > n$. The embedded correlation can be viewed as the correlation due to the presence of N_R relationships among the p variables, that is, as

$$K_{EMB} = \frac{N_R}{p-1}$$

The multivariate correlation index calculated by diagonalization of the $\rightarrow$ *WHIM weighted covariance matrices* derived from $\rightarrow$ *molecular geometry* is also among the $\rightarrow$ *WHIM shape* descriptors.

- **statistical weight** $\rightarrow$ statistical indices
- **stepwise regression methods** $\rightarrow$ variable selection
- **stereodynamic representation** $\rightarrow$ molecular descriptors
- **stereoelectronic representation** $\rightarrow$ molecular descriptors
- **steric density parameter** $\rightarrow$ steric descriptors

■ steric descriptors

Steric effects are among the most relevant in modeling $\rightarrow$ *physico-chemical properties* and biological activities, thus playing a fundamental role in QSAR/QSPR modeling.

For this reason, a huge number of steric parameters was defined and used from the beginning to represent the steric properties of a molecule. Steric properties influence molecule energy, reaction and conformational paths, reaction rates and equilibria, binding affinity between the ligand and receptor, and other thermodynamic properties.

Steric descriptors account for both the size and shape of molecules and substituents, and are thus contemporarily related to $\rightarrow$ *size descriptors* and $\rightarrow$ *shape descriptors*. However, in several QSAR models, size descriptors are encountered as estimates of steric molecular/fragment properties.

The term **bulk descriptors** can be used to denote steric descriptors of the whole molecule, referring to the measure of the hindrance of the molecule when it is considered as a part of a system and dense assembly, each molecule being constrained by its neighbors to a limited region in space.

Steric descriptors were obtained from $\rightarrow$ *experimental measurements* of equilibrium and rate constants, $\rightarrow$ *computational chemistry*, geometrical and structural characteristics, and the $\rightarrow$ *topological representation* of a molecule.

The most common steric descriptors are $\rightarrow$ *molar refractivity*, $\rightarrow$ *surface areas*, and several $\rightarrow$ *volume descriptors* such as $\rightarrow$ *molecular volume*. Other steric descriptors are $\rightarrow$ *steric interaction fields*, $\rightarrow$ *MTD descriptors*, $\rightarrow$ *common overlap steric volume*, and several topological descriptors accounting for both size and $\rightarrow$ *molecular branching*.

Other popular molecular steric descriptors are listed below.

Steric substituent constants (or **steric substituent parameters**) are descriptors of substituent groups that measure the substituent steric effects on the reactivity centers of a molecule, based on differences in the rate and equilibrium constants of selected chemical reactions.

The most popular steric substituent descriptors are listed below.

- **Taft steric constant (E_S)**

The Taft steric constant E_S was proposed as a measure of steric effects that a substituent X exerts on the acid-catalyzed hydrolytic rate of esters of substituted acetic acids XCH_2COOR [Taft, 1952]. The basic assumption is that the effect of X on acid hydrolysis is purely steric, as the reaction constant ρ for acid hydrolysis of substituted esters is close to zero.

Therefore, the Taft steric constant is calculated as the average value from four series of kinetic data (hydrolysis of ethyl esters in 79% aqueous acetone in volume at 25 °C, esterification of carboxylic acids in methanol at 25 °C, esterification of carboxylic acids in ethanol at 25 °C, and hydrolysis of ethyl esters in 69% aqueous acetone in volume at 25 °C):

$$E_S = \log(k_X)_A - \log(k_H)_A = \log\left(\frac{k_X}{k_H}\right)_A$$

where k_X and k_H are the rate constants for the substituted and unsubstituted esters or acids, respectively, and the subscript A denotes hydrolysis in acid solution. The bulkier the substituent, the more negative the E_S constant values.

The **sum of steric substituent constants** was proposed as a molecular steric descriptor obtained by summing the steric substituent constants of all the substituents present in the molecule, such as

$$\Sigma = \sum_k (E_S)_k$$

where E_S are the Taft steric constants.

A rescaled set of the Taft steric constants was defined for the series of esters of substituted formic acids XCOOR as

$$E_S'' = E_S - 1.24$$

where 1.24 corresponds to the E_S value of the formic acid or ester [Motoc and Balaban, 1982].

To correct the hyperconjugation effect due to the α -hydrogens, a **corrected Taft steric constant** E_S^C (**Hancock steric constant**) was proposed:

$$E_S^C = E_S + 0.306 \cdot (h_\alpha - 3)$$

where h_α represents the number of α -hydrogens [Hancock, Meyers *et al.*, 1961].

To account for both C–C and C–H hyperconjugation effects, a different correction for Taft steric constant was proposed [Palm, 1972], obtaining the **Palm steric constant**:

$$E_S^0 = E_S + 0.33 \cdot (h_\alpha - 3) + 0.13 \cdot N_C$$

where h_α represents the number of α -hydrogens and N_C the number of α -carbon atoms in the substituent.

The **Dubois steric constant** E'_S is a revised Taft steric constant defined using the acid-catalyzed esterification of carboxylic acids (at 40 °C) as reference reaction [MacPhee, Panaye *et al.*, 1978a, 1978b].

The **Taft–Kutter–Hansch steric constants** (TKH E_S) constitute a combined set of parameters of the original Taft steric constant and those extended by Kutter and Hansch [Kutter and Hansch, 1969; Hansch, 1970; Sotomatsu and Fujita, 1989] by using a correlation with the average of the minimum and maximum van der Waals radii, as defined by the equation

$$\hat{E}_S = 3.484 - 1.839 \cdot \bar{R}^{vdw}$$

The values of the van der Waals radii are taken from Charton. By using this equation, steric constants were also calculated for ortho-substituted and nonalkyl groups.

The **Fujita steric constant** E_S^F [Fujita, Takayama *et al.*, 1973; Fujita and Iwamura, 1983] was defined to evaluate the global steric effect for branched alkyl substituents of the type $CR_1R_2R_3$ by the following correlation equation:

$$E_S^F = -2.104 + 3.429 \cdot E_S^C(R_1) + 1.978 \cdot E_S^C(R_2) + 0.649 \cdot E_S^C(R_3)$$

A **Model of the Frontier Steric Effect** is a theoretical approach proposed to estimate the Taft steric constant of substituents on the basis of the fundamental characteristics of the constituent atoms [Cherkasov and Jonsson, 1998]:

$$R_S = -30 \cdot \log \left(1 - \sum_{i=1}^n \frac{R_i^2}{4 \cdot r_i^2} \right) \quad \text{and} \quad R'_S = \sum_{i=1}^n \frac{R_i^2}{r_i^2}$$

where the summations run over all the atoms of the substituent, r_i is the distance of the i th atom of the substituent to the reaction center, and R_i the atomic radius of the atom. The second definition R'_S is an approximated solution of the first one for small values of the ratio. Moreover, R'_S reflects the specific surface of the reaction center screened by the atoms of the substituent [Cherkasov and Jonsson, 1999].

The **Topological Steric Effect Index (TSEI)** was proposed to describe the steric effects of substituents in terms of the relative specific volume of the reaction center screened by the atoms

of the molecule [Cao and Liu, 2004; Randić, Basak *et al.*, 2004]. The definition of TSEI is based on the idea that the steric effect of a substituent can be expressed by the relative specific volume V_{rc} of the reaction center, which is defined by analogy with the steric effect parameter R'_S by replacing the specific surface of the reaction center with the specific volume as

$$V_{rc} = \sum_{i=1}^n \left(\frac{\frac{4}{3} \cdot \pi \cdot R_i^3}{\frac{4}{3} \cdot \pi \cdot r_i^3} \right) = \sum_{i=1}^n \left(\frac{R_i^3}{r_i^3} \right)$$

where the summation is over the atoms in the substituent, R_i is the covalent radius of the i th atom, and r_i the distance between the i th atom and the reaction center, which is calculated as the sum of the bond lengths of the bonds along the path joining the i th atom to the reaction center.

Then, the topological steric effect index TSEI for an alkyl substituent, ignoring hydrogen atoms, is defined as

$$TSEI = \frac{V_{rc}}{k_t} = \sum_{i=1}^n \frac{1}{d_i^3} \quad k_t = \frac{0.772^3}{(0.772 + 0.772)^3} = 0.125$$

where d_i is the topological distance between the i th atom of the substituent and the reaction center; k_t is a constant defined in terms of the covalent radius R_C of the carbon atom (i.e., 0.772 Å) and the bond length r_{C-C} of the C–C bond (i.e., 1.544 Å).

The topological steric effect index $TSEI_X$ for a substituent containing any heteroatom X is analogously calculated as

$$TSEI_X = \frac{V_{rc}}{k_t} = \sum_{i=1}^n \frac{R_{i,rel}^3}{r_{i,rel}^3} \quad R_{i,rel} = \frac{R_i}{R_C} \quad r_{i,rel} = \frac{r_i}{r_{C-C}}$$

where $R_{i,rel}$ is the covalent radius of the i th atom normalized on the carbon atom and $r_{i,rel}$ is the normalized sum of the bond lengths from the i th atom to the reaction center.

📖 [Taft, 1953a, 1953c; Lambert, 1966; Hopkinson, 1969; Unger and Hansch, 1976; Murray, 1977; Fujita, 1978; Panaye, MacPhee *et al.*, 1980; Gallo, 1983]

• Charton steric constant (ν_X)

Charton steric constant ν_X (or **Charton characteristic volume**) is defined as

$$\nu_X = R_X^{vdw} - R_H^{vdw} = R_X^{vdw} - 1.20$$

where R_X^{vdw} is the $\rightarrow$ *van der Waals radius* of the substituent X and R_H^{vdw} the hydrogen atom radius [Charton, 1969, 1975, 1983]. For symmetric top substituents of the type CX_3 , Charton assumed that the axis of the group is the extension of the C– G_i bond, where G_i is the skeletal atom to which the group is bonded; then, he defined the minimum and maximum van der Waals radii perpendicular to the group axis as minimum and maximum widths of the group and the van der Waals radius parallel to the axis as the length of the group. Thus, depending on the van der Waals radius, different sets of ν_X constants can be calculated to completely characterize most substituents. However, to take into account that groups of atoms assume conformations, which minimize their repulsive effects, it is suggested to select the minimum van der Waals radius.

The average of the minimum and maximum van der Waals radii, as defined by Charton, was correlated to the Taft steric constant (see above).

Directly obtained from the van der Waals radii is also the **Bowden–Young steric constant R** [Bowden and Young, 1970]. It measures the steric hindrance of the substituent, calculated as the distance (in Å) from the aromatic carbon atom to which the substituent is bonded to the periphery of the van der Waals radius of the substituent, using known distances and van der Waals radii, and referred to that of the unsubstituted compound, that is, $R_H = 0$.

📖 [Charton, 1971, 1976, 1978b; Charton and Charton, 1978; Lall, 1982]

- **number of atoms in substituent specific positions**

The number of atoms in specific positions of a substituent are among the simplest substituent steric descriptors and can be used as steric “correction site” parameters. The **six position number N_6** (or **rule of six**) was suggested as the major factor in steric effects by Newman [Newman, 1950; Taft, 1956], summing carbon and hydrogen atoms in the sixth position from the carbonyl oxygen considered as atom one. This empirical rule is assumed for reactions involving addition to an unsaturated function: the greater the number of atoms in position six, the greater the steric effect.

The **Idoux steric constant $\Delta 6$** is defined as the change in the six number, that is, the difference of the six number of the X substituent of the ester XCOOX' minus the six position number of the same substituent X in the part X' of the ester [Idoux, Hwang *et al.*, 1973; Idoux, Scandrett *et al.*, 1977].

Distinguishing the contributions of carbon and hydrogen atoms, the **Bowden–Wooldridge steric constant E_S^{BW}** is defined by a correlation equation with the Taft steric constant:

$$E_S^{\text{BW}} = 0.119 - 0.347 \cdot N_{\text{C}6} - 0.075 \cdot N_{\text{H}6}$$

where $N_{\text{C}6}$ and $N_{\text{H}6}$ are the carbon and hydrogen atoms, respectively, in sixth position with respect to the carbonyl oxygen atom in the ester used to define E_S [Bowden and Wooldridge, 1973].

Still based on the number of substituent carbon atoms in different positions, Charton proposed a general correlation equation with its steric constant ν defined as

$$\nu = b_0 + b_1 \cdot n_\alpha + b_2 \cdot n_\beta + b_3 \cdot n_\gamma + b_4 \cdot n_\delta$$

where n_α , n_β , and so on represent the number of atoms in α -, β -, γ -, and δ -position, respectively [Charton, 1978b].

- **steric density parameter (SD_X)** ($\equiv$ Dash–Behera steric density parameter)

A substituent descriptor that combines molecular weight and van der Waals volume according to equation

$$SD_X = \left(\frac{\text{MW}}{V^{\text{vdw}}} \right)_X - \left(\frac{\text{MW}}{V^{\text{vdw}}} \right)_H = \left(\frac{\text{MW}}{V^{\text{vdw}}} \right)_X - 0.29$$

where MW is the $\rightarrow$ molecular weight and V^{vdw} the $\rightarrow$ van der Waals volume. X denotes the X-substituted compound and H the unsubstituted compound, that is, with the hydrogen atom in

the substitution site [Dash and Behera, 1980]. Namely, the meaning of this parameter is related to the density of the substituent. Values of *SD* for several substituents are derived from substituted organic acids.

- **steric vertex topological index (*SVTI*)**

A substituent steric descriptor for alkyl groups only defined in terms of the $\rightarrow$ topological distance *d* from a $\rightarrow$ *H-depleted molecular graph* [Ivanciuc and Balaban, 1996b]:

$$SVTI = \sum_{j=1}^{A_X} d_{ij} \quad \forall d_{ij} \leq 3$$

where A_X is the number of non-hydrogen atoms of the substituent *X* and d_{ij} the topological distance of any vertex v_j of the substituent group from v_i , which is the atom of the attachment site of a common reference skeleton. This index was proposed to approximate the steric effects of substituents so that only distances not exceeding 3 were considered as the steric effect of a residue related to the volume of that portion of its body closer to the link site.

- **substituent front strain (S_f) ($\equiv F$ strain)**

Substituent steric descriptor calculated from standard enthalpy of formation obtained by $\rightarrow$ computational chemistry using empirical force field methods. It is defined by the following relationship [Beckaus, 1978; Giese and Beckaus, 1978]:

$$S_f = \Delta H_f^0[\text{XC}(\text{CH}_3)_3] - \Delta H_f^0[\text{XCH}_3] + 8.87 \text{ [10}^4 \text{ J/mol]}$$

where $\Delta H_f^0[\text{XC}(\text{CH}_3)_3]$ and $\Delta H_f^0[\text{XCH}_3]$ are the standard enthalpies of formation of the *X*-substituted *t*-butyl and methyl derivatives, respectively, methyl and *t*-butyl groups being chosen for their high symmetry. The additive term 8.87 appears on normalization of $S_f(\text{CH}_3) = 0$.

F strain constants reflect only the steric repulsion of the attacking or leaving group of a reaction and contain no additional conformational effects.

The **steric energy difference** ΔSE , quite well related to the difference in enthalpy of formation $\Delta\Delta H_f$, has been also used as steric descriptor [DeTar and Tenpas, 1976; DeTar and Delahunty, 1983]; it was defined as the difference between the steric energy (calculated by molecular mechanic methods) of the *R*-substituted orthoacid $\text{RC}(\text{OH})_3$ and the steric energy of the corresponding *R*-substituted carboxylic acid RCOOH .

- **Joshi steric descriptor**

Quantum-chemical descriptor proposed to measure the steric effects of one or more substituents in a congeneric series of compounds [Joshi, Meister *et al.*, 1993, 1994]. The Joshi steric parameter JM1 (or log (JM1)) is defined as

$$\text{JM1} = \frac{\Delta E_R}{\Delta E_H} \quad \log(\text{JM1}) = \log(\Delta E_R) - \log(\Delta E_H)$$

where ΔE_H and ΔE_R are differences in conformational energy values of the unsubstituted and *R*-substituted compound, respectively:

$$\Delta E_H = E_H(\text{global}) - E_H(\text{strained}) \quad \Delta E_R = E_R(\text{global}) - E_R(\text{strained})$$

$E(\text{global})$ is the global energy minimum of the most favorable conformation associated with the least steric interactions, and $E(\text{strained})$ is the energy of a randomly chosen reference conformation, which is fixed for all compounds of the series.

For the unsubstituted compound, $\log(JM1) = 0$; if $\Delta E_H = 0$, the reference conformation is changed. The steric parameter JM1 values are calculated by computational chemistry using both AM1 method and PCILO approximation.

The steric component is achieved by subtracting the energy $E(\text{strained})$ of a constrained or sterically hindered molecule conformation from the energy values $E(\text{global})$ of the sterically most favorable conformation (ΔE_R). The steric influences of all substituents in the compound are determined by referencing ΔE_R to ΔE_H .

• Kier steric descriptor (Ξ)

A linear combination of $\rightarrow$ Kier shape descriptors obtained from an empirical relationship between $\rightarrow$ Taft steric constant E_S and group κ indices [Kier, 1986b] and defined by the following [Kier, 1987d]:

$$\Xi = 2 \cdot {}^1\kappa_\alpha - {}^0\kappa - {}^3\kappa_\alpha$$

where ${}^1\kappa_\alpha$ and ${}^3\kappa_\alpha$ are the alpha-modified first- and third-order Kier shape descriptors and ${}^0\kappa$ the $\rightarrow$ Kier symmetry descriptor. This index can be calculated for both the whole molecule and the substituent groups; it is somewhat related to the radii of the atoms involved in the substituent, and is a measure of the spatial influence of the group operating through the attached atom in the group.

• Austel branching index (S_X)

A geometric steric parameter for the X substituent based on the non-hydrogen atoms of the substituent and their $\rightarrow$ topological distance from the atom to which the substituent is linked [Austel, Kutter *et al.*, 1979]; it is calculated as

$$S_X = \sum_i b_i \cdot \sum_j n_{ij} \cdot k_j$$

where the first summation runs over all shells of the substituent around the link point on the parent structure, where the first shell ($i = 0$) includes the link point, the second shell ($i = 1$) all the substituent atoms at a topological distance equal to 1 from the link point, and so on; b_i are coefficients accounting for the branching contribution in the i th shell to the steric effect of the substituent. n_{ij} is the number of atoms of j th type bonded to atoms in the i th shell (atoms in the previous shell $i - 1$ are not counted); k_j is a factor that accounts for the size of the j th atom type.

Both the b and k values were estimated by regression analysis from the Taft steric constant E_S and the Charton steric constant in an iterative procedure using 96 substituents. The k values are equal to 1.0 for the second period elements, except for the fluorine atom ($k = 0.8$); $k = 1.2, 1.3$, and 1.7 for the third, fourth, and fifth period elements, respectively.

In calculating n_{ij} , a correction is proposed for cyclic structures: the number n_{ij} for the atom in the i th shell is reduced by the number of ring bonds leading from the $i - 1$ shell to the current i th shell.

Assuming that the relative distance contributions to the steric effect of a substituent are not significantly different, all the regression coefficients b were settled equal to one and a simplified

expression was proposed as

$$S_X = \sum_j N_j \cdot k_j$$

where the sum runs over all the different atom types and N_j is the count of the j th atom types.

For example, over this approximation, S_X values for $-\text{CH}_3$, $-\text{C}=\text{OF}$, $-\text{SF}_5$ substituent groups are 1 (1×1), 2.8 ($1 \times 1 + 1 \times 1 + 1 \times 0.8$), and 5.2 ($1 \times 1.2 + 5 \times 0.8$), respectively.

• Jenkins steric parameter (S_{AFF})

A steric parameter based on the proton affinity A_{H} and methyl cation affinity A_{CH_3} , calculated by computational chemistry methods, for reaction at the nitrogen atom series of compounds in which the nitrogen atom was unhindered, for example, pyridines in which the 2- and 6-positions are unsubstituted [Jenkins, Kelly *et al.*, 1994; Jenkins, Samuel *et al.*, 1995; Baxter, Jenkins *et al.*, 1996]. Using these affinity values, a reference regression model was found, which correlates the methyl cation affinity A_{CH_3} to the proton affinity A_{H} for unhindered compounds:

$$\hat{A}_{\text{CH}_3} = -446.78 + 0.971 \cdot A_{\text{H}}$$

For nitrogen compounds where the nucleophilic nitrogen atom is hindered, the methyl cation affinity values calculated are less than those estimated by the reference model. The difference between the two values is assumed as the steric parameter, that is,

$$S_{\text{AFF}} = A_{\text{CH}_3} - \hat{A}_{\text{CH}_3}$$

📖 [Dubois, MacPhee *et al.*, 1980; Verloop, 1985; Kim, 1992b]

- **steric energy difference** → steric descriptors (⊙ substituent front strain)
- **steric interaction fields** → molecular interaction fields
- **steric interactions in biological systems** → minimal topological difference

■ steric misfit

The difference between the steric hindrance of the → *leading compound* (assumed as → *reference compound*) and the considered molecule. It is a measure of the similarity where low values of steric misfit correspond to high similarity between the molecule and reference compound, that is, a favorable condition for the considered molecule.

Methods for evaluating the steric misfit are → *minimal topological difference* and → *molecular shape analysis*.

- **steric substituent constants** → steric descriptors
- **steric substituent parameters** ≡ *steric substituent constants* → steric descriptors
- **steric vertex topological index** → steric descriptors
- **Sterimol B parameters** → Sterimol parameters
- **Sterimol length parameter** → Sterimol parameters

■ Sterimol parameters (≡ Verloop Sterimol parameters; Verloop parameters)

A set of parameters proposed by Verloop [Verloop, Hoogenstraaten *et al.*, 1976; Verloop, 1987] to describe the size and shape of substituents in a congeneric series. These parameters were

evaluated by measuring the dimensions of substituents in a restricted number of directions by a computer program (STERIMOL), which simulates 3D model building of substituent groups, using the $\rightarrow$ *Corey–Pauling–Koltun volume* (CPK atomic models). For flexible substituents, minimum energy conformations are considered.

The substituent attachment atom G_1 on the parent structure (e.g., benzene) is placed at the origin of the Cartesian coordinates (x, y, z), assuming that the bond connecting the substituent to G_1 defines the X axis. The **Sterimol length parameter L** is defined as the maximum length along the X axis, that is, it is the x coordinate of the intersection point on the X axis of the tangential plane to the substituent, perpendicular to the X axis (Figure S7a). For example, for the substituent H of the benzene molecule, $L = 2.06$, obtained as sum of the C–H bond (1.06 Å) plus the van der Waals radius of H (1.00 Å).

The **Sterimol B parameters** were proposed to characterize the widths of the substituent along the directions perpendicular to the X axis. The width parameters B_1 , B_2 , B_3 , and B_4 are, in ascending order, taken as the distances to the X axis of tangential planes to the substituent, perpendicular to the Z and Y axes. B_1 and B_4 are defined as the smallest and largest width along the Z axis, whereas B_2 and B_3 are defined as the smallest and largest width along the Y axis (Figure S7b). In other words, these parameters describe the positions, relative to the origin and the axes, of the five side planes of the box embedding the substituent, made in such a way that the distance of one of its sides to the axis has the smallest possible value. In most cases, the B_4 value is almost equal to the maximum length L of the substituent.

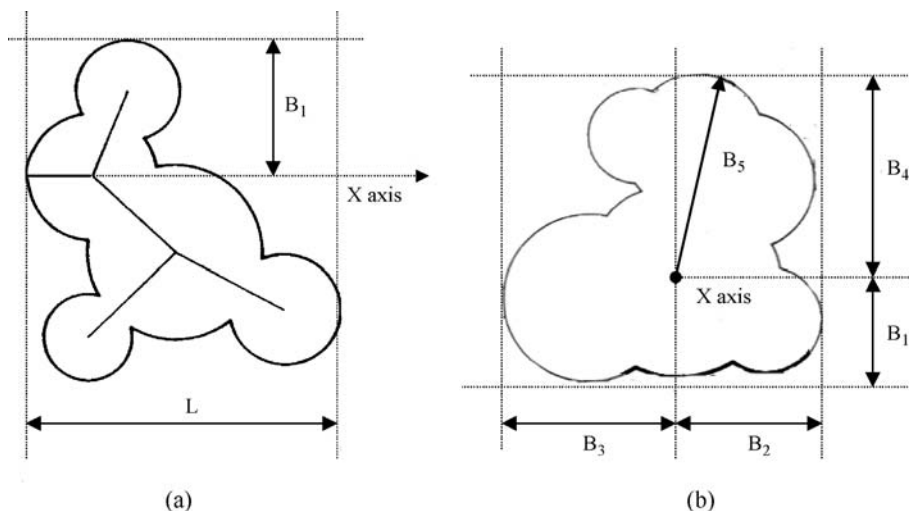


Figure S7 (a) Projection of a substituent along the X axis showing the parameters L and B_1 . (b) Projection of a substituent perpendicular to the X axis showing B parameters.

Later, a new Sterimol width parameter B_5 was introduced to replace the B_4 parameter, defined as the maximum width (i.e., the maximum distance from X axis) of the substituent in the Z–Y plane (perpendicular to the X axis).

Among the $\rightarrow$ *shape descriptors*, **Sterimol shape parameters** were proposed as the $\rightarrow$ *length-to-breadth ratio*, defined as L/B_1 and B_5/B_1 (or previously, B_4/B_1), giving information about the deviations of a substituent from a spherical shape.

Verloop *et al.* determined the Sterimol parameters for over 1000 substituent groups (Table S17).

📖 [Arnaud, Taillandier *et al.*, 1994; Kawashima, Yamada *et al.*, 1994; Draber, 1996; Singh, Gupta *et al.*, 1996]

- **Sterimol shape parameters** $\rightarrow$ Sterimol parameters
- **STIMS indices** $\rightarrow$ count descriptors
- **stochastic adjacency matrix of a general graph** $\rightarrow$ TOMOCOMD descriptors
- **stochastic edge adjacency matrix** $\rightarrow$ TOMOCOMD descriptors
- **stochastic matrices** $\rightarrow$ algebraic operators
- **Structure–Activity Landscape Index** $\equiv$ *SAL Index* $\rightarrow$ Structure/Response Correlations
- **Structure–Activity Relationship Index** $\equiv$ *SAR Index* $\rightarrow$ Structure/Response Correlations
- **structural alerts** $\rightarrow$ property filters ($\odot$ functional group filters)
- **structural code** $\equiv$ *vertex structural code* $\rightarrow$ self-returning walk counts
- **structural environment vector** $\rightarrow$ scoring functions ($\odot$ Klopman–Henderson cumulative substructure count)
- **structural graph** $\equiv$ *molecular graph*
- **structural information content** $\rightarrow$ indices of neighborhood symmetry
- **structural keys** $\rightarrow$ substructure descriptors
- **structure–activity relationships** $\rightarrow$ Structure/Response Correlations
- **structure–property relationships** $\rightarrow$ Structure/Response Correlations
- **structure–reactivity relationships** $\rightarrow$ Structure/Response Correlations

■ Structure/Response Correlations (SRC)

The term *Structure/Response Correlations* (SRC) is proposed in this book to collect under a unique framework all the approaches aimed at finding relationships between the molecular structure and measured (or calculated) molecular response.

The proposed term *structure/response correlations* is a further generalization of the proposal *Structure Property Correlation* (SPC) suggested by van de Waterbeemd [van de Waterbeemd, 1992]. The authors think that the term *property* has actually a too specific meaning and suggest the very general term *response* to encompass both physico-chemical properties and biological activities. *Response* is semantically related to the mathematical concept of γ *response* in modeling, that is, the dependent variable in any correlation equation that can be easily intended here as any experimental measured quantity, such as $\rightarrow$ *physico-chemical properties*, that is, **Structure–Property Relationships** (SPR), or biological activities, that is, **Structure–Activity Relationships** (SAR), or reactivity measures, that is, **Structure–Reactivity Relationships** (SRR).

In quantitative approaches to SRC studies, that is, **Quantitative Structure/Response Correlations** (QSRC), the aim is to obtain quantitative relationships like those represented by regression models and classification models. Thus, SRC studies involve both quantitative correlation studies and qualitative comparisons among $\rightarrow$ *molecular descriptors* and responses.

The class of the quantitative approaches to SRC studies includes all the well-known approaches called **Quantitative Structure–Activity Relationships** (QSAR), **Quantitative Struc-**

ture–Property Relationships (QSPR), Quantitative Structure–Reactivity Relationships (QSRR), Quantitative Shape–Activity Relationships (QShAR), the molecular shape being considered as a component of the molecular structure, Quantitative Structure–Chromatographic Relationships (QSCR), Quantitative Structure–Toxicity Relationships (QSTR), → *Quantitative Similarity–Activity Relationships* (QSiAR), Quantitative Structure–Enantioselective Retention Relationships (QSERR), and so on.

The proposed slash between the two terms “structure” and “response” denotes both “and” and “or,” thus accounting also for property–property relationships as well as → *similarity/diversity* correlations. Therefore, quantitative property–property relationships (QPPR), property–activity relationships (QPAR), activity–activity relationships (QAAR), and similarity/diversity correlations fall, in a broader sense, within SRC studies (Figure S8).

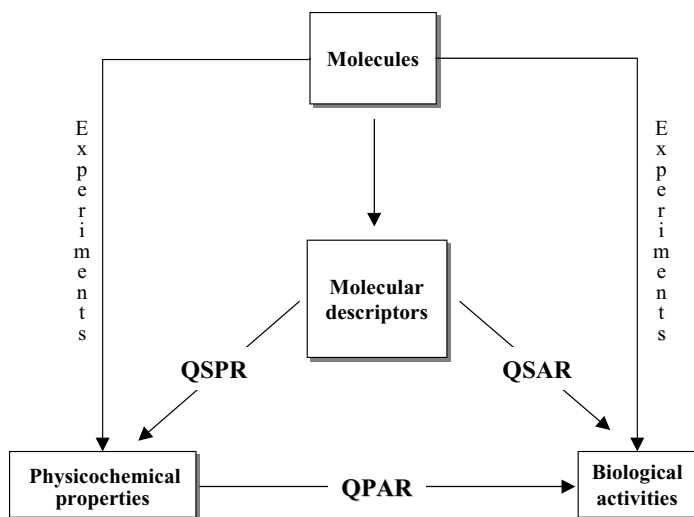


Figure S8 Scheme of the relationships of the molecular descriptors with molecules and structure/response correlations approaches.

Quantitative Information Analysis is the term proposed by Kier to denote structure/response correlations, where the word “analysis” is chosen to avoid any restriction to QSAR/QSPR models, but naturally includes similarity/diversity analysis as well as any explorative analysis or model that not refers only to relationships with the molecular structure.

The term **classical QSAR** is often used to denote → *Hansch analysis*, → *Free–Wilson analysis*, → *Linear Free Energy Relationships* (LFER), and → *Linear Solvation Energy Relationships* (LSER), that is, those SRC approaches developed between 1960 and 1980 that can be considered the beginning of the modern QSAR/QSPR methods.

QSAR approaches based on the → *topological representation* of the molecules are often called **2D-QSAR**. The term **3D-QSAR** is often, though improperly, used to denote only → *grid-based QSAR techniques* and *receptor mapping techniques* (see below), but the present relevance of other kinds of descriptors obtained from 3D → *molecular geometry* naturally leads to the enlarging this term to include all the SRC methods based on the → *geometrical representation* of the molecules.

Proposed by Basak and coworkers [Gute, Grunwald *et al.*, 1999; Basak, Grunwald *et al.*, 2000; Basak, Mills *et al.*, 2001], **Hierarchical-QSAR (HiQSAR)** is a QSAR approach that partitions the whole set of molecular descriptors into four different logical blocks (topostructural descriptors, topochemical descriptors, geometric descriptors, and quantum-chemical descriptors) and uses them sequentially in the formulation of QSAR models for predicting physical, biomedical, and toxicological properties. → *Variable selection* is accomplished by searching for different clusters of molecular descriptors based on their mutual correlations; from each cluster, the molecular descriptors most correlated with the cluster centroid are selected for modeling as well as any molecular descriptors that are poorly correlated with their cluster centroids. Applications of the HiQSAR approach are discussed in: [Basak and Mills, 2001a, 2001b; Basak, Mills *et al.*, 2002; Basak, Mills *et al.*, 2003b; Basak and Mills, 2005; Basak, Gute *et al.*, 2006].

With a different meaning, the term hierarchical QSAR was also used to denote the application of Partial Least Squares (PLS) and Principal Component Analysis (PCA) to different logical blocks of molecular descriptors to summarize descriptors of each block into a few latent variables or components, which were called “supervariables” [Eriksson, Johansson *et al.*, 2002].

Dynamic QSAR (also called **4D-QSAR**) denotes those SRC techniques that take conformation variability of the molecules into account [Mekenyan, Ivanov *et al.*, 1994; Dimitrov and Mekenyan, 1997].

→ *Binary QSAR analysis* is an approach to screening of chemical libraries aimed at identifying possible lead compounds on the basis of a probability distribution function for active and inactive compounds [Labute, 1999; Gao, Williams *et al.*, 1999].

In binary QSAR, the biological activity, expressed in a binary form (1 for active and 0 for inactive) is correlated with molecular descriptors of compounds, and a probability distribution for active and inactive compounds in a training set is estimated. The derived binary QSAR model can subsequently be used to predict the probability of new compounds to be active against a given biological target.

Due to the importance of the design of new active compounds in contemporary → *drug design*, several methods have been proposed to measure the binding affinity of a set of ligand compounds without a deep knowledge of the three-dimensional structure of the receptor site. A number of methods, called **receptor mapping techniques**, attempt to provide insight into the receptor active site and characterize receptor binding requirements to design a desirable ligand.

Usually in these methods, the most used molecular descriptors are → *interatomic distances* and different kinds of interaction energies. → *Distance Geometry* (DG) and → *minimal topological difference* (MTD) are the earliest examples of such an approach. Among them, other popular methods are *Active Analogue Approach* (AAP) [Marshall, Barry *et al.*, 1979], *LOCON*, and *LOGANA methods* [Franke and Streich, 1985a, 1985b; Franke, Hübel *et al.*, 1985; Streich and Franke, 1985], *Receptor Surface Model* (RSM) [Hahn and Rogers, 1995, 1998; Hahn, 1995, 1997], *Receptor Binding Site Model* (RBSM) [Höltje, Baranowski *et al.*, 1985; Höltje, Anzali *et al.*, 1993], *Hypothetical Active-Site Lattice* (HASL) [Doweyko, 1991; Doweyko and Mattes, 1992; Wiese, 1993], *Genetically Evolved Receptor Models* (GERM) [Walters and Hinds, 1994; Walters, 1998; Chen, Zhou *et al.*, 1998], and → *Probabilistic Receptor Potential* (PRP) [Labute, 2001].

Proteo-chemometrics approach was the name proposed for a QSAR approach based on the combined analysis of series of receptors and ligands, wherein description of ligands, proteins, and the so-called ligand–protein cross-terms are correlated with interaction activities [Lapinsch, Prusis *et al.*, 2003].

Reviews and general discussions about QSAR/QSPR strategies can be found in Refs. [Craig, 1984; Charton, 1996; Devillers, 1998; Charton and Charton, 1999, 2002; Devillers and Balaban, 1999; Devillers, 1999b; Gundertofte and Jørgensen, 2000; McKinney, Richard *et al.*, 2000; Diudea, 2001; Kubinyi, 2002; Cronin and Schultz, 2003; Schultz, Cronin *et al.*, 2003a, 2003b; Stanton, 2003; Clark, 2004; García-Domenech, Gálvez *et al.*, 2008].

The development of QSAR/QSPR models is a quite complex process, as outlined in Figure S9.

Important steps of this process are (a) selection of the set of molecules the modeling procedure is applied to, and the set of molecular descriptors that will define the model chemical space; (b) selection of the training set for the model estimation and the test set for model validation; (c) application of the validated model(s) to design new molecules with desirable properties and/or predict the response of interest for future molecules, paying attention to the → *applicability domain* of the model.

The **chemical space** is defined as the p -dimensional space constituted by a set of p molecular descriptors selected to represent the studied compounds; chemical space design is generally recognized as a crucial step for the successful application of QSPR/QSAR methods [Oprea, Zamora *et al.*, 2002; Dutta, Guha *et al.*, 2006; Eckert, Vogt *et al.*, 2006; Landon and Schaus, 2006].

Another relevant aspect in structure/response correlations is the ability to obtain information about molecular structure from QSAR/QSPR models. In particular, the term **reversible decoding** (or **inverse QSAR**) denotes any procedure capable to reconstruct the molecular structure or fragment starting from molecular descriptor values, that is, once molecular

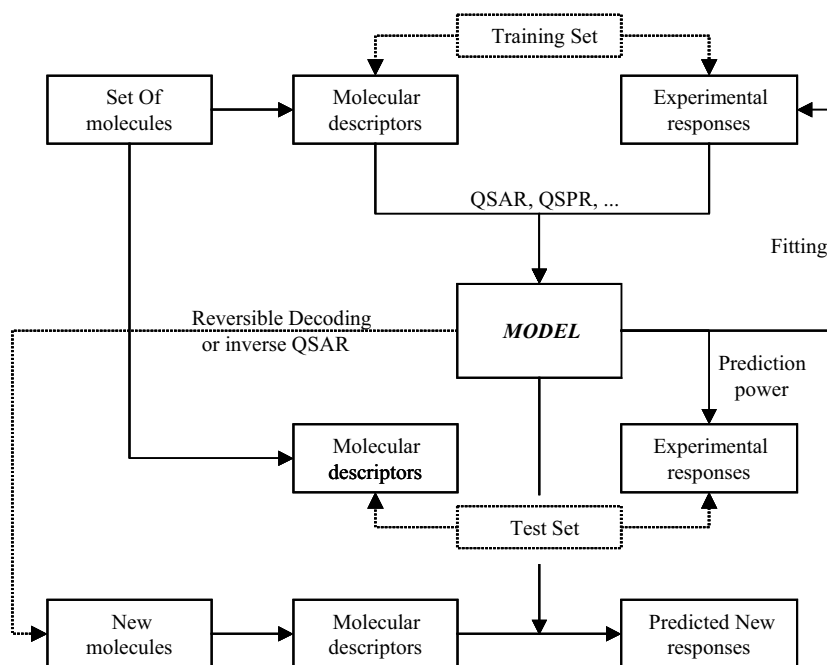


Figure S9 Role of a model in structure/response correlations.

descriptors from a structure representation are obtained, reversibility would lead to structures from molecular descriptors [Baskin, Gordeeva *et al.*, 1989; Gordeeva, Molchanova *et al.*, 1990; Zefirov, Palyulin *et al.*, 1991; Hall, Kier *et al.*, 1993; Hall, Dailey *et al.*, 1993; Kier, Hall *et al.*, 1993; Kier and Hall, 1993; Skvortsova, Baskin *et al.*, 1993; Zefirov, Palyulin *et al.*, 1995; Cho, Zheng *et al.*, 1998; Lukovits, 1998a; Brüggemann, Pudenz *et al.*, 2001; Gozalbes, Doucet *et al.*, 2002; Churchwell, Rintoul *et al.*, 2004; Weis, Faulon *et al.*, 2005; Brown, McKay *et al.*, 2006].

Reversible decoding is of great importance since once a SRC model is established, optimal values of the response can be chosen and values of the model molecular descriptors calculated by using the estimated SRC model. Then, the possible molecular structures corresponding to the optimized descriptor values can be designed (and synthesized). This last operation is troublesome task if the molecular descriptors of the model are not simple and easily interpretable.

Reversibility is a highly desired property of a descriptor, but is not strictly essential for structure–response studies; it is closely related to the uniqueness of the descriptor, that is, to its degree of degeneration.

Due to the large availability of different models predicting the same molecular property, such as those models provided by genetic algorithms, a particular QSAR strategy, called **consensus analysis**, was proposed [Charifson, Corkery *et al.*, 1999; Clark, Strizhev *et al.*, 2002; Sutherland and Weaver, 2003; MobyDigs – Talete s.r.l., 2003; van Rhee, 2003]. This QSAR approach consists in selecting not just one model, but more than one. Predictions are performed contemporarily using the average response obtained from all the selected models or, better, using the weighted average response, considering as the statistical weight the leverage h_{ik} of the i th molecule from each k th model, as

$$\bar{y}_i = \frac{\sum_{k=1}^M w_k (\hat{y}_{ik}/h_{ik})}{\sum_{k=1}^M 1/h_{ik}}$$

where M is the number of selected models, $\hat{y}_{ik}$ the response estimated for the i th object by the k th model, and h_{ik} the diagonal elements of the $\rightarrow$ *leverage matrix* [MobyDigs – Talete s.r.l., 2003]. The leverage is a measure of the “distance” of the molecule from the chemical space defined by the selected model, that is, small leverage corresponds to a molecule well represented in the model chemical space, whereas high leverage is obtained for a molecule on the boundary of the model space; thus, the response likely being extrapolated and less reliable. Moreover, the weight w_k can be a unitary weight or can take into account the quality (e.g., the Q^2 leave-one-out) of each model as

$$w_k = \frac{Q_k^2}{\sum_{k=1}^M Q_k^2} \quad \wedge \quad \sum_{k=1}^M w_k = 1$$

A consensus modeling can provide for each molecule the standard deviation of the responses predicted by the selected models, which is a measure of the convergence of all the selected models toward a unique response.

Applications of consensus analysis are discussed in Refs. [Asikainen, Ruuskanen *et al.*, 2004; Baurin, Mozziconacci *et al.*, 2004; Godden, Furr *et al.*, 2004; Gramatica, Pilutti *et al.*, 2004a;

Klon, Glick *et al.*, 2004; Votano, Parham *et al.*, 2004b; Dutta, Dutta *et al.*, 2007; Gramatica, Giani *et al.*, 2007; Hewitt, Cronin *et al.*, 2007; Zhang, Golbraikh *et al.*, 2007].

Another important and well-known aspect in structure/response correlation studies is the concept of *congenericity*. Congenericity is a fuzzy concept related to the structures of molecules in a data set. With respect to some molecular structural characteristics, chemical analogues can be considered congeneric if their structural differences are the interesting part of the study. Monosubstituted benzenes, polychlorobiphenyls, triazines, and polyaromatic hydrocarbons are all examples of the families of congeneric compounds.

The **congenericity principle** is the assumption that “similar compounds have similar activities” and is the basic requirement for several structure/response correlations. According to this principle, activity changes gently in the chemical space, that is, small changes in the structure of molecules lead to small changes in the activity.

Failure of the congenericity principle has been recognized as one of the major problems for producing reliable QSAR models, unless the presence of the so-called “activity cliffs” is accounted for. The *activity cliff* was defined as the ratio of the difference in activity of two compounds to their distance in a given chemical space [Maggiora, 2006]. Activity cliffs arise when very similar compounds, whose similarity is measured by the set of molecular descriptors used to define the chemical space, possess very different activities. The presence of activity cliffs leads to some important implications for QSAR modeling:

“First, purely linear models, even very local ones, in which neither the parameters nor the variables are nonlinear, are unlikely to satisfactorily account for activity landscapes with significant numbers of cliffs. Second, outliers in the data may not be due to statistical fluctuations or to measurement errors but rather may reflect the presence of activity cliffs. Thus, perfectly valid data points located in cliff regions may appear to be outliers. Third, the presence of activity cliffs requires the assay of additional compounds in the neighborhoods around these cliffs to ensure that activity landscapes are adequately represented in these rapidly varying regions and, thus, that QSAR models can faithfully represent the SAR data. Another crucial issue that arises here is the lack of invariance of chemical space to changes in the set of descriptors used to represent the molecular information in the model” [Maggiora, 2006].

A quantification of the concept of presence of activity cliffs was proposed in terms of the **SAL Index** (or **Structure–Activity Landscape Index**), which for a pair of compounds is defined as [Maggiora, 2006; Guha and Ven Drie, 2008]

$$SALI_{ij} = \frac{|A_i - A_j|}{1 - s_{ij}}$$

where A_i and A_j are the biological activities of two compounds i and j and s_{ij} their similarity in a given chemical space. Large values of this index are obtained when different activities are observed for a pair of very similar compounds, thus indicating a lack of information in describing the compound diversity by the selected set of descriptors.

Another index for evaluating presence of activity cliffs is the **SAR Index** (or **Structure–Activity Relationship Index**), defined as a function of two separately calculated scores that assess intraclass diversity and activity differences of similar compounds [Peltason and Bajorath, 2007]:

$$SARI = \frac{1}{2} \cdot [score_{\text{cont}} - (1 - score_{\text{disc}})]$$

where $score_{\text{cont}}$ and $score_{\text{disc}}$ are the *continuity* and *discontinuity score*, respectively.

The **continuity score** measures activity-weighted structural diversity within a class of active compounds to generate the continuity score; first, the following quantity is calculated as

$$q_{\text{cont}} = 1 - \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n w_{ij} \cdot s_{ij}}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n w_{ij}}$$

where n is the number of compounds in the data set and s_{ij} the similarities between pairs of compounds. The pairwise weights are defined as

$$w_{ij} = \frac{A_i \cdot A_j}{1 + |A_i - A_j|}$$

where A indicates the compound activity. High values of continuity scores reflect the presence of structurally diverse molecules having comparable activity.

The **discontinuity score** determines the average activity difference for pairs of similar compounds, which reveals the presence of activity cliffs as a major determinant of discontinuous SARs. To generate the discontinuity score, the following quantity is defined:

$$q_{\text{disc}} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n |A_i - A_j| \cdot s_{ij}}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n s_{ij}} \quad \forall (i, j) : s_{ij} > 0.6$$

where only pairwise similarities greater than 0.6 are taken into account.

Finally, the two quantities are transformed into the corresponding scores by standardization:

$$\text{score}_{\text{cont}} = \frac{q_{\text{cont}} - \bar{q}_{\text{cont}}}{s(q_{\text{cont}})} \quad \text{and} \quad \text{score}_{\text{disc}} = \frac{q_{\text{disc}} - \bar{q}_{\text{disc}}}{s(q_{\text{disc}})}$$

where $\bar{q}$ and $s(q)$ are the mean and the standard deviations, respectively, calculated over a set of reference activity classes.

Small values of the SAR index indicate discontinuous relationships between the activity and descriptors, large values indicate continuous, and intermediate values indicate heterogeneous relationships that combine continuous and discontinuous elements.

 Additional references are collected in the thematic bibliography (see Introduction).

- **subgraph** → graph
- **subgraph centrality** → spectral indices
- **subgraph count set** → count descriptors
- **subgraph ID numbers** $\equiv$ *fragment ID numbers* → ID numbers
- **subgraph property indices** $\equiv$ *SP indices*
- **SubMat binary descriptors** → substructure descriptors (⊙ structural keys)
- **submolecular polarity parameter** → charge descriptors
- **subspectral graphs** → graph

■ substituent constants

These are experimentally determined descriptors of molecular substituents in congeneric series of compounds representing the variation of a measured molecular property Φ when a substituent X replaces a reference group or atom (usually hydrogen) on the skeletal structure.

Substituent constants ϕ_X are estimated as

$$\phi_X = \log \Phi_X - \log \Phi_0$$

where Φ_X is the global property value of the X-substituted compound and Φ_0 is the property value of the reference (parent) compound. The commonly measured $\rightarrow$ *physico-chemical properties* Φ are the equilibrium and rate constants of specific reactions.

The most popular substituent constants are $\rightarrow$ *electronic substituent constants*, $\rightarrow$ *Hansch–Fujita hydrophobic substituent constants*, and $\rightarrow$ *steric substituent constants* such as $\rightarrow$ *Taft steric constant*, $\rightarrow$ *Charton steric constant*, $\rightarrow$ *substituent front strain*, and $\rightarrow$ *steric density parameter*.

📖 [McDaniel and Brown, 1955; Hancock and Falls, 1961; Charton, 1969, 1971; Hansch, Unger *et al.*, 1973; Fujita and Nishioka, 1976; Unger and Hansch, 1976; Hansch, Rockwell *et al.*, 1977; Fujita, 1981, 1983; Sasaki, Takagi *et al.*, 1981, 1992, 1993; Alunni, Clementi *et al.*, 1983; Buydens, Massart *et al.*, 1983; Schultz and Moulton, 1985; van de Waterbeemd, El Tayar *et al.*, 1989; Livingstone, Evans *et al.*, 1992; Stegeman, Peijnenburg *et al.*, 1993; Hansch, Leo *et al.*, 1995; Hasegawa, Kimura *et al.*, 1996; Exner, Ingr *et al.*, 1997; Amić, Davidović-Amić *et al.*, 1998; Hansch, Gao *et al.*, 1998]

■ substituent descriptors

Substituent descriptors represent $\rightarrow$ *physico-chemical properties* of substituents or functional groups or simply their presence in specific positions of a parent molecule. All $\rightarrow$ *structure/response correlations* based on the substituent descriptors involve the hypothesis that modeled property closely depends on the characteristics more of the molecule substituents than the whole molecule, thus the validity of the $\rightarrow$ *congenericity principle* and response additivity are assumed.

Well-known substituent descriptors are the $\rightarrow$ *substituent constants* that are experimentally determined descriptors; among them, $\rightarrow$ *electronic substituent constants*, $\rightarrow$ *steric substituent descriptors*, and lipophilicity substituent descriptors such as $\rightarrow$ *Hansch–Fujita hydrophobic constants* are the most commonly used in QSAR/QSPR modeling.

Moreover, in spite of their holistic character, several molecular physico-chemical properties have also been calculated for substituents; examples are $\rightarrow$ *molar refractivity*, $\rightarrow$ *lipophilicity*, and $\rightarrow$ *surface areas*; size properties of the substituents are often represented by $\rightarrow$ *Sterimol parameters*.

Among the $\rightarrow$ *graph invariants*, several substituent descriptors have been defined, for example, $\rightarrow$ *Kier steric descriptor*, $\rightarrow$ *steric vertex topological index*, and $\rightarrow$ *fragment molecular connectivity indices*.

A great advantage of substituent descriptors is that they are calculated just once and then can be used in a congeneric series of compounds whenever the substituent is present in a site of the parent molecule, independently of the different molecules they are attached, without further calculations. Approaches based on the substituent descriptors usually allow to perform a $\rightarrow$

reversible decoding; however, they need the congenericity of the training set compounds and in most cases do not account for substituent–substituent and substituent–molecule skeleton interactions.

The most known QSAR approaches based on substituent descriptors are the $\rightarrow$ *Hansch analysis* and $\rightarrow$ *Free–Wilson analysis*; in the latter technique, the substituents are defined by $\rightarrow$ *indicator variables* representing their presence/absence in the substitution sites of the parent molecule.

Table S17 Values of some substituent descriptors.

Substituent	MR	π	σ_m	σ_p	$\mathcal{R}$	$\mathcal{F}$	L	B1	B5	SD
H	1.03	0.00	0.00	0.00	0.00	0.00	2.06	1.00	1.00	0.000
CH ₃	5.65	0.56	−0.07	−0.17	−0.13	−0.04	2.87	1.52	2.04	0.807
C ₂ H ₅	10.30	1.02	−0.07	−0.15	−0.10	−0.05	4.11	1.52	3.17	0.923
<i>n</i> -C ₃ H ₇	14.96	1.55	−0.07	−0.13	−0.08	−0.06	4.92	1.52	3.49	–
CH(CH ₃) ₂	14.96	1.53	−0.07	−0.15	−0.10	−0.05	4.11	1.90	3.17	0.970
F	0.92	0.14	0.34	0.06	−0.34	0.43	2.65	1.35	1.35	2.986
Cl	6.03	0.71	0.37	0.23	−0.15	0.41	3.52	1.80	1.80	2.664
Br	8.88	0.86	0.39	0.23	−0.17	0.44	3.82	1.95	1.95	4.994
I	13.94	1.12	0.35	0.18	−0.24	0.40	4.23	2.15	2.15	6.171
NO ₂	7.36	−0.28	0.71	0.78	0.16	0.67	3.44	1.70	2.44	2.448
OH	2.85	−0.67	0.12	−0.37	−0.64	0.29	2.74	1.35	1.93	2.270
SH	9.22	0.39	0.25	0.15	−0.11	0.28	3.47	1.70	2.33	–
NH ₂	5.42	−1.23	−0.16	−0.66	−0.68	0.02	2.78	1.35	1.97	1.228
NHCH ₃	10.33	−0.47	−0.30	−0.84	−0.74	−0.11	3.53	1.35	3.08	–
CHO	6.88	−0.65	0.35	0.42	0.13	0.31	3.53	1.60	2.36	1.507
COOH	6.93	−0.32	0.37	0.45	0.15	0.33	3.91	1.60	2.66	2.037
OCH ₃	7.87	−0.02	0.12	−0.27	−0.51	0.26	3.98	1.35	3.07	–
CH ₂ OH	7.19	−1.03	0.00	0.00	0.00	0.00	3.97	1.52	2.7	1.549
CN	6.33	−0.57	0.56	0.66	0.19	0.51	4.23	1.60	1.60	3.589
SCN	13.40	0.41	0.41	0.52	0.19	0.36	4.08	1.70	4.45	–
NCS	17.24	1.15	0.48	0.38	−0.09	0.51	4.29	1.50	4.24	–
COCH ₃	11.18	−0.55	0.38	0.50	0.20	0.32	4.06	1.60	3.13	1.341
CH=CH ₂	10.99	0.82	0.05	−0.02	−0.08	0.07	4.29	1.60	3.09	–

MR, molar refractivity; π , hydrophobic substituent constant; σ_m and σ_p , overall electronic constants for *meta*- and *para*-position; $\mathcal{R}$ and $\mathcal{F}$, Swain–Lupton resonance and field constants; L, Sterimol length parameter; B1 and B5: Sterimol B parameters; and SD, Dash–Behera steric density parameter.

📖 [Wootton, Cranfield *et al.*, 1975; Borth and McKay, 1985; van de Waterbeemd, El Tayar *et al.*, 1989; Breyer, Strasters *et al.*, 1991; Harada, Hanzawa *et al.*, 1992; Peijnenburg, Thart *et al.*, 1992; Peijnenburg, Debeer *et al.*, 1992; Sello, 1992; Delaney, Mullaley *et al.*, 1993; Hansch, Leo *et al.*, 1995; Jurs, Dixon *et al.*, 1995]

- **substituent drug-likeness index** $\rightarrow$ scoring functions (⊙ biological activity profile score)
- **substituent front strain** $\rightarrow$ steric descriptors
- **substituent steric constant sum** $\rightarrow$ steric descriptors
- **substructural analysis** $\rightarrow$ scoring functions

■ substructure descriptors ($\equiv$ fragment-based descriptors)

Substructure descriptors are $\rightarrow$ *vectorial descriptors* collecting counts of occurrences of pre-defined structural features (functional groups, augmented atoms, pharmacophore point pairs, atom pairs and triangles, surface triangles, etc.) in molecules or binary variables specifying their presence/absence [Crowe, Lynch *et al.*, 1970; Adamson, Lynch *et al.*, 1971]. These string representations of chemical structures are usually designed to enhance the efficiency of chemical database screening and analysis. Each bin or set of bins of the string is associated with a structural feature or pattern. The string length can vary depending on the amount of structural information to be encoded.

Characterization of a molecule by a set of substructures is evident to chemists and directly related to similarity/diversity of chemical structures [Varmuza, Demuth *et al.*, 2005]. Substructure descriptors are mostly used in chemical database handling for exact structure/substructure and similarity searching, in combinatorial chemistry for $\rightarrow$ *similarity/diversity* analysis of compound libraries, and in $\rightarrow$ *group contribution methods* for the evaluation of molecular properties. Some of them were designed specifically to describe molecular shape.

Substructure searching is the most common molecule retrieval mechanism involving the retrieval of all molecules in a database that contain a user-defined substructure query [Hagadone, 1992; Barnard, 1993; Pearlman, 1993; Pepperrell, 1994; Clark and Murray, 1995; Brown and Martin, 1997; Lipkus, 1997; Gillet, Willett *et al.*, 1998; Pickett, Luttmann *et al.*, 1998; Wang and Zhou, 1998; Willett, Barnard *et al.*, 1998; Downs and Willett, 1999; Gao, Williams *et al.*, 1999; Lipkus, 1999; Chen and Reynolds, 2002; Badreddin Abolmaali, Wegner *et al.*, 2003; Xu, 2003; Xue, Godden *et al.*, 2003c; Tovar, Eckert *et al.*, 2008].

$\rightarrow$ *Substructural analysis* is substructure searching where weights are calculated, relating the presence of a specific substructure moiety in a molecule to the probability that the molecule is active in some biological test system [Cramer III, Redl *et al.*, 1974; Hodes, Hazard *et al.*, 1977; Ormerod, Willett *et al.*, 1989, 1990; Craig, 1990; Klopman, 1992; Gillet, Willett *et al.*, 1998]. Approaches similar to substructural analysis are the $\rightarrow$ *Klopman–Henderson cumulative substructure count* and $\rightarrow$ *Hodes statistical–heuristic method*; a further development of substructural analysis is the definition of quantitative $\rightarrow$ *scoring functions*.

Substructure descriptors can be divided into two main classes, 2D substructure descriptors that are based on a $\rightarrow$ *topological representation* of molecules and 3D substructure descriptors that encode spatial relationships, most usually distances and/or angles, between molecular features [Sheridan, Nilakantan *et al.*, 1989; Pepperrell and Willett, 1991; Bath, Poirrette *et al.*, 1994; Allen, Bath *et al.*, 1995]. Moreover, rigid 3D substructure descriptors consider only a single conformation of the molecule, whereas flexible 3D substructure descriptors account for all the geometric features that can be achieved during a conformational exploration of the considered molecule, usually based on the incremental rotation of all of the rotatable bonds [Good and Kuntz, 1995; McGregor and Muskal, 1999; Mason, Morize *et al.*, 1999; Bradley, Beroza *et al.*, 2000; Unity Chemical – Tripos Associates Inc., 2008]. There are a number of ways for decomposing a molecular structure into its constituent substructures. These can be key functional groups, atom types and/or bond types, atom environment descriptors defined by a set of rules, or all the combinations of two, three, or four molecular centers for which distances are evaluated, either topological or geometrical. Molecular centers can be all the constituent elements of molecules such as atom and bond types, points on molecular surface, atom group or ring system centroids, or selected classes of potential pharmacophore points. Many approaches were proposed that transform the structural chemical information of each molecule into a

simplified representation, that is, the $\rightarrow$ *reduced graph*. In a reduced graph, groups of atoms within the structure are collapsed together to form single centers of interest.

When substructure descriptors are collected into a long molecular vector, which is not very effective and time consuming for some applications, they usually need to be compressed to a shorter vector of fixed or variable length. The advantage of the compression is that less storage space is required and the compressed representation of molecules can be searched faster than the uncompressed counterpart. Equivalent representations of binary vectors, which can be considered already a form of compressed representation if the initial bit vector is sparse, are the *index representation* and the *run-length representation* [Baldi, Benz *et al.*, 2007].

The index representation indexes the vector components that are set 1, whereas the run-length representation indexes the length of the corresponding runs (series of 0 bits followed by a 1 bit). An example of these representations is given in Figure S10.

Original 16 bit-string	1	0	0	0	0	1	1	0	1	0	1	0	0	0	1	0
Index representation	1	6	7	9	11	15										
Run-length Representation	0	4	0	1	1	3										

Figure S10 Index representation and run-length representation of a binary vector of 16 bits.

The most widely used compression algorithm for molecular vector representation is the **modulo compression algorithm** (*bit folding*). Vectors of length p are “folded” using a modulo L operator into K shorter vectors of fixed length L with $p = K \times L$. In the binary case, for a given molecule, a bit in position k of the compressed vector is set 1 if and only if there is at least one bit set 1 in any position j that holds the relationship $k = j \bmod L$ in the full vector of length p , where $j = 1, \dots, p$. In other words, folded vectors are obtained when the full vector of length p is divided, for example, into two distinct vectors of the same length L and the logical union operator (OR) is applied to merge them. The result of folding is a shorter vector with higher bit density (Figure S11), the vector density being defined as the ratio of the number of nonzero elements over the vector length. The folding procedure can be repeated until the desired information density or vector length is reached.

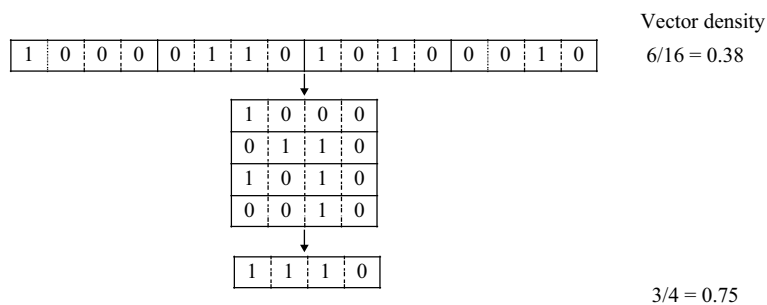


Figure S11 Application of the modulo compression algorithm to a 16-bit string. The string is folded into a binary vector of length 4 modulo 4.

Folding aids in facing problems due to the sparseness of a bit string, which is directly related to its information density, and allows reducing the length of large strings. A drawback of the folding compression is, however, that it is *lossy*: some information is lost during the compression

[Baldi, Benz *et al.*, 2007]. In effect, in folded vectors, different fragments can set the same bit causing fragment collisions. In any case, this does not affect the effectiveness of screening because false positives can be returned for an atom-by-atom matching, but compounds with the searched substructure will never be rejected.

Augmented Atoms (AA, or atom-centered fragment descriptors) are short-range substructure descriptors that count the number of occurrences of augmented atoms in the molecule. The augmented atom is defined in terms of its chemical element together with its connected non-hydrogen atoms [Crowe, Lynch *et al.*, 1970; Adamson, Lynch *et al.*, 1971]. Atoms with connectivities of up to four (excluding bonds to hydrogen atoms) are usually considered. Thus, the largest fragments contain five atoms and four bonds, spanning a distance of two bonds and three connected atoms in a molecule. Augmented Atom codes representing all augmented atoms occurring in the molecules of the chemical data set are recorded in an arbitrary, but fixed, way to give a uniform-length vector.

Augmented Atom keys are defined in a similar way as AAs, but only nonterminal atoms are considered as fragment center and also the bond multiplicity is accounted for [Hodes, 1981b].

Augmented pair descriptors were proposed similarly to augmented atoms descriptors, but focusing on bonds instead of atoms [Crowe, Lynch *et al.*, 1970; Adamson, Lynch *et al.*, 1971]. The augmented pair is defined in terms of two bonded atoms, specifying the bond type, the linked atoms, and the number of external bonds at each end of the pair. **Bonded pair descriptors** were also proposed as an extension of the augmented pairs, further distinguishing pairs of bonded atoms according to the multiplicity of the external bonds.

An extension of the augmented atom keys are the **ganglia-augmented atom keys (gAA)**, which are counts of ganglia-augmented atoms, each one consisting of an augmented atom plus the additional bonds on all the atoms of the fragment [Hodes, 1981b; Tinker, 1981].

Moreover, a generalization of augmented atoms is obtained by a **hierarchical fragment description** of molecules, consisting in representing each molecule as a set of ordered substructures determined concentrically around a focus (e.g., atom, bond, and group centroid). In this approach, the fragment definition is not limited to the first atom neighbors, but proceeds from the focus via the number of neighbors it has to the atom type and bond orders for each neighbor in turn. Atom environment is hierarchically exploited layer-by-layer and the number of considered layers is user-defined or automatically determined by the molecular diameter [Barnard, 1993; Takahashi, Sukekawa *et al.*, 1992; Xiao, Qiao *et al.*, 1997]. Molecular descriptors derived from this approach are sometimes referred to as **circular substructure descriptors**, a circular substructure being a fully explored labeled tree of a particular depth, rooted at a particular vertex. Examples of these descriptors are $\rightarrow$ *indices of neighborhood symmetry*, $\rightarrow$ *structural environment vectors (SEV)*, $\rightarrow$ *multilevel neighborhoods of atoms descriptors (MNA)*, $\rightarrow$ *signature descriptors*, $\rightarrow$ *extended connectivity fingerPrints (ECFPs)*, and $\rightarrow$ *atom environment descriptors*. Moreover, $\rightarrow$ *DARC/PELCO analysis* is based on this hierarchical fragment description of molecules.

Other atom-centered fragments were proposed on the basis of different sets of rules and mainly used in $\rightarrow$ *group contribution methods* for estimation of molecular properties such as $\rightarrow$ *lipophilicity* and $\rightarrow$ *molar refractivity* [Ghose and Crippen, 1986; Viswanadhan, Ghose *et al.*, 1989; Mekenyan, Bonchev *et al.*, 1987; Meylan, Howard *et al.*, 1992; Lohninger, 1994; Baumann and Clerc, 1997; Wildman and Crippen, 1999].

Linear subfragment descriptors are counts of linear chains containing between 3 and 12 interconnected heavy atoms, each described by its chemical element and number of bonded

hydrogen atoms; moreover, two sets of labels are associated with the sequence, one indicating the multiplicity of the bonds that join the atoms of the sequence and the other indicating the presence of a side chain consisting of a terminal functionality such as a halogen, NH_2 , COOH , and so on [Klopman, 1984].

The **topological torsion descriptor** (TT) is related to the 4-atom linear subfragment descriptor of Klopman because it is defined as a Boolean variable for the presence/absence of a linear sequence of four consecutively bonded non-hydrogen atoms $k-i-j-l$, each described by its atom type (TYPE), the number of π electrons (NPI) on each atom, and the number of non-hydrogen atoms (NBR) bonded to it [Nilakantan, Bauman *et al.*, 1987]. Usually NBR does not include $k-i-j-l$ atoms that go to make the torsion itself; therefore, it is -1 for k and l atoms and -2 for the two central atoms i and j . The torsion around the $i-j$ bond and defined by the four indices $k-i-j-l$ is represented by the following TT descriptor:

$$\text{TT} = \{[\text{NPI}-\text{TYPE}-\text{NBR}]_k[\text{NPI}-\text{TYPE}-\text{NBR}]_i[\text{NPI}-\text{TYPE}-\text{NBR}]_j[\text{NPI}-\text{TYPE}-\text{NBR}]_l\}$$

The TT descriptor is a topological analogue of the 3D torsion angle, defined by four consecutively bonded atoms. The topological torsion is a short-range descriptor, that is, it is sensitive only to local changes in the molecule and is independent of the total number of atoms in the molecule.

The use of atom-centered fragments and related descriptors greatly increases the specific chemical information concerning different functional groups, but cannot discriminate between different arrangements of functional groups within a molecule.

Instead, **atom pairs** are substructure descriptors defined in terms of any pair of atoms and bond types connecting them. An atom pair is composed of two non-hydrogen atoms and an interatomic separation:

$$\text{AP} = \{[ith \text{ atom description}][\text{separation}][jth \text{ atom description}]\}$$

The two considered atoms need not be directly connected and the *separation* can be the $\rightarrow$ *topological distance* between them [Carhart, Smith *et al.*, 1985]; these descriptors are usually called **topological atom pairs** being based on the topological representation of the molecules. Atom type is defined by the element itself, the number of heavy-atom connections and number of π electron pairs on each atom.

Unlike topological torsions, atom pairs are sensitive to long-range correlations between the atoms in molecules and therefore to small changes in one part of even large molecules. Atom pair descriptors usually are Boolean variables encoding the presence or absence of a particular atom pair in each molecule.

Using the $\rightarrow$ *geometric distance* in place of the topological distance between any pair of atom types, **geometric atom pair descriptors** were analogously defined [Sheridan, Nilakantan *et al.*, 1989; Sheridan, Miller *et al.*, 1996]. Atom types are defined here by the chemical element, number of heavy-atom connections, number of π electron pairs, number of attached hydrogens, and formal charge.

For a particular conformation of a molecule, all geometrical distances between any pair of atom types are calculated and distributed into a number of discrete bins. For each combination of two atom types, the same number of distance bins is allocated in the final bit string.

Distance bins of equifrequent occurrence are used instead of fixed width bins. To even out the distribution of the number of distances falling into each bin, narrower bins are used around

3–5 Å, which is the region where the frequency distribution of interatomic distances reaches the maximum value.

A suitable empirical formula for the calculation of the position of a distance bin in the string segment corresponding to a particular atom pair is the following:

$$\text{Bin Number} = \text{int} \left[5 \cdot \tan^{-1} \left(\frac{r_{ij}-3}{2} \right) + k \right]$$

where $\tan^{-1}$ is given in radians, r_{ij} is the geometric distance between the i - j atoms, and k is an adjustable parameter. A value of $k = 20$ was suggested by Sheridan [Sheridan, Nilakantan *et al.*, 1989].

→ *Distance-counting descriptors (SE vectors)* are a particular implementation of topological atom pairs proposed by Clerc and Terkovich in 1990. These are holographic vectors encoding information on the occurrence frequency of any combination of two atom types and a distance relationship between them. All the paths and not only the shortest one between any pair of atom types are considered in the original proposal. Based on the shortest path, revised SE vectors were proposed by Baumann in 2002 and called → *SESP-Top vectors* and → *SESP-Geo vectors*.

Similar to atom pairs, **REX descriptors** are defined in terms of pairs of “terminators” and the link between them [Judson, 1992b]:

$$\text{REX} = \{[i\text{th terminator}][j\text{th terminator}][\text{link length}]\}$$

A terminator may be an atom, a lone pair, or a bond; the link is derived from a topological representation of the molecules as the length of the path between the considered terminators. For each pair of terminators, different REX descriptors are defined according to each different link between them, that is, all paths and not only the shortest path may be evaluated.

As terminators, all the atoms of interest (i.e., C, H, N, P, O, S, halogens) can be selected; terminators of the molecule not explicitly considered are classed together as a single dummy terminator. Links having two carbon terminators are usually excluded from REX analysis, considerably reducing the total number of links associated with each structure. However, for an analysis of saturated hydrocarbons, the links between carbon atoms are included.

Individual hydrogen atoms are included as terminators only if they are attached to specific atom types such as oxygen, nitrogen, and sulfur atoms. When distinguishing REX descriptors, the type of atoms and bonds connecting the terminators is not taken into account.

Unlike → *atom pairs*, REX descriptors use more generalized linear fragments, allowing more complex substructures to be dealt with.

Distance Profiles descriptors (or DiP descriptors) are a particular implementation of → *geometric atom pairs*, which is based on simplified atom types [Baumann, 2002b]. Atoms are distinguished according to their chemical type and three additional types that are the generic heavy atom type, the sp^2 -hybridized atom type, and the sp -hybridized atom type. Each atom can be assigned up to three different atom types. Carbon and hydrogen atom types are typically not considered because of their abundance. However, carbon can be assigned the sp^2 - or sp -hybridized type. Geometrical distances are evaluated between any pair of atom types and divided into a number of equal-width bins. Finally, the DiP descriptor is obtained by incrementing by one the bin corresponding to each combination of two atom types and a distance range.

Triangular descriptors (or triplet descriptors) describe the relative positions of three atoms, group centroids, pharmacophore points, or surface points in the molecule [Pepperrell and

Willett, 1991; Bemis and Kuntz, 1992; Nilakantan, Bauman *et al.*, 1993; Fisanick, Cross *et al.*, 1993; Bath, Poirrette *et al.*, 1994; Norel, Fisher *et al.*, 1994; Good and Kuntz, 1995; Good, Ewing *et al.*, 1995]. Each possible triplet of non-hydrogen atoms (or molecular surface points) is taken as a triangle and different geometrical measures can be used to describe them, such as individual triangle side lengths and triangular perimeter and area; these measures are digitized and transformed into bit strings of defined length by different procedures, their distribution is used to describe the molecule.

For instance, triangle side lengths are calculated for all combinations of atom triplets; these lengths are then digitized using a certain distance range (e.g., 1.0 Å) and triplets containing distances greater than a maximum value (e.g., 30 Å) are ignored. The side lengths of valid triangles are sorted (long, intermediate, and short) and their combination is used in the final vector as the identification of the triangle [Good, Ewing *et al.*, 1995]. Alternatively, the side lengths are sorted and scaled by size and subsequently assigned a unique triangle value according to the following equation:

$$t = l_1 + 1000 \cdot l_2 + 1000000 \cdot l_3$$

where t is the final triangle value and l_1 , l_2 , and l_3 are the three digitized and sorted triangle side lengths [Nilakantan, Bauman *et al.*, 1993]. Moreover, each triangle can be assigned a value calculated as the sum of the squares of the side lengths, which can be either 3D geometric distances or 2D topological distances [Bemis and Kuntz, 1992]. This triangle measure can be also combined with another quantity that corresponds to the deviation of that triangle from equality; this quantity is defined as the deviation of the area for a given triangular perimeter from the maximum possible area for the same perimeter [Good and Kuntz, 1995; Good, Ewing *et al.*, 1995]:

$$t = \frac{Area}{Max Area} \cdot \exp \left[- \left(\sqrt{(P/3-l_1)^2} + \sqrt{(P/3-l_2)^2} + 3 \cdot \sqrt{(P/3-l_3)^2/2P} \right) \right]$$

$$Area = \sqrt{P/2 \cdot (P/2-l_1) \cdot (P/2-l_2) \cdot (P/2-l_3)} \quad Max Area = \sqrt{P^4/432}$$

where P is the perimeter and l_1 , l_2 , and l_3 the three side lengths of the triangle.

Triangular descriptors are used both for characterizing molecular shape and for 3D database searching. Making use of geometric interatomic distances, most of the triangular descriptors are specifically designed to account for conformational variation of molecules.

Other common descriptors derived from substructure-based methods are discussed below. Among these, hash structural codes, structural keys, and fingerprints are mostly applied in virtual screening and substructure searching, whereas pharmacophore-based descriptors are more successful in similarity/diversity analysis and QSAR/QSPR studies.

• hash structural codes

Hash structural codes are string representations derived from hashing algorithms and aimed at characterizing molecular structures and speeding up access to molecules in chemical databases [Wipke, Krishnan *et al.*, 1978; Ihlenfeldt and Gasteiger, 1994; Tomczak, 2003]. Instead of searching the database in serial order, the hash code specifies the likeliest location of the searched molecule. A hash code is typically a highly compressed, generally one-way, encoding of a molecular structure with a fixed value range and therefore a fixed bit/byte length. Due to the limited value range, hash codes may collide, or, in other words, two different molecules may

yield the same hash code. When collisions of molecules occur in the hash table, a further step to solve collisions is required. The most common hash codes were calculated based on molecular topology, stereochemistry, charges, and so on [Wipke, Krishnan *et al.*, 1978; Freeland, Funk *et al.*, 1979; Bemis and Kuntz, 1992; Ihlenfeldt and Gasteiger, 1994]. Whatever the specific algorithm used, a highly desired property for structural hash codes is uniqueness. This makes them very effective in exact structure searching, but not very useful in substructure and similarity searching. For these last purposes, the most common molecule string representations are structural keys and fingerprints.

• structural keys

Structural keys are binary vectors in which each element is “true” or “false” and denotes presence or absence of a corresponding structural feature. Examples of such structural features are common functional groups, rings and ring systems, specific atom types, and so forth. Occurrence frequency of specific features may also be encoded by mapping each feature to a set of bits. Structural keys rely on the use of a predefined fragment dictionary, which specifies which feature is encoded by every bit or bit combination of the key.

For an efficient use of the structural keys in library searching, the set of selected fragments should obey two principles [Feldman and Hodes, 1975; Hodes, 1976; Brown and Martin, 1996]:

- (1) Independence of occurrence, avoiding that two fragments occur together thus resulting in redundant information.
- (2) Fragments should be of approximately equiproportional distribution to avoid lack of generality.

Examples of structural keys are $\rightarrow$ *Augmented Atoms* (AA), $\rightarrow$ *atom pairs* and related descriptors, and $\rightarrow$ *atom-type E-state counts*. However, the most common structural keys implemented in specific automated tools are *MACCS keys*, *BCI keys*, and **CACTVS screen vectors** [Ihlenfeldt, Takahashi *et al.*, 1994; Voigt, Bienfait *et al.*, 2001].

Two different **MACCS keys** (or **MDL keys**) [MACCS keys – MDL Information Systems Inc., 2008; Durant, Leland *et al.*, 2002] are commonly encountered, one containing 960 bits and the other, which is public, containing a subset of 166 bits (**ISIS keys**). The fragment dictionary is based on a number of atom types, atom pairs, and custom atom environments. There can be a one-to-one relationship between the structural features and bits, or hashing can be used to create a many-to-one or many-to-many relationship between the features and bits.

A structural feature is defined by nine numbers. The first four numbers (n1–n4) serve to identify the specific atom type, atom pair, or atom environment by means of a predefined set of properties, while the remaining five numbers (n5–n9) determine which bits of the whole key are set by the feature. Specifically, in the case of single atom descriptors, n1 is 0 and n2 and n3 encode one or two properties of the atom; for atom pair descriptors, n1 encodes the number of bonds (topological distance) between the atoms, while n2 and n3 encode the property values of the two atoms; finally, for custom atom environment descriptors, n1 is equal to 7, while n2 encodes the specific atom environment and n3 encodes the property of the atom in the center of that environment. The number n4 encodes the number of occurrences in the molecule of the considered feature. The number n5 is used to specify the number of bits that are set, while n6 is a flag indicating whether or not hashing is allowed; the final three numbers, n7, n8, and n9 identify the bits in the structural key.

Some applications of the MACCS keys discussed in literature are: [McGregor and Pallai, 1997; Ajay, Walters *et al.*, 1998; Xue and Bajorath, 2002; Hert, Willett *et al.*, 2004a].

SSKey-type descriptors are 57 structural keys containing 41 of the 166 MACCS keys, which represent small molecular fragments, with 16 additional structural fragments [Xue, Godden *et al.*, 1999b].

In **BCI keys** (or **Barnard keys**), six different families of fragments are used to define the fragment dictionary [BCI fingerprints – Barnard Chemical Information Ltd, 2008]: Augmented Atoms, Atom/Bond Sequences (atom/bond paths up to a user-specified maximum length, with stereospecific bond types to distinguish configurations around double bonds and at adjacent ring substitution positions), Atom Pairs, Ring Composition Fragments (atom/bond sequences around rings in the Extended Set of Smallest Rings, ESSR), Ring Fusion Fragments (sequences of ring connectivities around ESSR rings), and Ring Ortho Fragments (stereo configuration at nonplanar orthofusion junctions). Fragments are initially generated with fully specified atoms and bonds and are then progressively generalized by using user-defined intermediate-level atom and bond types (e.g., “any ring bond” or “any halogen” and “any atom” and “any bond” types).

Mini-FingerPrints (or **MFP descriptors**) are short binary bit-string representations of molecular structure and properties, constituted by three numerically encoded descriptors (number of hydrogen-bond acceptors *HBA*, number of aromatic bonds *AB*, and fraction of rotatable bonds *RBF*) and a number of structural keys (e.g., specific atom types and bond patterns, structural fragments, functional groups, etc.) [Xue, Godden *et al.*, 1999a, 2000; Xue, Stahura *et al.*, 2001b]. Two types of mini-fingerprints were originally defined: MFP1 and MFP2, consisting of the 3 numerical descriptors and a set of 32 and 40 structural keys, respectively.

To generate a binary string, the hydrogen-bond descriptor value is encoded by using 10 bins, the number of aromatic bonds by 7 bins, and the fraction of rotatable bonds by 5 bins (Table S18). Descriptor values are encoded incrementally: for instance, if there is no hydrogen-bond acceptor, all bits are set off; if one hydrogen-bond acceptor is present, bit 1 is set on; if two are present, both bit 1 and 2 are set on; and so on.

Table S18 MFP scheme for encoding the three numerical descriptors.

Bin	1	2	3	4	5	6	7	8	9	10
<i>HBA</i>	1	2	3	4	5	6	7	8	9	10
<i>AB</i>	1	2	3	4	5	6	7	8	9	≥10
<i>RBF</i>	2–7	8–15	16–19	20–25	26–31	32–37	≥38	—	—	—
	>0	>0.1	>0.2	>0.3	>0.4	—	—	—	—	—
	≤0.1	≤0.2	≤0.3	≤0.4						

HBA, hydrogen-bond acceptor count; *AB*, number of aromatic bonds; and *RBF*, rotatable bond fraction.

The structural keys included in the Mini-Fingerprints encode information about the presence/absence of predefined molecular fragments selected specifically on the basis of their ability to classify compounds according to their biological activity. MFP1 and MFP2 have only four structural keys in common.

MFP consensus profiles are obtained by comparing the bit settings for each active compound belonging to a specific activity class and identifying the bit positions that are set on in all the mini-fingerprints of the active compounds [Xue, Stahura *et al.*, 2001a]. The consensus profile is

then calculated by summing up all bits at each position and dividing the sum by the total number of compounds. Thus, the profile contains the relative frequency (value from 0 to 1) of each bit for the considered activity class. The **consensus fingerprint** is a binary vector where only those bits that have a frequency greater than a cut-off value (e.g., 0.8) are set on (*consensus bits*).

Moreover, to enhance the probability of identifying compounds with similar activity by database searching, mini-fingerprints of all molecules are scaled by using a scaling factor s_k applied only to consensus bits. The scaling factor is linearly weighted within a frequency interval (e.g., 0.8–1.0) as follows:

$$s_k = \begin{cases} 4.0 + \frac{f_k - 0.8}{1.0 - 0.8} & \forall f_k \geq 0.8 \quad 4.0 \leq s_k \leq 5 \\ 1 & \forall f_k < 0.8 \end{cases}$$

where f_k are the relative frequencies of the fingerprint bits [Xue, Godden *et al.*, 2003c]. For instance, bits with frequency of one in the consensus profile are multiplied by 5.0 in the scaled mini-fingerprint, bits with a frequency of 0.9 are multiplied by 4.5, and bits with a frequency of 0.8 are multiplied by 4.0. All bits not set in the consensus fingerprint are left unchanged in the scaled mini-fingerprint.

Mini-FingerPrints of second generation were also proposed that encode a greater number of numerical property descriptors and structural keys [Xue, Godden *et al.*, 2003c]. In **SE-MFP descriptors**, there are 14 binned numerical descriptors and 110 structural keys; to encode the value of each numerical descriptor, 8 bits are used resulting into a string of 222 bits [Xue, Godden *et al.*, 2003a]. In **MP-MFP descriptors** [Xue, Godden *et al.*, 2003b], there are 61 binary encoded numerical descriptors and 110 structural keys for a total of 171 bits. In this case, each descriptor is assigned only one bit position and the bit is set on if the descriptor value for a given molecule is equal to or greater than a reference value (i.e., the descriptor median calculated on selected databases).

SubMat binary descriptors are a set of binary descriptors of different substructure groups covering a large diversity of organic molecules [Scsibraný, Karlovits *et al.*, 2003; Varmuza, Demuth *et al.*, 2005]. The different kinds of substructure groups are collected in Table S19, giving a total number of 1365 substructures. These descriptors are restricted to the most common elements: C, N, O, S, P, F, Cl, Br, I, B, and Si. Moreover, two pseudo-elements have

Table S19 SubMat binary descriptors: substructure groups and number of substructures per group.

Group number	Group definition	No. of substructures
1	Elements (single atom substructures)	46
2	Two-atom substructures	78
3	Single, not aromatic rings	404
4	Condensed, not aromatic rings	130
5	Aromatic rings	97
6	Other rings	39
7	Trees (chains and branches)	418
8	Functional groups	153

been defined: pseudo-element A for heteroatoms (any atom except C or H) and pseudo-element Q for non-hydrogen atoms.

MolDiA descriptors (*Molecular Diversity Analysis descriptors*) are string representations encoding information about cyclic and acyclic fragments present in the molecules [Maldonado, Doucet *et al.*, 2007]. These fragments are identified by breaking down each molecule into its constituent independent cycles and atomic units defined by their specific environment and matching them with fragments of a predefined dictionary.

The fragment dictionary contains high frequency/common fragments and functional groups and specific fragments of pharmaceutical and medicinal interest. Different families of fragments are also created by grouping structurally similar fragments, with the aim of implementing different levels of exact and fuzzy comparison among substructures when analyzing molecules.

Unlike a common structural key, the MolDiA descriptor is a vector of codes of only those fragments generated by the molecule and, thus, its length strictly depends on the molecule size and the meaning of each single vector element also depends on the molecule itself. As a consequence, MolDiA descriptor can be used to measure similarity/diversity among molecules, but not to evaluate structure–property correlations. Moreover, in $\rightarrow$ *similarity/diversity* analysis, structural and property weights can be used for calculating a numerical score associated to each fragment and for encoding structural and physico-chemical information of the fragment.

Activity Class–Characteristic Substructures (ACCS) are those molecular fragments that occur in the randomly generated fragment populations of at least two active molecules but not in the population of the compounds of other activity classes [Lounkine, Batista *et al.*, 2007]. These are specific unique combinations of fragments, whose presence depends on each other, and seem to define the core structures in compound activity classes.

ACCSs for each activity class are determined by an analysis of conditional co-occurrence in populations of random fragments. These, called **MolBlaster fragments**, are obtained through series of random deletions of bonds in connectivity tables of H-depleted molecular graphs of the molecules in analysis.

Once ACCS subset is assigned to a specific activity class, characteristic substructures are mapped back on each molecule by performing a subgraph search for each fragment and whenever a fragment matches an atom of the molecule, a counter for that atom is increased by 1. For each atom, division of its final counter state by the total number of matched substructures gives its match rate. Then, for each active molecule, core structures are defined as the set of all atoms of the molecule that have an ACCS match rate greater than a threshold value.

To quantitatively compare the relationships between core structures in each activity class, bit string representations were generated, where each bit position accounts for the presence or absence of an individual fragment of the corresponding ACCS set.

• fingerprints

Unlike structural keys, which make use of a fragment dictionary, fingerprints are Boolean vectors that define a set of patterns to index. A pattern may be, for example, a path of predefined length, each path being characterized by the nature of atoms and bonds along the path. Fingerprints are generated in such a way to capture the common chemical features present in the molecules of the training set [Shemetulskis, Weininger *et al.*, 1996].

The patterns are generated from the molecule itself, and therefore since the patterns differ from one molecule to another, the meaning of any particular bit is not well defined. Depending on the type of selected pattern and the size of the molecule, the number of distinct fragments produced from any structure may be very large. To reduce their length, fingerprints are usually submitted to a hashing algorithm. Each pattern is used as a seed to a pseudo-random number generator, the output of which is a set of bits (typically 4 or 5 bits per pattern). The binary representations obtained are then combined by the logical OR operator to produce the final fingerprint.

Whereas in a traditional structural key there is a straightforward correspondence between a single bin and a single fragment, in hashed fingerprints different fragments may be encoded into the same bin. Due to these fragment collisions, accuracy of hashed fingerprints is lower than accuracy of structural keys, but the overall efficiency of structure characterization is increased due to the much greater number of encoded fragments.

The **Daylight fingerprints** (DFP) are hashed fingerprints encoding each atom type, all Augmented Atoms and all paths of length 2–7 atoms, giving a total string of 1024 bits [Daylight-James, Weininger *et al.*, 2008]. → *Augmented Atoms* (AA) are defined by a central atom and the nature of atoms and bonds incident to it. Applications of Daylight fingerprints discussed in literature are: [Shemetulskis, Weininger *et al.*, 1996; Dixon and Koehler, 1999; Gillet, Willett *et al.*, 1999, 2003; Raymond and Willett, 2002, 2003; Salim, Holliday *et al.*, 2003; Olah, Bologa *et al.*, 2004b; Stahl and Mauser, 2005; Capelli, Feriani *et al.*, 2006; Fechner and Schneider, 2006; Batista and Bajorath, 2007; Gardiner, Gillet *et al.*, 2007].

Other hashed fingerprints are **Unity fingerprints** that keep information from different length path distinct [Unity Chemical – Tripos Associates Inc., 2008]. Thus, separate regions of a bit string record information from paths of lengths 2, 3, and 4 including hydrogens and 4–6 excluding hydrogens. A few generic structural keys are added for some common atoms and ring system counts, producing a total string of 988 bits. Unity 3D fingerprints are based on pairs of the following features: oxygen, nitrogen, generic oxygen or nitrogen, phenyl ring centroid, point on the normal to the plane of a phenyl ring, and carbonyl extension point. To produce a bit string, a distance range is first defined (e.g., 2–10 Å) and then divided into a number of bins by specifying a bin width (e.g., 0.5 Å). The first bin records all distances smaller than the minimum distance (e.g., 2 Å) and the last all distances greater than the maximum (e.g., 10 Å). Finally, pairs of features between which the distance is to be indexed are derived from the molecule and each is allocated a number of bits in the bit string depending on the distance value.

Applications of UNITY fingerprints discussed in literature are: [Brown and Martin, 1996; Matter, 1997; Matter and Pötter, 1999; Schuffenhauer, Gillet *et al.*, 2000; Wild and Blankley, 2000; Wintner and Moallemi, 2000; Makara, 2001; Cruciani, Pastor *et al.*, 2002; Keseru and Molnár, 2002; Martin, Kofron *et al.*, 2002; Raymond and Willett, 2002; Abrahamian, Fox *et al.*, 2003; Holliday, Salim *et al.*, 2003; Salim, Holliday *et al.*, 2003; Wilton and Willett, 2003; Schuffenhauer, Floersheim *et al.*, 2003; Bender, Mussa *et al.*, 2004c; Bender and Glen, 2005; Schuffenhauer, Brown *et al.*, 2006].

Molecular holograms (or **holograms**) are 2D hashed → *holographic vectors* containing counts of various molecular fragments and substructures occurring in the molecule, defined by a set of rules [HQSAR – Lowis, 1997; Seel, Turner *et al.*, 1999]. To generate molecular holograms, all the molecular structures of the input data set are broken down into all the possible linear

and branched fragments of connected atoms of size between a user-defined minimum (M) and maximum (N) number of atoms. Each unique fragment in the data set is assigned a specific pseudo-random positive integer by means of a cyclic redundancy check (CRC) algorithm. The integer defines the location of a bin in an integer array of fixed length L . Then, the molecular hologram of each molecule is obtained by mapping its specific fragments to the integer array. Specifically, the integer assigned to a fragment by the CRC algorithm is used to select and increment by one the bin in the integer array corresponding to that fragment.

Identical fragments are always hashed to the same bin and, hence, the final bin value is the number of occurrences of a specific fragment in the molecule. Typically, as the total number of unique fragments contained in the data set is much larger than the number (L) of hologram bins, the folding procedure is used to map different unique fragments to the same bin, causing collision among fragments.

Molecular holograms are thus very similar to hashed fingerprints, but rather than using a binary bit string containing either 0 or 1 in each bin, the bins of molecular holograms contain information about the number of fragments hashed to each bin.

In hologram generation, some fragment parameters have to be settled: two fragment size parameters determine the maximum and minimum number of atoms in any one fragment (the default values are $M=4$ and $N=7$ atoms) and five fragment distinction parameters allow fragments to be distinguished according to elemental atom types, bond orders, atomic hybridization states within fragments, hydrogens, and fragment chirality. Also, the hologram length needs to be optimized as it affects the pattern of fragment collisions. Typical hologram lengths range from 50 to 500, even if the number of distinct fragments generated according to the above-cited parameters is around 1000.

The QSAR approach based on molecular holograms, which are correlated with activity and physico-chemical properties by means of PLS analysis, was called **Hologram QSAR** (HQSAR) [Tong, Lowis *et al.*, 1998; Winkler, Burden *et al.*, 1998; Winkler and Burden, 1998; Burden and Winkler, 1999b; So and Karplus, 1999; Viswanadhan, Ghose *et al.*, 1999, 2000; Ducrot, Andrianjara *et al.*, 2001; Mannhold and van de Waterbeemd, 2001; Shi, Fang *et al.*, 2001; Viswanadhan, Mueller *et al.*, 2001; Barker, Gardiner *et al.*, 2003; Wang, Tang *et al.*, 2003; Hirons, Holliday *et al.*, 2005; Melville and Hirts, 2007]. This approach aims at identifying substructural fragments relevant to biological activity in sets of active compounds.

Similar to molecular holograms, but based on a different set of rules, are **ISIDA descriptors** (or **SMF descriptors**, standing for *Substructural Molecular Fragment descriptors*) [Solov'ev, Varnek *et al.*, 2000; Varnek, Wipff *et al.*, 2002]. These are holographic descriptors representing the occurrence frequency of different molecular fragments generated by breaking down all the molecules in the training set.

Two main types of fragments are defined: sequences (i.e., paths) up to a specified length and $\rightarrow$ *Augmented Atoms* (AA). Sequences are fragments comprised of interconnected atoms between a user-defined minimum and maximum number of atoms. These can be distinguished either according to atom types (A) or bond types (B), or both (AB). Only the shortest path between the two atoms is used to define sequences. Augmented atoms are single atoms defined by their element type and environment; also for augmented atom definition, one can consider either connected atoms and bonds (AB) or only types of

connected atoms (A) or only bonds (B). To obtain more discriminating augmented atoms, the hybridization state of all the atoms in the fragment can be taken into account (option available for type A).

Parameters for fragment generation that need to be settled are type of fragments, minimum and maximum number of atoms (or bonds) in the sequences (between 2 and 6), and type of environment for augmented atoms. Once all the molecules in the data set have been analyzed, only independent fragments are retained, considering, for example, the C–C–N and N–C–C sequences as one fragment. Linearly dependent fragments form a single group defined as an extended fragment. Moreover, fragments of rare occurrence (i.e., less than two in the molecule data set) are excluded from the final vectorial descriptor.

ISIDA descriptors were originally proposed for molecular property estimation by using both linear and nonlinear $\rightarrow$ *group contribution methods*. Applications of ISIDA descriptors discussed in literature are: [Solov'ev and Varnek, 2003, 2004; Katritzky, Kuanar *et al.*, 2005; Varnek, Fourches *et al.*, 2005; Horvath, Bonachera *et al.*, 2007; Konovalov, Coomans *et al.*, 2007; Varnek, Kireeva *et al.*, 2007].

NASAWIN descriptors are $\rightarrow$ *holographic vectors* encoding information on a number of molecular fragments, distinguished into chains (1–6 atoms), cycles (3–6 member), and several types of branched fragments [Zefirov and Palyulin, 2002; Artemenko, Baskin *et al.*, 2003; Baskin, Halberstam *et al.*, 2003].

Each atom in a fragment can be described using up to four levels of classification in accordance with its neighborhood providing significant flexibility to account for heteroatoms, functionality, bond types, and so on. Specification of each fragment also includes sequence of bond orders. These substructure descriptors were called *FRAGMENT descriptors*.

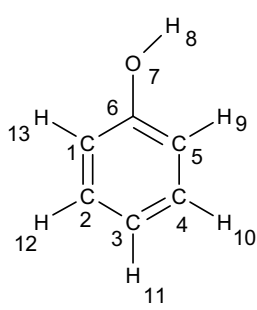
FRAGPROP descriptors were further proposed by combining values of atomic properties (the number of electrons and lone pairs, atomic radius, electronegativity, ionization potential, etc.) within different substructural fragments (chains containing up to five atoms) where each combination has some physical meaning. Hydrogen atoms are represented explicitly in FRAGPROP but not in FRAGMENT type descriptors [Varnek, Kireeva *et al.*, 2007].

Multilevel Neighborhoods of Atom descriptors (or **MNA descriptors**) are $\rightarrow$ *holographic vectors* comprised of the occurrence frequencies of all the different fragments present in the molecules of the data set and identified by a method based on a $\rightarrow$ *hierarchical fragment description* [Filimonov, Poroikov *et al.*, 1999; Poroikov and Filimonov, 2001].

To generate MNA descriptors, an iterative procedure is used. First, atom types are assigned based on the chemical element and distinguished depending on whether atoms belong or not to cyclic systems. Moreover, in defining atom types, bond types are not accounted for, while hydrogens are according to valencies and partial charges of atoms. Then, each atom is described by its atom type in the descriptor of level 0 (MNA/0), by its atom type and atom types of the first neighbors in the descriptor of level 1 (MNA/1). In the descriptor of level 2 (MNA/2), atom types of the atoms bonded to first neighbors of the focused atom are also considered. Therefore, the descriptor of each successive level is a concatenation of the zero-level descriptor of the atom and, enclosed in parentheses, a lexicographically ordered list of descriptors of the previous level of its nearest neighbors. MNA descriptors are implemented in the computer system $\rightarrow$ PASS.

Example S8

Representation of phenol by MNA descriptors of zero (MNA/0), first (MNA/1), and second (MNA/2) levels. Hyphen is the chain marker for the atoms in the chain.

	Atom	MNA/0	MNA/1	MNA/2
	1	C	C(CC-H)	C(C(CC-H)C(CC-O)-H(C))
	2	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
	3	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
	4	C	C(CC-H)	C(C(CC-H)C(CC-H)-H(C))
	5	C	C(CC-H)	C(C(CC-H)C(CC-O)-H(C))
	6	C	C(CC-O)	C(C(CC-H)C(CC-H)-O(C-H))
	7	-O	-O(C-H)	-O(C(CC-O)-H(-O))
	8	-H	-H(C)	-H(-O(C-H))
	9	-H	-H(C)	-H(C(CC-H))
	10	-H	-H(C)	-H(C(CC-H))
	11	-H	-H(C)	-H(C(CC-H))
	12	-H	-H(C)	-H(C(CC-H))
	13	-H	-H(C)	-H(C(CC-H))

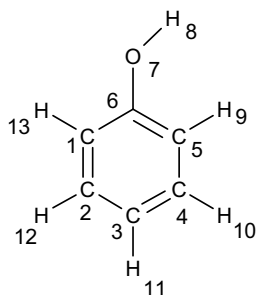
Applications of MNA descriptors discussed in literature are: [Poroikov, Filimonov *et al.*, 2000, 2003; Fomenko, Filimonov *et al.*, 2006].

MOLPRINT-2D fingerprints [Bender, Mussa *et al.*, 2004b, 2004c] are bit-string representations encoding information on the presence/absence of count vectors of atom types. These count vectors are called **Atom Environment descriptors** and are derived from $\rightarrow$ *hierarchical fragment description* approach [Xing and Glen, 2002]. To generate MOLPRINT-2D fingerprints, all atom environment descriptors are first calculated for each non-hydrogen atom in all the molecules of the data set and then a bit string is constructed where each bin is assigned a unique atom environment descriptor in the data set.

Atom environment descriptors are calculated by a two-step procedure: in the first step, the Sybyl atom types are assigned to all heavy atoms; in the second step, count vectors of the atom

Example S9

Atom environment descriptor for carbon C₁ of phenol. Atoms are assigned Sybyl atom types (C₁ = C.ar type). A topological distance up to four is considered. The final count vector is obtained by linking the five layers (0–4) into a unique string.

	Layer	C.3	C.2	...	C.ar	...	O.3	O.2	...	H	...
	0	0	0	...	1	...	0	0	...	0	...
	1	0	0	...	2	...	0	0	...	1	...
	2	0	0	...	2	...	1	0	...	1	...
	3	0	0	...	1	...	0	0	...	3	...
	4	0	0	...	0	...	0	0	...	1	...

types from the central atom up to a given topological distance are constructed for each heavy atom in the molecule. A count vector is a string comprising the number of atoms of each type in each layer, that is, at each topological distance from the focused atom. Therefore, the count vector describing each atom consists of a number of bins equal to the number of atom types multiplied by the number of layers.

Applications of the MOLPRINT-2D descriptors discussed in literature are: [Bender and Glen, 2005; Cannon, Amini *et al.*, 2007; Givvehchi, Bender *et al.*, 2006; Batista and Bajorath, 2007; Li, Bender *et al.*, 2007; Tovar, Eckert *et al.*, 2008].

Similar to $\rightarrow$ MNA descriptors, **signature descriptors** are holographic vectors of count numbers of all the atomic signatures present in the data set molecules, the signature of an atom being a canonical representation of the atom environment up to a predefined depth, that is, a circular substructure descriptor [Visco, Pophale *et al.*, 2002; Faulon, Visco *et al.*, 2003].

The signature of an atom of depth h takes the form of a tree rooted at the atom itself and obtained by a five-step procedure: (1) The subgraph containing all atoms at topological distance h from the focused atom x is extracted; (2) the subgraph vertices are labeled in a canonical order, atom x having label 1; (3) a tree spanning all edges of the subgraph is constructed, the root of the tree being the atom x ; the first layer of the tree is composed of the first neighbor of x , the second layer of the vertices at topological distance 2, and so on up to topological distance h . The neighbors of each vertex are sorted in decreasing lexicographic order; moreover, each edge can appear only once. (4) Once the tree has been constructed up to layer h , all canonical labels that appear only once are removed and the remaining labels are renumbered in the order they appear. (5) The signature is finally written by reading the tree in a depth first; the character "(" is entered each time an edge parent-child is read, while the character ")" when the edge is read from child to parent.

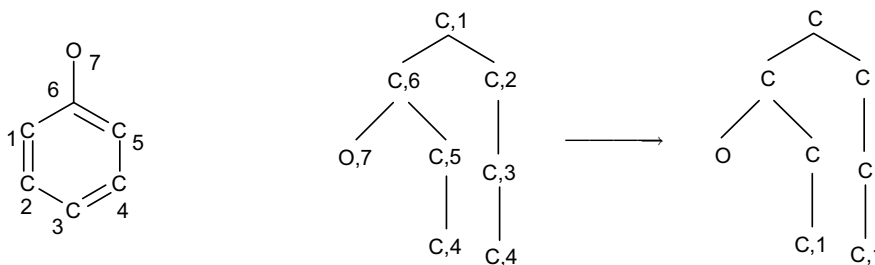
Therefore, the signature of an atom can be thought of as a string of characters. The signature of a molecule $^h\sigma$ of depth h is a linear combination of the signatures $^h\sigma_i$ of the molecule atoms:

$$^h\sigma = \sum_{i=1}^A {}^h\sigma_i = \sum_{g=1}^G {}^hn_g \cdot {}^h\sigma_g$$

where A is the number of atoms, G the number of different atomic signatures, and ${}^hn_g$ the number of occurrences of the g th atomic signature. These occurrence numbers are then collected in the final signature descriptor.

Example S10

Signature descriptor up to depth 3 for atom C_1 of chlorophenol.



$$\begin{aligned} {}^0\sigma(C_1) &= C & {}^2\sigma(C_1) &= C(C(O\,C)C(C)) \\ {}^1\sigma(C_1) &= C(C\,C) & {}^3\sigma(C_1) &= C(C(O\,C(C,1))C(C(C,1))) \end{aligned}$$

Molecular signatures were demonstrated to be useful for $\rightarrow$ reversible decoding of QSAR models based on topological indices: as any topological index can be computed from the molecular signature, a QSAR model can be replaced with an equivalent model involving occurrence numbers of substructural fragments (i.e., atomic signatures).

Applications of the signature descriptors discussed in literature are: [Faulon, Churchwell *et al.*, 2003; Churchwell, Rintoul *et al.*, 2004; Faulon, Collins *et al.*, 2004; Weis, Faulon *et al.*, 2005].

Extended Connectivity FingerPrints (or **ECFP fingerprints**) encode information on atom-centered fragments identified by a variant of the $\rightarrow$ Morgan's extended connectivity algorithm [PipeLine Pilot – Scitegic Inc., 2005; Rogers, Brown *et al.*, 2005]. First, atom types present in each molecule are identified on the basis of the following features: number of connections (bonds), element type, charge, and atomic mass. At iteration 0, the fingerprint ECFP_0 is obtained that encodes information on individual atoms. Then, an iterative process is used to evaluate atom environment to generate larger substructural fragments. At iteration 1, the information of all atoms directly bonded to each atom (within a diameter of 2 chemical bonds, and hence termed ECFP_2) is encoded. At iteration 2 (ECFP_4), the information of all atoms within a diameter of 4 chemical bonds is also encoded. When the desired neighborhood size (6 by default) is reached, the process is complete. All the fragments present in the data set molecules are stored in the final ECFP fingerprint as single bins. Extended Connectivity FingerPrints are conceptually the same as $\rightarrow$ MOLPRINT 2D fingerprints, but different atom types are used.

A variant of the ECFP fingerprints is represented by the **Functional Connectivity FingerPrints** (or **FCFP fingerprints**), where atoms are characterized by functional types: hydrogen-bond acceptor, hydrogen-bond donor, positively ionizable, negatively ionizable, aromatic, and halogen [Hert, Willett *et al.*, 2004b; Rogers, Brown *et al.*, 2005; Hassan, Brown *et al.*, 2006]. For instance, in FCFP, all halogens give the same atom bit codes, whereas, in ECFP, they are characterized by different bit codes.

Modified ECFP fingerprints, called **ECFC fingerprints**, are based on the use of counts of how many times each fragment is present in the molecules instead of binary variables [Liu and Zhou, 2008]. Still based on fragment counts are **FCFC fingerprints**, which are vectors of Counts of Fragments based on Connectivity of atoms belonging to specified Functional Classes [Gedeck, Rohde *et al.*, 2006].

Applications of Extended Connectivity fingerprints discussed in literature are: [Schuffenhauer, Brown *et al.*, 2006; Jensen, Vind *et al.*, 2007; Baldi, Benz *et al.*, 2007].

Avalon fingerprints are hashed fingerprints including atoms, augmented atoms, atom triplets, connection paths, and ring description, divided into 16 feature classes [Gedeck, Rohde *et al.*, 2006; Baldi, Benz *et al.*, 2007]. The defined feature classes are listed in Table S20; more detailed information is given in the supplementary material of [Gedeck, Rohde *et al.*, 2006].

Table S20 Avalon feature classes.

Feature class	Description	Average no. of bits
ATOM_COUNT	Count ranges of certain atom types, bond types, and special atom environments	5.7
ATOM_SYMOL_PATH	Paths of atom bond sequences of varying length and specificity	38.8
AUGMENTED_ATOM	Indicators of different single shell atom environments	13.9

(Continued)

Table S20 (Continued)

Feature class	Description	Average no. of bits
AUGMENTED_BOND	Combined indicators of bond end environments	4.0
HCOUNT_PAIR	Hydrogen presence at bond ends	5.8
HCOUNT_PATH	Paths starting at hydrogen bearing atoms	18.3
RING_PATH	Paths restricted to ring bonds	4.2
BOND_PATH	Paths ignoring atom type	21.1
HCOUNT_CLASS_PATH	Similar to HCOUNT_PATH but only distinguishing carbon from heteroatoms	9.5
ATOM_CLASS_PATH	Selected paths of carbon/heteroatoms ignoring bond order	13.9
RING_PATTERN	Paths indicating uncommon ring features	7.5
RING_SIZE_COUNTS	Counts of bonds contained in rings of different sizes	4.5
DEGREE_PATHS	Paths indicating graph degree	25.1
CLASS_SPIDERS	Graph distance triples to a central atom for special central and leaf atom types	8.9
FEATURE_PAIRS	Graph distance pairs for special end point atom types	20.8
ALL_PATTERNS	Bits from all feature classes together	157.6

• pharmacophore-based descriptors

These are substructure descriptors based on the concept of pharmacophore, which can be thought of as a set of structural features in a spatial arrangement, which represents the interactions made in common by a set of ligands with a protein receptor. Provided that a pharmacophore is necessary for molecule–receptor bonding, it is considered to encode important information about bioactive shape and electronic properties.

Several pharmacophore-based descriptors were proposed, which differ from each other by how structural features are defined, relationships among features are accounted for, and this information is encoded into a final vectorial representation. However, the basic philosophy underlying all these descriptors is that atoms are distinguished according to how they behave in biological systems. Thus, for instance, the nitrogen in a guanidium and the nitrogen of a nitro are not considered to be equivalent [Brown and Martin, 1996]. Moreover, once the pharmacophore-based features have been defined, different molecular descriptors can be derived by all the combinations of two, three, four, or more features (Figure S12).

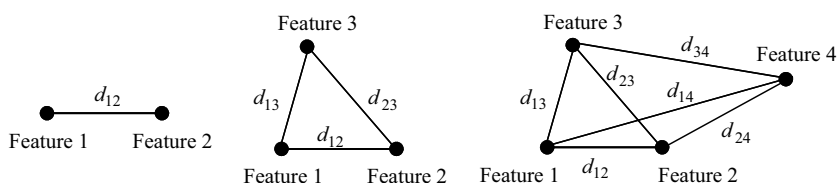


Figure S12 Examples of 2-, 3-, and 4-point pharmacophores.

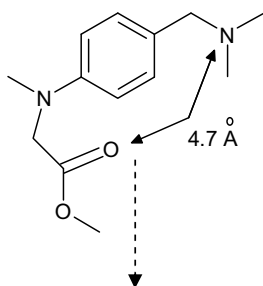
Binding property torsions (bt) are binary descriptors for the presence/absence of linear sequences of four interconnected non-hydrogen atoms. These descriptors were proposed as an extension of $\rightarrow$ *topological torsions* by assigning atoms to seven different types of potential pharmacophore points: cations, anions, neutral hydrogen-bond donors and acceptors, polar atoms (both donor and acceptor, e.g., hydroxyl oxygen), and hydrophobic atoms and others [Kearsley, Sallamack *et al.*, 1996].

Binding property pairs (bp) were also proposed by the same authors as an extension of $\rightarrow$ atom pairs by assigning atoms to the seven pharmacophore point types defined for binding property torsions [Bush and Sheridan, 1993; Kearsley, Sallamack *et al.*, 1996; Sheridan, Miller *et al.*, 1996, 1998]. Binding property pairs can be either topological (i.e., **topological binding property pairs**) or geometric (i.e., **geometric binding property pairs**) depending on the method used to measure the separation between atom pairs.

In calculating hydrophobic pairs and torsions and charge pairs and torsions, since hydrophobic contribution and charge are continuous rather than discrete entities, a set of overlapping bins numbered 1–7 is used to represent ranges of these properties: (1) ≤ -0.50 , (2) -0.75 to -0.25 , (3) -0.50 to 0.00 , (4) -0.25 to 0.25 , (5) 0.00 to 0.50 , (6) 0.25 to 0.75 , and (7) ≥ 0.50 . Each atom is allowed to be in two different bins; for instance, an atom with a charge of -0.33 would be in the bin 2 and 3. Moreover, to generate geometric binding property pairs, interfeature distances are divided into a number of discrete bins. To allow an overlap of the distance bins, a particular pair of features is allowed to occupy more than one bin in proportion to where its distance falls between the bin centers [Sheridan, Miller *et al.*, 1996]. Distances were divided into 30 bins, starting at $1 \text{ \AA}$ and ending at $75.3 \text{ \AA}$. The interval between the first and second bin starts at $0.5 \text{ \AA}$ and thereafter it increases, thus the centers of the bins are approximately <1.0 , 1.5 , 2.1 , 2.7 , 3.3 , 4.1 , 4.9 , 5.7 , 6.7 , 7.8 , 9.0 , 10.3 , 11.7 , 13.3 , 15.5 , and so on.

Example S11

Calculation of geometric binding property pairs. For each bin, the distance center and the bin number are shown. The atom pair highlighted (NX3-OX1) is characterized by a distance of $4.7 \text{ \AA}$ and therefore it contributes $1 - \frac{3}{4} = 0.25$ to the total fuzzy count of the bin NX3-(6)-OX1 and 0.75 to the total count of the bin NX3-(7)-OX1.



Bin centers		3.3	4.1	4.9	5.7	6.7	7.8	9.0	
....	bin 5	bin 6	bin 7	bin 8	bin 9	bin 10	bin 11		

According to this binning scheme, the two distance bins, whose centers are the closest ones to the distance r_{ij} of a given combination of two features, are incremented proportionally leading to *fuzzy counts* of the binding property pairs. A simple formula for the calculation of the increments Δ for the main bin m , that is, the bin where the r_{ij} distance actually falls, and the adjacent bin a is

$$\Delta_m = 1 - \left| \frac{\bar{b}_m - r_{ij}}{\bar{b}_a - \bar{b}_m} \right| \quad \Delta_a = 1 - \Delta_m$$

where $\bar{b}$ is the bin center. If $\bar{b}_m - r_{ij}$ is negative, the adjacent bin is the left one with respect to the main bin, otherwise it is the right one [Stiefl and Baumann, 2003].

TGD fingerprints, consisting of 420 bits, are topological binding property pairs calculated covering 15 distance values. Distances between any two types of the seven pharmacophore features are determined as the shortest connecting paths in the molecular graph representation.

TGT fingerprints, consisting of 1704 bits, are an extension of TGD fingerprints encoding triplets of pharmacophore features. Interfeature distance is yet topological but only six distance values are used. Moreover, 3-point pharmacophores are determined by only four features. Both TGD and TGT fingerprints are implemented in the program MOE [MOE – Chemical Computing Group Inc., 1999]. Applications of these fingerprints are discussed in: [Eckert and Bajorath, 2006a, 2007a; Vogt, Godden *et al.*, 2007; Wang, Eckert *et al.*, 2007; Tovar, Eckert *et al.*, 2008].

Potential Pharmacophore Point pairs (or PPP pairs) are 3D string representations of molecules encoding the geometrical distance information between all possible combinations of two potential pharmacophore points [Brown and Martin, 1996]. PPP pairs are a particular case of $\rightarrow$ *geometric atom pairs* where only pairs of potential pharmacophore points are considered, thus resulting very similar to $\rightarrow$ *geometric binding property pairs*.

Potential pharmacophore points are hydrogen-bond donors and acceptors, sites of potential negative and positive charge interaction, and hydrophobic atoms. All atoms of the molecule are analyzed to see whether they can be classed potentially as one of these point types. Atoms can be assigned to one or more types of pharmacophore point. Groups of connected hydrophobic atoms are considered as a single geometric centroid. Finally, each molecular structure is reduced to its pharmacophore points and geometric distances between all such points are calculated. Then, in PPP pair descriptors, the distance information between all the 15 combinations of pairs (donor–donor, donor–acceptor, etc.) of the five point types is encoded into a bit string. Each combination of pairs is assigned a number of bins in the bit string, each bin corresponding to a range of distance values. As for geometric atom pairs, bins of quasi-equifrequent occurrence are used instead of fixed width bins. Then, each distance value selects and increments by one the bin whose location in the bit string is given by

$$\text{Bin Number} = \text{int} \left[5 \cdot \tan^{-1} \left(\frac{r_{ij} - 3}{2} \right) + 6 \right]$$

where $\tan^{-1}$ is given in radians and r_{ij} is the geometric distance between the i – j atoms.

There are other 3D string representations similar to PPP pairs that are based on extended sets of 17 potential pharmacophore points containing also polar hydrogen, nonpolar hydrogen, nitrogen, and oxygen together with aromatic ring center and triple and double bond centers [Chen, Rusinko III *et al.*, 1998]. Definitions of these potential pharmacophore points are given in Table S21. An atom can give rise to more than one PPP type; for instance, the oxygen atom in carbonyl group is classified to both type 3 (hydrogen-bond acceptor) and type 12 (oxygen atom).

The distance between any two PPP types is measured and assigned into one or two distance bins. The width of each distance bin is usually 1.0 Å. However, since adjacent bins are allowed to have 10% overlap with each other, the actual bin width is 1.2 Å. Thus, as for $\rightarrow$ *geometric binding property pairs*, any distance located in the overlap region is assigned to both of the bins; for instance, a distance of 2.05 Å, under fuzzy boundary conditions, belongs to both bin 1 and bin 2.

All the distances larger than 20 Å are assigned to the last bin. Thus, 0.0–1.1 Å is the bin 0, 0.9–2.1 Å is the bin 1, 1.9–3.1 Å is the bin 2, . . . , and ≥ 19.9 Å is the bin 20.

When multiple conformations are available to represent a compound, all the PPP pairs that exist in any of the possible conformations are used to describe that compound, thus leading to a 3D flexible vectorial descriptor.

Table S21 Potential pharmacophore point types from [Chen, Rusinko III *et al.* 1998].

ID	PPP type	ID	PPP type
1	Negative charge center, including carboxylic group, sulfinic group, phosphinic group, etc.	10	C atom
2	Positive charge center: all N in primary, secondary, and tertiary amines	11	N atom
3	Hydrogen-bond acceptor: all N, O, and S with at least one available lone pair electron	12	O atom
4	Polar hydrogen atom: H atoms linked on N, O, S, or the terminal of a triple bond	13	S atom
5	Nonpolar hydrogen atom: all H atoms linked on the carbon	14	P atom
6	Hydrogen atom, including both polar and nonpolar H atoms	15	F atom
7	Triple bond center	16	Cl, Br, or I
8	Double bond center	17	Other element
9	Aromatic ring center		

CATS descriptors are very similar to the PPP pair descriptors, the main difference being the topological distance between any pair of pharmacophore point types used in place of the geometrical distance [Schneider, Neidhart *et al.*, 1999; Fechner, Franke *et al.*, 2003]. Moreover, whereas PPP pair descriptors are bit strings, CATS descriptors are $\rightarrow$ *holographic vectors* where each bin encodes the number of times a PPP pair occurs in the molecule.

The five defined potential pharmacophore points are hydrogen-bond donor (D), hydrogen-bond acceptor (A), positively charged or ionizable (P), negatively charged or ionizable (N), and lipophilic (L). If an atom does not belong to any of the five PPP types, it is not considered. Moreover, an atom is allowed to be assigned to one or two PPP types (Figure S13).

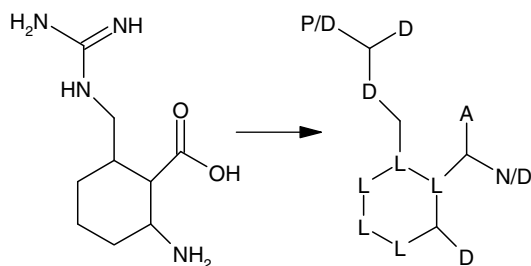


Figure S13 Conversion of a two-dimensional molecular representation into the molecular graph, in which pharmacophore point types are assigned as implemented in CATS.

For each molecule, the number of occurrences of all 15 possible pharmacophore point pairs (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) is determined and then associated with the number of intervening bonds between the two considered points, whereby the shortest path length is used. Topological distances of 0–9 bonds are considered leading to a 150-dimensional autocorrelation vector.

Finally, PPP pair counts are scaled by the total frequency in the molecule.

CATS descriptors defined above are better named **CATS2D descriptors** because they are based on topological distances.

CATS3D descriptors [Renner, Noeske *et al.*, 2005] are based on geometrical distances between PPPs. Hydrogens are also considered in PPPs pairs calculation and 20 equal-spaced bins from 0 to 20 Å are used. Unlike CATS2D, multiple potential pharmacophore point assignments of one atom are not allowed. Moreover, an additional type is defined to account for atoms assigned to none of the five PPP types: a total of 28 possible PPP pairs is thus obtained and to each of them, 20 distance bins are assigned, resulting into a 560-dimensional vector.

SURFCATS descriptors [Renner and Schneider, 2006] are based on the spatial distance between PPPs on the → *Connolly surface area*. Surface points are calculated with a spacing of 2 Å and assigned to the pharmacophore type of the nearest atoms.

CATS-Charge descriptors [Fechner, Franke *et al.*, 2003] map the partial atom charges of a molecule to predefined spatial distance bins. The geometrical distances of all atom pair combinations in one molecule are calculated. Distances within a certain range (0.1 Å) are allocated to the same bin. The charges of the two atoms that form a pair are multiplied to yield a single charge value per pair. Charge values that are assigned to the same bin are summed up. Distances from 0 to 10 Å are considered with increments of 0.1 Å. All distances greater than 10 Å are associated with the last bin. The output is a 100-dimensional vector, which characterizes the molecule by means of its atom partial charge distribution. This vectorial descriptor is based on the same approach as the 3D → *autocorrelation descriptors*.

Applications of CATS descriptors discussed in literature are: [Zuegge, Fechner *et al.*, 2002; Byvatov, Fechner *et al.*, 2003; Fechner and Schneider, 2004a, 2007; Merkwirth, Mauser *et al.*, 2004; Evers, Hessler *et al.*, 2005; Fechner, Paetz *et al.*, 2005; Renner, Ludwig *et al.*, 2005; Schneider and Fechner, 2005; Noeske, Sasse *et al.*, 2006; Franke, Schwarz *et al.*, 2007].

MaP descriptors (*Mapping Property distributions of molecular surfaces*) are count vectors of 2-point pharmacophores evaluated on the molecular surface, thus resulting conceptually similar to the → *SURFCATS descriptors* [Stiefl and Baumann, 2003; Stiefl, Bringmann *et al.*, 2003]. The procedure to calculate MaP descriptors is threefold. First, an approximation to the molecular surface with equally distributed surface points is computed. Next, pharmacophoric properties are projected onto the molecular surface according to the following rules: (1) Atoms in the molecule are assigned a pharmacophore point type (hydrophobic (L), hydrophilic (H), hydrogen-bond acceptor (A), and hydrogen-bond donor (D)); (2) the atom closest to a given surface point is defined as its base atom; (3) if the base atom is classified as an hydrogen-bond acceptor (A) or hydrogen-bond donor (D), the surface point is assigned the same pharmacophore point type as the base atom (i.e., A or D); and (4) only surface points that are not classified as the latter can be assigned the hydrophobic (L) or hydrophilic (H) type; in this case, hydrophobic type is assigned according to the → *hydrophobic potential* of the surface point.

Finally, geometrical distances between any two surface points are calculated and mapped to a count vector where a certain number of bins are allocated for each combination of two

pharmacophore types. For each 2-point combination, the considered distance ranges (bin_k) are

$$(k-1) \cdot res + res/2 \leq bin_k < k \cdot res + res/2 \quad k = 0, 1, 2, \dots$$

where res is the bin width (set at 1 Å by default) and k is increased up to the maximum number of bins selected to encode each 2-point combination. The matching bins of each surface point pair are incremented according to the scheme of the fuzzy counts used for $\rightarrow$ *geometric binding property pairs*. An application of MaP descriptors discussed in literature is: [Baumann and Stiefl, 2004].

PPP triangle descriptors [Brown and Martin, 1996] are bit strings encoding the geometrical distance information about all the 35 possible combinations of three out of the five pharmacophore point types defined for the $\rightarrow$ *PPP pairs* (donor–donor–donor, donor–donor–acceptor, etc.). Each combination of three pharmacophore types defines a triangle and individual side lengths are taken as triangle measure used to identify a bin in the bit string.

For each PPP triangle, all combinations of the side lengths between a user-defined minimum (e.g., 2 Å) and maximum distance (e.g., 15 Å), given a distance bin width (e.g., 1 Å), are allowed if the triangle inequality is satisfied. Typically, the first bin of each PPP triangle represents all the combinations of the three side lengths smaller than 2, that is, $\{<2, <2, <2\}$, whereas the last bin is associated with triangle having all the distances larger than or equal to 15 Å, that is, $\{\geq 15, \geq 15, \geq 15\}$. Therefore, for a defined conformation of a molecule, each bit represents a particular combination of pharmacophore points and distance. As the number of all the possible PPP triangles, provided 5 PPP types and 15 distance ranges for each triangle side length, is very large (48 000), the hashing procedure is usually applied to give a shorter bit string (e.g., 2048). To this end, each bin is randomly assigned a number of bits in the shorter string.

The **Pharmacophore-Derived Query descriptors** (or **PDQ descriptors**) are based, as the PPP triangle descriptors, on 3-point pharmacophores [Pickett, Mason *et al.*, 1996; Ashton, Jaye *et al.*, 1996] and are derived from 3D queries of the ChemDBS-3D database of the program Chem-X [ChemDiverse – Chemical Design Ltd, 2008]. The procedure for PDQ descriptor calculation first implies assignment of atoms and group centroids to six PPP types: hydrogen-bond donor, hydrogen-bond acceptor, acidic center, basic center, hydrophobe, and aromatic ring centroid. Six distance ranges (Å) are used to cover most expected pharmacophore sizes: 2–4.5, 4.5–7, 7–10, 10–14, 14–19, and 19–24.

Then, queries are generated covering all possible triplets of PPP types and geometrical interatomic distances (i.e., 5916 geometrically valid queries) and used to search the molecules in the analyzed data set. To account for conformational flexibility of molecules, a tolerance is associated with each point–point distance. Finally, the molecular descriptor is obtained as a bit string indicating the queries (i.e., pharmacophores) hit by the compound. For each pharmacophore, it is possible to identify not only if a compound or database of compounds can express it, but also how many times it can express.

PharmPrint descriptors are other binary descriptors encoding geometrical distance information about all possible combinations of three PPP types [McGregor and Muskal, 1999, 2000]. In PharmPrint descriptors generation, atoms are first assigned a pharmacophore type; hydrogens, if present, are ignored. Then, bonds are rotated to generate multiple conformations. Finally, geometric interatomic distances are calculated in each valid conformation.

The pharmacophore point types used to derive PharmPrint descriptors are hydrogen-bond acceptor (A), hydrogen-bond donor (D), sites of formal negative (N) and positive (P) charges, and hydrophobic (H) and aromatic (R) groups. A seventh type (X) is used to label all the atoms that are not assigned to any PPP type.

Geometrical interfeature distances are considered if falling into one of the following distance ranges (Å): 2.0–4.5, 4.5–7.0, 7.0–10.0, 10.0–14.0, 14.0–19.0, and 19.0–24.0.

For each 3-point pharmacophore, all the combinations of three interatomic distances are determined, considering only those combinations valid that satisfy the triangle rule, that is, the length of each side of a triangle cannot exceed the sum of the lengths of the other two sides, otherwise this would produce a geometrically impossible object. Moreover, redundant pharmacophores related to symmetry are usually eliminated. The resulting number of descriptors is 10 549.

PDT fingerprints (or **Pharmacophore Definition Triplets fingerprints**) are 3-point pharmacophore vectorial descriptors, where each bit is set to one or zero, depending on the presence or absence of a specific combination of three pharmacophoric points and three geometric distances [Matter and Pötter, 1999].

Five pharmacophore point types are used to generate PDT fingerprints: hydrogen-bond acceptor atom, hydrogen-bond donor atom, acceptor site, donor site, and hydrophobic center. While donor and acceptor atoms are part of the molecule, site points refer to interaction points located on a “virtual” receptor defined by geometrical criteria [Martin, Bures *et al.*, 1993]. Interfeature distances from 2.5 to 15.0 Å are divided into 27 distance bins of equal width (i.e., 0.5 Å), leading to a final string of 307 020 bits. Pharmacophore geometries from acceptable conformers are combined into a union fingerprint.

Similog keys are $\rightarrow$ *holographic vectors* containing the occurrence numbers of 3-point pharmacophores defined in terms of combinations of four features: hydrogen-bond donor (D), hydrogen-bond acceptor (A), bulkiness (B), and electropositivity (E) [Schuffenhauer, Floer-sheim *et al.*, 2003; Schuffenhauer, Brown *et al.*, 2006]. For each triplet of features, all possible combinations of three topological distances are calculated. The topological distances are mapped to four ranges: (2–3), (4–5), (6–7), and (≥ 8). To determine the pharmacophoric properties of atoms, the Sybyl atom types are applied to the neutral, uncharged molecules (Table S22).

Table S22 Atom-typing scheme applied to generate Similog keys.

Sybyl type	D	A	R^{vdw}	χ^{PA}	Sybyl type	D	A	R^{vdw}	χ^{PA}
H	–	–	1.08	2.1	S.o	–	–	1.7	2.5
Li	–	–	0.6	1.0	S.o2	–	–	1.7	2.5
B.2	–	–	1.60	2.5	Cl	–	–	1.65	3.0
B.3	–	–	1.60	2.5	K	–	–	1.33	0.8
C.3	–	–	1.52	2.5	Ca	–	–	0.99	1.0
C.2	–	–	1.53	2.5	Fe.3	–	–	2.00	1.0
C.1	–	–	1.54	2.5	Fe.2	–	–	2.00	1.0
C.ar	–	–	1.53	2.5	Co.3	–	–	2.00	1.0
N.3	+	+	1.45	3.0	Co.2	–	–	2.00	1.0
N.2	+	+	1.48	3.0	Zn.2	–	–	2.00	1.0
N.1	–	+	1.5	3.0	Zn.1	–	–	2.00	1.0

(Continued)

Table S22 (Continued)

Sybyl type	D	A	R^{vdw}	χ^{PA}	Sybyl type	D	A	R^{vdw}	χ^{PA}
N.ar	—	+	1.48	3.0	As.5	—	—	1.8	2.8
N.am	+	—	1.45	3.0	As.3	—	—	1.8	2.8
N.4	+	—	1.45	1.0	Se.3	—	—	1.9	2.5
N.2 +	+	—	1.48	1.0	Se.o	—	—	1.9	2.5
N.o	—	—	1.5	3.0	Se.o2	—	—	1.9	2.5
N.o2	+	—	1.5	3.0	Br	—	—	1.8	2.8
N.pl3	+	—	1.5	3.0	Ag.2	—	—	2.00	1.0
N.lin	—	—	1.48	1.0	Ag.1	—	—	2.00	1.0
O.3	+	+	1.36	3.5	Cd.2	—	—	2.00	1.0
O.2	—	+	1.36	3.5	Cd.1	—	—	2.00	1.0
O.2 +	—	—	1.36	1.0	Sn.2	—	—	3.00	1.0
F	—	—	1.3	4.0	Sn	—	—	3.00	1.0
Na	—	—	0.95	0.9	Te.2	—	—	2.05	2.5
Al	—	—	2.05	1.5	Te	—	—	2.05	2.5
Si	—	—	2.1	2.8	I	—	—	2.05	2.5
Si.2	—	—	2.1	2.8	Au.3	—	—	2.00	1.0
P.5	—	—	1.75	2.1	Hg.2	—	—	3.00	1.0
P.3	—	—	1.75	2.1	Hg.1	—	—	3.00	1.0
P.o2	—	—	1.75	2.1	Tl	—	—	3.00	1.0
S.3	—	—	1.7	2.5	Du	—	—	0.0	0.0
S.2	—	—	1.72	2.5	LP	—	—	0.85	0.0

Hydrogen-bond donors (D) and acceptors (A), van der Waals radii (R^{vdw}), and Pauling electronegativity (χ^{PA}) assigned to the Sybyl atom types. Hydrogen-bond donor property is assigned only if there is at least one hydrogen attached.

The **bulkiness of an atom** is calculated from the van der Waals radii R^{vdw} of the atom types; namely, the i th atom has the property of bulkiness if the following condition holds:

$$(R_i^{vdw})^3 + \sum_j (R_j^{vdw})^3 > 10^3 \text{ \AA}^3$$

where the summation runs over all the non-hydrogen first neighbors.

The **electropositivity of an atom** is derived from Pauling's electronegativity associated with the Sybyl atom types; the i th atom has the property of electropositivity if the following condition holds:

$$(\chi_i^{PA} \leq 2.5) \wedge (\chi_j^{PA} \leq 2.5 \quad \forall j)$$

where j refers to all the non-hydrogen first neighbors. The electropositivity is used as an estimate of the ability to undergo lipophilic interactions.

To obtain a unique key, form the six equivalent notations of a key, the lexicographically smallest one is taken in the final vectorial descriptor.

Triplets of Pharmacophoric Point descriptors (or **TOPP descriptors**) encode information on presence/absence or occurrence frequencies of 3-point pharmacophores derived from $\rightarrow$ *molecular interaction fields* (MIFs) [Sciabola, Morao *et al.*, 2007; Lamanna, Catalano *et al.*, 2007].

To calculate TOPP descriptors, the atoms of each molecule are first classified by the $\rightarrow$ GRID force field parametrization in four different categories according to their charge and hydrogen bond properties: DRY (hydrophobic), DONN (hydrogen-bond donor, HBD), ACPT (hydrogen-bond acceptor, HBA), and DNAC (both HBD and HBA).

For ordering purposes, type DRY (A) prevails over type DONN (B), which prevails over type ACPT (C), which in turn, prevails over type DNAC (D). All the possible 3-point combinations using four different atom types are encoded as follows: AAA, AAB, AAC, AAD, ABB, ABC, ABD, ACC, ACD, ADD, BBB, BBC, BBD, BCC, BCD, BDD, CCC, CCD, CDD, and DDD.

For each of these combinations, all possible combinations of three interatomic distances are calculated by using a maximum distance of 20 Å and a user-defined bin distance width, usually set to 1 Å. Only 3-point pharmacophores present in the molecule data set are stored in the final bit string. Moreover, a subset of TOPP descriptors is obtained by fractional factorial design implemented into $\rightarrow$ GOLPE.

TOPP descriptors can also be calculated by considering only 3-point pharmacophores, which include a common anchor point (an atom or functional group in the molecule). This option is particularly advantageous with data sets where some substituents are crucial for the activity.

4-Potential Pharmacophore Point keys (or 4-PPP keys) are binary string representations of molecules, which extend the concept of PPP triangle to combinations of four pharmacophore point types, that is, 4-point pharmacophores [Mason, Morize *et al.*, 1999].

4-PPP keys were designed with the aim of increasing the amount of shape information and including the ability of distinguishing chirality. Moreover, they account for conformational flexibility and can be calculated to describe the molecule relatively to “privileged substructures” of interest present in the molecule data set. “Privileged substructures” are selected substructures able to provide high-affinity ligands for more than one type of receptor or enzyme [Evans, Rittle *et al.*, 1988].

To account for the “privileged substructures,” one of the four pharmacophore points is forced to be a feature associated with the “privileged substructure.” In practice, this feature is a dummy atom defined as the centroid of the substructure. Then, **privileged pharmacophore keys** are 4-point pharmacophore keys, where only pharmacophores including the dummy atom representing the substructure of interest are stored.

Generation of 4-PPP keys implies transformation of the molecular structure into a $\rightarrow$ *reduced graph* by using dummy atoms and assignment of PPP types. The basic six PPP types are hydrogen-bond donor (D), hydrogen-bond acceptor (A), acidic center (C), that is, site of negative charge at pH 7, basic center (B), that is, site of positive charge at pH 7, hydrophobe (L), and aromatic ring centroid (R). In addition, quaternary nitrogen can be optionally assigned to the basic center type or an extra (seventh) PPP type. Dummy atoms, with associated PPP types, are used to represent hydrophobic regions, geometric ring centroids, and “privileged substructures.” Atoms can be assigned to one or more PPP types.

After PPP type assignments, all interatomic distances are calculated and represented by the bins into whose distance ranges they fall. Six distances are needed to describe each combination of four PPP types. Seven or 10 nonuniform distance ranges (Table S23) were proposed to calculate 4-PPP keys since they were found to be the best trade-off among differentiation, resolution, and manageable descriptor size (e.g., around 5.6 million pharmacophore/molecule with 7 ranges and 24.4 million with 10 ranges).

Table S23 Definitions of 7 and 10 distance ranges (Å) for 4-PPP keys.

Bin										
	0	1	2	3	4	5	6	7	8	9
7	0–2.5	2.5–4.0	4.0–6.0	6.0–9.0	9.0–13.0	13.0–18.0	>18.0	—	—	—
10	0–2.0	2.0–2.5	2.5–3.2	3.2–4.3	4.3–5.8	5.8–7.9	7.9–10.6	10.6–14.3	14.3–19.5	>19.5

Only geometrical valid distance combinations, checked by the triangle inequality rule, are stored into the final bit string where each bit represents a 4-point pharmacophore, that is, a specific combination of PPP types and geometrical distances. Other two filters, “volume check” and “accessibility check,” are applied to select potentially valid pharmacophores. A number of additional filters were also proposed to allow the removal of unwanted pharmacophores, which add noise to the final 4-PPP keys [Good, Cho *et al.*, 2004].

The 4-point pharmacophore keys can encode information either on the presence/absence of the possible pharmacophores or on their occurrence frequency. In the latter case, the occurrence frequency of each pharmacophore is normalized by the conformational ensemble count [Good, Cho *et al.*, 2004]. Moreover, to account for chirality, for all chiral 4-point pharmacophores, separate bins are set in the bit strings for the two enantiomers. Finally, the similarity/diversity measure between the two 4-PPP keys is based on the $\rightarrow$ *Tversky association coefficient*.

ChemDiverse pharmacophore descriptors can be either 3-point or 4-point pharmacophore descriptors [ChemDiverse – Chemical Design Ltd, 2008; Pickett, Mason *et al.*, 1996; Pickett, McLay *et al.*, 2000]. Seven pharmacophore types are allowed: hydrogen-bond donor, hydrogen-bond acceptor, basic center, acidic center, hydrophobe, aromatic center, and Nplus. These types are user-definable and, in particular, Nplus can be redefined to identify groups that can act both as donor and acceptors or to define a specific atom or molecular region as a reference point (i.e., privileged substructures). To generate ChemDiverse descriptors, up to 15 nonuniform distance ranges are usually covered: 1.7–3.0 in 0.1 Å increments, 3.0–7.0 in 0.5 Å increments, and 7.0–15.0 in 1 Å increments. Thus, provided 84 combinations of three PPP types and 30 distance ranges, the final ChemDiverse 3-PPP bit string consists of 848 925 valid 3-point pharmacophores, while the number increases around 350 million for 4-point pharmacophore descriptors.

The 4-point pharmacophore and 3-point pharmacophore descriptors are often generally referred to as **multiple pharmacophore descriptors**. A general strategy to describe molecules in terms of pharmacophoric features was proposed by contemporarily using all possible combinations of two, three, and four PPP types [Bradley, Beroza *et al.*, 2000]. After a full conformational analysis, the presence/absence of all possible pharmacophores in each molecule conformer are mapped to the final bit string (**pharmacophore signature**). In this strategy, the considered PPP types are hydrogen-bond acceptors and donors, hydrophobes, negative and positive charges, and aromatic centers; to reduce the total number of pharmacophores, only a few interfeature distance ranges are taken into account.

FLAP fingerprints (FLAP stands for *Fingerprints for Ligands and Proteins*) are vectorial descriptors encoding information about 3- and 4-point pharmacophores [Perruccio, Mason *et al.*, 2006; Baroni, Cruciani *et al.*, 2007]. The theory underpinning FLAP is similar to that of the $\rightarrow$ *TOPP descriptors*.

Atoms in molecules are first classified into different pharmacophoric types by the $\rightarrow$ *GRID* force field parametrization. Features considered are hydrophobicity, hydrogen-bond donor and acceptor capabilities, and charge. Then, all accessible geometries for all the combinations of three or four features are calculated and encoded in the final vector. Fingerprints can be calculated both for ligands and proteins. In small ligands, pharmacophores are defined by triplets or quartets of atoms, which have critical interactions with a receptor.

In proteins, a pharmacophore is defined as a combined set of all site points located in the receptor active site, which describe the location of the most favorable interaction between the given probe and the protein structure. Site points correspond to interaction energy points showing the best interaction energy (local minima). Thus, they define pharmacophoric features in proteins with a common frame of reference with pharmacophoric features in ligands.

MOLPRINT-3D fingerprints are binary strings encoding information on presence/absence of surface point environments, calculated in a way similar to $\rightarrow$ *atom environment descriptors* and derived from $\rightarrow$ *molecular interaction fields* [Bender, Mussa *et al.*, 2004a]. **Surface point environment descriptors** are generated for every point on the molecular surface in a three-step procedure and encode molecular interaction energies at each point of the surface and its neighboring points. First, points on a solvent-excluded molecular surface are defined. Second, interaction energies at surface points are calculated by using the different probes implemented in the program $\rightarrow$ *GRID* and selected to cover a variety of possible interactions between the ligand and receptor. Third, interaction energies at every surface point are binned according to a binning scheme (Table S24), which is calculated to give equally occurring bit frequencies in a selected database. More specifically, for each point on the molecular surface, its topologically adjacent neighbors are arranged in layers. For instance, points that are adjacent to the central point (layer 0) belong to layer 1; points that are externally adjacent to points in layer 1 belong to layer 2; and points in layer k are those that are adjacent to points in layer $k - 1$ and that have not been assigned to a layer of lower order. Bits are set in the final surface point environment descriptor if interaction energies within a bin range are present in the particular layer.

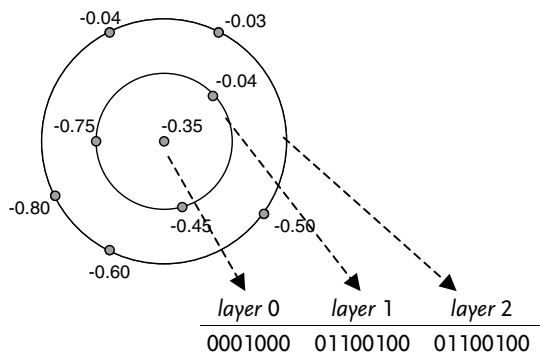
Table S24 Ranges of the different probes used in MOLPRINT 3D fingerprints.

bin	C3	DRY	N1 +	N2	O	O-
1	<-1.45	<-1.12	< 4.30	<-5.20	<-1.90	<-2.80
2	-1.45; -1.08	-1.12; -0.72	-4.30; -3.20	-5.20; -3.70	-1.90; -1.20	-2.80; -2.10
3	-1.08; -0.85	-0.72; -0.40	-3.20; -2.30	-3.70; -2.45	-1.20; -0.95	-2.10; -1.75
4	-0.85; -0.65	-0.40; -0.08	-2.30; -1.70	-2.45; -1.80	-0.95; -0.80	-1.75; -1.45
5	-0.65; -0.50	-0.08; -0.05	-1.70; -1.30	-1.80; -1.38	-0.80; -0.65	-1.45; -1.22
6	-0.50; -0.35	-0.05; -0.01	-1.30; -0.90	-1.38; -1.00	-0.65; -0.52	-1.22; -0.90
7	-0.35; 0.72	-0.01; -0.001	-0.90; -0.55	-1.00; -0.75	-0.52; -0.42	-0.90; -0.60
8	> 0.72	>-0.001	>-0.55	$\geq$ 0.75	>-0.42	>-0.60

Finally, the MOLPRINT 3D fingerprint is based on the set of all surface point environment descriptors. Usually descriptors obtained from different probes are treated independently.

Example S12

Surface point environment descriptor for the surface point having the DRY probe interaction energy of -0.35 eu (energy units).



Bin	DRY
1	≤ 1.12
2	$-1.12; -0.72$
3	$-0.72; -0.40$
4	$-0.40; -0.08$
5	$-0.08; -0.05$
6	$-0.05; -0.01$
7	$-0.01; -0.001$
8	≥ 0.001

ToPD fingerprints (or **Total Pharmacophore Diversity fingerprints**) are $\rightarrow$ *vectorial descriptors* encoding information about 3D shape and functionality of molecules [Makara, 2001]. Unlike the common pharmacophore-based descriptors that encode information on distances between a few potential pharmacophore points, ToPD fingerprints are derived from pairwise distances between a pharmacophore point type and all heavy atoms in a molecule (Figure S14). In this way, the relative position of each pharmacophore feature is mapped on the overall shape of the molecule, which is here described by the positions of all heavy atoms. Moreover, a ToPD fingerprint describing the molecular shape is derived from all pairwise distances between all heavy atoms of the molecules.

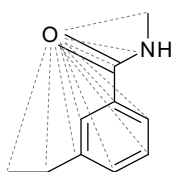


Figure S14 Pairwise geometric distances for a hydrogen-bond acceptor oxygen.

Five pharmacophore types are used: hydrophobic centers (sp^3 carbons, aromatic carbons, chlorines, bromines, iodines, sp^3 sulfurs, and sp and sp^2 carbons if not attached to a heteroatom); hydrogen-bond donors (nitrogens or oxygens bonded to a hydrogen); hydrogen-bond acceptors (sp^2 and sp^3 oxygens, sp and sp^2 nitrogens, sp^3 nitrogens if not quaternary or positively charged, sp^2 and sp^3 sulfurs, and fluorines); positively charged (quaternary nitrogens, amidines, guanidines, and carbocations); and negatively charged (oxygens of monocharged carboxylic acids, monocharged sulfuric acids, discharged phosphonic acids, monocharged nitrogen at position 2 in tetrazoles, and nitrogen in imides of $C(=O)NS(=O)(=O)$ type).

Interatomic distances can be calculated from a single conformation of a molecule or as an average distance from all possible conformations. Then, distances relative to a particular property, that is, molecular shape or a pharmacophore feature, are sorted in descending order to give a distance function from which various statistical parameters are extracted and collected in the final ToPD fingerprint. Distance function parameters are slope and intercept of the linear region, the median distance value in the linear region, the slope and the intercept of the logarithm function of the nonlinear region, the distance value at the end of the nonlinear region, and the number of distances greater than 3 Å.

There are a number of approaches for substructure searching, which are based on a matrix representation of molecules and pharmacophores used in place of the most common bit string. For instance, the $\rightarrow$ *Electronic-Conformational method* was proposed for pharmacophore identification and quantitative bioactivity prediction in drug design and toxicology by using the so-called matrices of congruity. By comparison of matrices of congruity of training set molecules, the EC submatrix of activity is derived that represents the pharmacophore.

Another approach is based on the **Compressed Feature Matrix (CFM)**, which is either a topological distance or geometrical distance matrix, whose elements represent distance relationships between atoms or pharmacophore types [Badreddin Abolmaali, Wegner *et al.*, 2003].

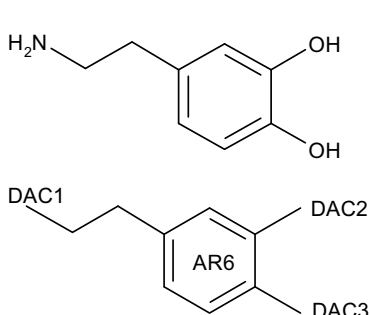
Unlike other molecular matrices, CFM is not restricted to the representation of atoms but may be built on the basis of any user-defined set of structural features.

Three basic sets of features were proposed for CFM: (1) chemical elements without information about bond types, (2) chemical elements with information about bond types, and (3) 15 pharmacophore characteristics (terminal carbon atom, hydrogen-bond donor, hydrogen-bond acceptor, atoms that may occur as either hydrogen-bond donor or acceptor, positive and negative sites, aromatic rings of size 3, 5, and 6, and nonaromatic rings from size 3 to 8).

To perform a substructure search, both the query substructure and the molecules are represented by their Compressed Feature Matrices based on the same set of structural features.

Example S13

Compressed feature matrix (CFM) for dopamine.



CFM =

	DAC ₁	DAC ₂	DAC ₃	AR6
DAC ₁	0	6	7	3
DAC ₂	6	0	3	1
DAC ₃	7	3	0	1
AR6	3	1	1	0

📖 [Sheridan and Venkataraghavan, 1987; Martin, Danaher *et al.*, 1988; Sheridan, Nilakantan *et al.*, 1989; Hicks and Jochum, 1990; Randić, 1992d; Tratch, Lomova *et al.*, 1992; Attias and Petitjean, 1993; Zhou, Xie *et al.*, 1993; Jackel and Nendza, 1994; Merschmundermann,

Rosenkranz *et al.*, 1994; Takihi, Rosenkranz *et al.*, 1994; Jordan, Leach *et al.*, 1995; Brown, 1997; Matter, 1997; Xiao, Qiao *et al.*, 1997; Brown and Martin, 1998; Flower, 1998; Molchanova and Zefirov, 1998; Matter and Pötter, 1999; Cho, Shen *et al.*, 2000; Varmuza and Scsibraný, 2000]

- **substructure list representation** → molecular descriptors
- **substructure searching** → substructure descriptors
- **sum–delta connectivity indices** → Zagreb indices
- **sum of bond length** $\equiv$ *contour length* → size descriptors ($\odot$ Kuhn length)
- **sum layer matrix** → layer matrices
- **sum matrices** → matrices of molecules
- **sum of matrices** → algebraic operators
- **sum of steric substituent constants** → steric descriptors ($\odot$ Taft steric constant)
- **superadjacency index** → eccentricity-based Madan indices ($\odot$ Table E1)
- **superadjacency topochemical indices** → eccentricity-based Madan indices ($\odot$ Table E1)
- **supercode** → superindices
- **superdelocalizability** → quantum-chemical descriptors
- **superdelocalizability indices** → quantum-chemical descriptors

■ superindices

Superindices were proposed as ordered sequences of → *molecular descriptors* providing different chemical information. To obtain indices with higher discrimination power among isomers and molecular structures, different superindices were proposed [Balaban, 1979].

The most popular superindices are → *uniform-length descriptors* of → *topological information indices* such as [Bonchev, Mekenyan *et al.*, 1981c]

$$SI = \{ {}^V\bar{I}_D^M; {}^E\bar{I}_X^E; \bar{I}_Z; {}^V\bar{I}_{C,R}; \bar{I}_{ORB}; {}^E\bar{I}_{CHR} \}$$

where the sequence terms are → *mean information content on the distance magnitude*, → *mean information content on the edge equality*, → *Hosoya mean information index*, → *radial centric information index*, → *vertex orbital information content*, and → *edge chromatic information index*, respectively.

To preserve the information contained in the element partition of a graph, a **supercode** is obtained by replacing each information index of the superindex by the cardinalities of the equivalence classes used to calculate it.

Simple sums of superindex elements can also be considered superindices belonging to the class of → *combined descriptors*; some of them were proposed as measures of → *molecular complexity*.

📖 [Motoc and Balaban, 1981; Bonchev, 1983]

■ superpendentic index

A molecular descriptor derived from the → *H-depleted molecular graph* and proposed to enhance the role of terminal vertices in QSAR and QSPR studies [Gupta, Singh *et al.*, 1999; Bakken and Jurs, 1999a]. It is calculated from the **pendent matrix**, which is a submatrix of the → *distance matrix* **D** with A rows and a number *m* of columns corresponding to the number of terminal vertices. The superpendentic index is calculated as the square root of the sum of the products of

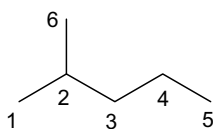
the nonzero row elements (the topological distances d) in the pendent matrix:

$$\int p = \left(\sum_{i=1}^A \prod_m d_{im} \right)^{1/2}$$

where m is over the terminal vertices, that is, the columns of the pendent matrix.

Example S14

Pendent matrix and superpendentic index for 2-methylpentane.



$$\int^P = (4 \cdot 2 + 1 \cdot 3 \cdot 1 + 2 \cdot 2 \cdot 2 + 3 \cdot 1 \cdot 3 + 4 \cdot 4 + 2 \cdot 4)^{1/2} = 7.2111$$

Pendent matrix

Atom	1	5	6
1	0	4	2
2	1	3	1
3	2	2	2
4	3	1	3
5	4	0	4
6	2	4	0

- **SURFCATS descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **surface areas** → molecular surface
- **Surface Autocorrelation Vector** → autocorrelation descriptors (⊙ Autocorrelation of Molecular Surface Properties)
- **surface electrostatic potential variance** → GIPF approach
- **surface factor** → shape descriptors (⊙ shape factor)

■ surface integral models

Surface integral models were proposed for the estimates of free energies of solvation in water, *n*-octanol, and chloroform and the enthalpy of solvation in water [Ehresmann, De Groot *et al.*, 2005]. They are based on AM1 semiempirical molecular orbital calculations leading to a parametrized nonlinear function f of four local properties calculated at the isodensity surface (the → *molecular electrostatic potential* V , local → *ionization energy* IP , local → *electron affinity* EA , and local → *polarizability* α), which is integrated over the triangulated surface area SA to obtain the target property P :

$$P = \sum_{k=1}^{N_{tri}} f(V_i, IP_i^L, EA_i^L, \alpha_i^L, \eta_i^L) \times SA_i$$

where the summation goes over N_{tri} triangles that make up the molecular surface, and the subscript i denotes the value of the relevant local property at the center of the i th surface triangle of area SA_i . The quantity η^L is the local → *hardness*, defined as

$$\eta_i^L = \frac{IP_i^L - EA_i^L}{2}$$

The function f is determined by multiple linear regression using precalculated sums of the individual components of the functions.

- **surface point environment descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **surface profiles** → molecular profiles
- **surface tension** → physico-chemical descriptors
- **surface-volume ratio** → shape descriptors (⊙ ovality index)
- **surface weighted charged partial negative surface areas** → charged partial surface area descriptors
- **surface weighted charged partial positive surface areas** → charged partial surface area descriptors
- **SURFCATS descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **susceptibility of molecular descriptors** → molecular descriptors
- **Suzuki-Kudo hydrophobic fragmental constants** → lipophilicity descriptors
- **Swain-Lupton approach** → electronic substituent constants (⊙ field/resonance effect separation)
- **Swain-Lupton field constant** $\equiv$ *field-inductive constant* → electronic substituent constants (⊙ field/resonance effect separation)
- **Swain-Lupton resonance constant** $\equiv$ *resonance constant* → electronic substituent constants (⊙ field/resonance effect separation)
- **SWIM descriptors** $\equiv$ *spectral weighted invariant molecular descriptors* → SWM signals

■ SWM signals ($\equiv$ *Spectral Weighted Molecular signals*)

Spectral weighted molecular signals are calculated within the framework on which the → *WHIM descriptors* are defined. For each molecule, the i th atom score of each m th principal axis t_{im} represents the atom projection along the axis and the weight w_{ij} of the atom is taken as the signal intensity in that position [Todeschini, Consonni *et al.*, 1999]. The weights w are atomic properties, for example, the weights used in the WHIM approach.

When more than one atom falls in the same position, that is, has the same projection, the signal intensity is the sum of the weights of the overlapping atoms having the same score. To avoid score values that differ only in numerical approximation due to the calculation procedure, such values are approximated to the first decimal digit (0.1). Therefore, a weighted spectral representation of a molecule is given by a set of ordered signals defined by score–intensity pairs.

$$\{t_{s(i),m}, w_{s(i),m}\} \quad m = 1, 2, 3$$

where m represents the m th principal axis and $s(i)$ the ordered sequence of the scores for each principal axis, from the most negative to the most positive values. Each score in the $s(i)$ position corresponds to the i th set of atoms having the same score. The complete spectral representation is given by fixing the range of t vector scores (e.g., between -20 and $+20$) and then juxtaposing the three axial spectral representations, from the first to the third principal axis.

The maximum total number of signals is $A \times 3$, where A is the number of molecule atoms; the actual total number of signals depends on the molecular symmetry (Figure S15).

Due to the properties of PCA, the projection along each principal axis is not totally invariant as only direction is uniquely defined; thus, changing the signs of the scores $t \rightarrow -t$, the reversed

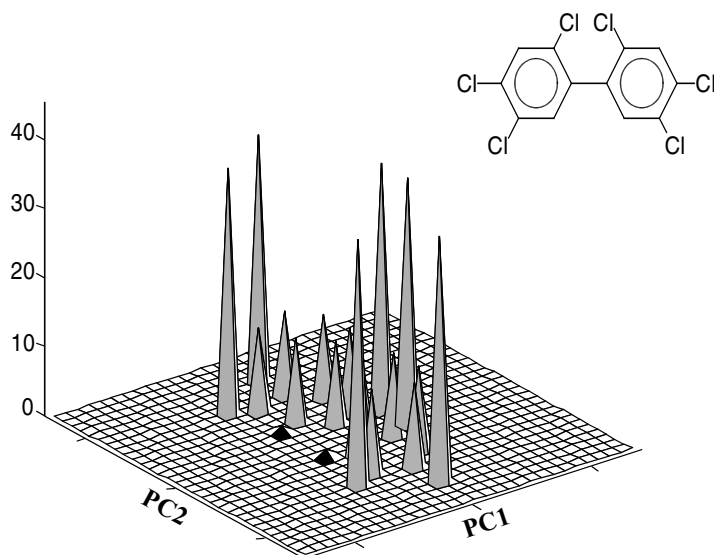


Figure S15 SWM signals of a hexachloro-biphenyl derivative.

axial spectrum is equally valid. The principal axes being three, 2^3 different spectra of a molecule can be obtained. Therefore, to compare different spectra, there are three possibilities:

- to take the two reversed spectra for each principal axis into account;
- to derive invariant molecular descriptors from their spectra; and
- to first adopt $\rightarrow$ *alignment rules*.

Procedure (a) was adopted for the similarity analysis of compounds that are each represented by a complete spectral representation.

Given two molecules *a* and *b*, each axial spectrum is compared by considering the two orientation possibilities for one of the two molecules; then, for each orientation, a rigid shift of spectral signals of one molecule with respect to the other is performed by a defined step (0.1). The maximum value of similarity found by this procedure is taken as the measure of similarity between the two molecules relative to the considered axial spectrum.

The most appropriate measure of distance was taken to be the average $\rightarrow$ *Camberra distance* $d_{ab,m}$ between the molecules *a* and *b* along the *m*th principal axis:

$$d_{ab,m} = \frac{1}{n_m} \cdot \sum_i \frac{|w_{ai,m} - w_{bi,m}|}{(w_{ai,m} + w_{bi,m})}$$

where $w_{ai,m}$ and $w_{bi,m}$ are the *i*th intensities of the molecules *a* and *b*, respectively, along the *m*th principal axis; n_m is the number of nonoverlapping signals along the *m*th axis; the sum runs over all the n_m nonoverlapping signals that are

$$n_m = n_{am} + n_{bm} - n_{m(\text{overlap})}$$

where n_{am} and n_{bm} are the number of signals along the *m*th axis for the two molecules, respectively, and $n_{m(\text{overlap})}$ the number of overlapping signals of the two molecular axial spectra, given the actual relative position due to the shift procedure.

If, in correspondence to a signal of a molecule, the other molecule does not present a signal, the denominator reaches the maximum value, that is, $d=1$. The Camberra distance is transformed into the corresponding similarity measure:

$$s_{ab,m} = 1 - d_{ab,m}$$

the average Camberra distance being normalized between zero and one.

The total similarity is calculated as the geometric mean of the axial similarities as

$$S_{ab}(w_j) = \sqrt[3]{s_{ab,1} \cdot s_{ab,2} \cdot s_{ab,3}}$$

If more than one weight is considered, the final measure of similarity is given as the geometric mean of the k total similarities obtained separately for each weight:

$$S_{ab} = \sqrt[k]{S_{ab}(w_1) \cdot \dots \cdot S_{ab}(w_k)}$$

Procedure (b) was adopted to obtain a set of molecular $\rightarrow$ autocorrelation descriptors calculated separately from each axial spectral representation. These descriptors were called **Spectral Weighted Invariant Molecular descriptors** (or **SWIM descriptors**).

As along each component the SWM signals of a molecule are naturally ordered by the scores, the autocorrelation of lag l for each m th axis AC_{lm} is calculated as

$$AC_{lm} = \sum_{i=1}^{N_m-l} w_{i,m} \cdot w_{i+l,m}$$

where N_m is the number of signals of the molecule on the m th axis; $w_{i,m}$ and $w_{i+l,m}$ are the weights of the molecule at positions i and $i+l$.

For each considered weight, a molecule is represented by a vector of autocorrelation descriptors:

$$\{AC_{01}, AC_{02}, AC_{03}; AC_{11}, \dots, AC_{L1}, AC_{L2}, AC_{L3}\}_w$$

where L is the maximum user-defined lag (e.g., $L=5$). The autocorrelations for lag zero ($l=0$) correspond to the sums of the squared weights, that is, the square intensity of each signal.

These descriptors are invariant to the direction of the axial spectral representations and constitute $\rightarrow$ *vectorial descriptors*.

■ symmetry descriptors

Symmetry is an important but often elusive molecular property when numerical values are to be assigned. However, symmetry plays an important role in the quantum-mechanical interpretation of atomic and molecular states, NMR spectrum, and several physico-chemical properties. Symmetry is closely related to those molecular properties that also depend on entropy contributions, such as melting points, vapor pressure, surface tension, and $\rightarrow$ *dipole moment*. Moreover, the nature of overall molecular shape depends on molecular symmetry.

Each molecule (or conformation) belongs to a definite point group of symmetry and each point group of symmetry includes a set of symmetry operations that are transformations leaving the whole system in a position equivalent to the initial one: identity, rotation, mirror reflection, inversion, and mirror rotation. The various groups of symmetry are

$$C_1, C_s, C_i, C_n, C_{nv}, C_{nh}, D_{nh}, T_d, \text{ etc.}$$

Some $\rightarrow$ *shape descriptors* and $\rightarrow$ *centric indices* contain information about symmetry, whereas $\rightarrow$ *WHIM symmetry*, $\rightarrow$ *Bertz complexity index*, $\rightarrow$ *indices of neighborhood symmetry*, and $\rightarrow$ *symmetry number* obtained by the $\rightarrow$ *automorphism group* of a graph are explicitly related to the symmetry.

Other specific symmetry descriptors are listed below.

- **information index on molecular symmetry (I_{SYM})**

A molecular symmetry descriptor calculated as $\rightarrow$ *total information content*:

$$I_{SYM} = A \cdot \log_2 A - \sum_{g=1}^G A_g \log_2 A_g$$

where A is the number of atoms, A_g the number of atoms belonging to the g th class of symmetry, and G the number of different classes of symmetry in the molecule [Bonchev, Kamenski *et al.*, 1976]. Each class of symmetry includes atoms able to exchange position through operations of the symmetry point group to which the molecule belongs.

Both the number of atoms and the degree of symmetry influence the increase in information content. In practice, for molecules with the same number of atoms, the information content increases with decrease in symmetry; for example, molecules with symmetry groups C_1 or C_s have the largest information content.

In general, molecular symmetry is closely related to the 3D geometry of the molecules and can be determined only from atomic coordinates; moreover, topological equivalence does not necessarily reflect all the symmetry of the 3D molecular geometry; however, it may be in some cases a useful approximation. In fact, the information index I_{SYM} , in principle based on knowledge of the molecular geometry, is related to the $\rightarrow$ *total information index on atomic composition* I_{AC} and the $\rightarrow$ *topological information content* I_{TOP} ; I_{SYM} is usually greater than these two indices. However, I_{SYM} coincides with I_{AC} when the symmetry group contains a number of symmetry operations for which all the atoms of the same chemical elements are equivalent; I_{SYM} is coincident with I_{TOP} when each different valency in the molecule is realized by atoms of only one chemical element.

I_{SYM} was used in combination with other information indices to define general measures of molecular complexity, that is, $\rightarrow$ *Dosmorov complexity index* and $\rightarrow$ *Bonchev complexity index*.

- **Kier symmetry index (${}^0\kappa$)**

Proposed as an extension of the $\rightarrow$ *Kier shape descriptors* to account for zero-order paths, that is, the atoms A , it is defined as total information content of the molecule [Kier, 1987c]:

$${}^0\kappa = -A \cdot \sum_{g=1}^G \frac{A_g}{A} \cdot \log_2 \frac{A_g}{A}$$

where A_g is the number of topologically equivalent atoms in the g th class. Each equivalence class is constituted by atoms having the same $\rightarrow$ *valence topological state* S .

It was proposed with the aim of measuring molecular symmetry in terms of atom topological uniqueness; the lower the ${}^0\kappa$ value, the greater the topological symmetry. The same index with opposite trend can also be calculated as $\rightarrow$ *redundancy index*.

• **Merrifield–Simmons index (σ)**

It is defined as the number of open sets in the topology $\mathcal{T}$ of the graph G , namely, $\sigma(G)$ [Merrifield and Simmons, 1980, 1998; Gutman, 1991b; Li, Zhao *et al.*, 2005; Zhao and Liu, 2006].

The topology $\mathcal{T}$ is the collection of all unions of *basis sets*, remembering that to each vertex v_i there can be associated a unique irreducible basis set $\mathcal{B}_i$. Thus, to produce only distinct unions, it is required that no two elements entering the union be comparable. Since comparable basis elements are those corresponding to the adjacent vertices of G , distinct open sets are guaranteed only if basis element unions, corresponding to sets of vertices of G no two of which are adjacent, are formed. In graph theory, such a set of vertices is called *stable set* of G . Thus, $\sigma(G)$ is the cardinality of a graph topology and is equal to the number of stable sets of the graph G . In other words, it is defined as

$$\sigma(G) = \sum_k q(G, k)$$

where $q(G, k)$ is the number of ways of choosing k disjoint vertices from G . It was found to be inversely correlated to the $\rightarrow$ Hosoya *Z index* [Gutman, Hosoya *et al.*, 1992; Hosoya, Gotoh *et al.*, 1999].

The method to calculate $\sigma(G)$ is to examine every subset S of $V(G)$, but this is impractical even for small graphs. In practice, $\sigma(G)$ is usually recursively calculated, considering the relation between the stable sets of G and those of G with a vertex v_i removed, that is, $G' \equiv G - v_i$. Since, each stable set of G' is also a stable set of G , the recursion formula is defined as

$$\sigma(G) = \sigma(G') + \sigma(G' - S_{\delta_i})$$

where $G' - S_{\delta_i}$ is the subgraph obtained from G by deleting the vertex v_i together with the set S_{δ_i} of all the vertices adjacent to v_i , δ_i being the vertex degree of the i th atom.

The calculation is performed considering that the recursion formula,

$$\sigma(G_{n+1}) = \sigma(G_n) + \sigma(G_{n-1})$$

holds for a linear graph of n vertices with a vertex added to one end of the graph.

This recursion formula is identical to the definition of **Fibonacci numbers**:

$$F_{n+1} = F_n + F_{n-1} \quad \text{where } F_0 = F_1 = 1$$

which are integers with a simple combinatorial meaning: F_{n+1} is the number of subsets of the set $\{1, \dots, n\}$ such that no two elements are adjacent [Gutman and El-Basil, 1986; Randić, Morales *et al.*, 1996].

Hence, for linear graphs L_n , the Merrifield–Simmons index is given by

$$\sigma(L_n) = F_{n+1}$$

Moreover, again for linear graphs (i.e., n -alkanes), the Hosoya *Z index* coincides with the Fibonacci number F_n and thus Hosoya and Merrifield–Simmons indices are both closely and directly related. For monocyclic graphs C_n and isopath graphs $i - L_n$, the following relationships between Fibonacci numbers and the Merrifield–Simmons index hold:

$$\sigma(C_n) = F_n + F_{n-2} \quad \sigma(i - L_n) = F_{n+1} + F_{n-4}$$

Table S25 collects Fibonacci numbers and the corresponding Merrifield–Simmons indices for number n of vertices between 0 and 16.

Table S25 Fibonacci numbers (F_n) and corresponding Merrifield–Simmons indices for linear (L_n), cyclic (C_n), and isographs (iL_n).

n	F_n	$\sigma(L_n)$	$\sigma(C_n)$	$\sigma(i - L_n)$
0	1	1		
1	1	2		
2	2	3	3	
3	3	5	4	
4	5	8	7	9
5	8	13	11	14
6	13	21	18	23
7	21	34	29	37
8	34	55	47	60
9	55	89	76	97
10	89	144	123	157
11	144	233	199	254
12	233	377	322	411
13	377	610	521	665
14	610	987	843	1076
15	987	1597	1364	1741
16	1597	2584	2207	2817

The Merrifield–Simmons index $\sigma(\mathcal{G})$ is a molecular descriptor quite sensitive to the molecular topology, in particular to symmetry, branching, and cyclicity; $\sigma(\mathcal{G})$ increases with branching and decreases with cyclicity.

From the graph topology, quantitative measures of the contribution of each bond in the molecule to the topological space can be obtained. Such a measure is the total number of open sets containing the bond i – j , namely n_{ij} , and the sum of their cardinalities, namely, c_{ij} . In particular, the quantity n_{ij} was found related to the $\rightarrow$ bond order by normalizing n_{ij} with respect to $\sigma(\mathcal{G})$. The **Merrifield–Simmons bond order** is defined as the best found nonlinear relationship between the bond orders and the quantity n_{ij}/σ :

$$\pi_{ij} = \exp \left[3.81 \cdot \left(\frac{n_{ij}}{\sigma} - 0.67 \right) \right]$$

• symmetry factor (σ_{SYM})

The symmetry factor was used in the expression of the entropy of acyclic saturated hydrocarbons proposed by Pitzer in terms of molecular constants by enumerating the partition functions for the molecular motions [Pitzer, 1940; Pitzer and Scott, 1941]. It is defined as

$$\sigma_{\text{SYM}} = \frac{\sigma_e \cdot \sigma_i}{I_R}$$

where σ_e is the symmetry number for external rotations, σ_i the symmetry number for internal rotations, and I_R the number of racemic isomers ($I_R = 2$ for racemic mixture, $I_R = 1$ otherwise) (Table S26).

Table S26 Values of the symmetry numbers for some compounds.

Compound	σ_e	σ_i	I_R	σ_{SYM}
<i>n</i> -Butane	2	1	1	2
2-Methyl-propane	3	1	1	3
<i>n</i> -Pentane	2	1	1	2
1-Methyl-butane	1	1	1	1
2,2-Dimethyl-propane	12	1	1	12
<i>n</i> -Hexane	2	1	1	2
2-Methyl-pentane	1	1	1	1
3-Methyl-pentane	1	1	1	1
2,3-Dimethyl-butane	2	1	1	2
2,2-Dimethyl-butane	1	3	1	3
3-Methyl-hexane	1	1	2	0.5

📖 [Kitaigorodsky, 1973; Cohen, Lee *et al.*, 1974; Shelley and Munk, 1977; Carhart, 1978; Schubert and Ugi, 1978; Randić, Brissey *et al.*, 1981; Narumi and Hosoya, 1985; Balaban, 1986b; Randić, Oakland *et al.*, 1986; Mezey, 1990a; Rücker and Rücker, 1990, 1991b; Figueras, 1992; Dannenfelser, Surendran *et al.*, 1993; Bonchev and Rouvray, 1994, 1995; Lin, 1996a; Mezey, 1997b; Caporossi and Hansen, 1998; Ivanov and Schüürmann, 1999; Hefferlin and Matus, 2001]

- **symmetry factor** → symmetry descriptors
- **symmetry number** → graph
- **Szeged difference matrix** → Szeged matrices
- **Szeged fragmental indices** → Szeged matrices
- **Szeged index** → Szeged matrices

■ Szeged matrices (SZ)

The unsymmetrical Szeged matrix **USZ** of a $\rightarrow$ *H-depleted molecular graph* G is a square unsymmetrical $A \times A$ matrix, A being the number of graph vertices, whose off-diagonal entry $i-j$ is the number of vertices $N_{i,p}$ lying closer to the focused vertex v_i [Gutman, 1994a; Dobrynin and Gutman, 1994; Dobrynin, Gutman *et al.*, 1995; Dobrynin, 1995; Gutman, Khadikar *et al.*, 1995; Gutman and Klavžar, 1995; Khadikar, Deshpande *et al.*, 1995; Diudea, Minailiuc *et al.*, 1997b; Diudea *et al.*, 1997b; Diudea, Pârv *et al.*, 1997b; Diudea, 1997b]:

$$[\mathbf{USZ}]_{ij} = N_{i,p_{ij}}$$

where

$$N_{i,p_{ij}} = |\{v | v \in V(G); d_{iv} < d_{jv}\}|$$

d being the $\rightarrow$ *topological distance* and $V(G)$ the set of graph vertices; note that the vertices equidistant to v_i and v_j are not counted and the value of $N_{i,p}$ depends on both v_i and v_j . The diagonal elements are equal to zero by definition. The Szeged matrix **USZ** is similar to the unsymmetrical $\rightarrow$ *Cluj-distance matrix* **UCJD** except for the supplementary condition in the Cluj matrix:

$$p_{iv} \cap p_{ij} = \{i\}$$

where p_{ij} is the shortest path between the considered vertices. This means that in Szeged matrix, unlike the Cluj matrix, vertices along the path p_{ij} are also counted if they are closer to vertex v_i than vertex v_j .

The square symmetric Szeged matrix $\mathbf{SZ}$, of dimension $A \times A$, is obtained from the unsymmetrical Szeged matrix $\mathbf{USZ}$ by a symmetrization procedure:

$$\mathbf{SZ} = \mathbf{USZ} \otimes \mathbf{USZ}^T \quad \text{or} \quad [\mathbf{SZ}]_{ij} = [\mathbf{USZ}]_{ij}[\mathbf{USZ}]_{ji}$$

where the symbol $\otimes$ indicates the $\rightarrow$ *Hadamard matrix product*.

There are two types of symmetric Szeged matrices. One is the **path-Szeged matrix**, denoted by $\mathbf{SZ}_p$, obtained when all the off-diagonal entries are calculated as the product of the numbers $N_{i,p}$ and $N_{j,p}$ of the vertices lying closer to the vertices v_i and v_j , respectively, for all the pairs (i, j) of vertices. The path-Szeged matrix $\mathbf{SZ}_p$ is formally defined as

$$[\mathbf{SZ}_p]_{ij} = \begin{cases} N_{i,p_{ij}} \cdot N_{j,p_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where

$$N_{i,p_{ij}} = |\{v | v \in V(\mathcal{G}); d_{iv} < d_{jv}\}|$$

$$N_{j,p_{ij}} = |\{v | v \in V(\mathcal{G}); d_{jv} < d_{iv}\}|$$

Note that vertices equidistant to vertices i and j are not counted.

The **edge-Szeged matrix**, denoted as $\mathbf{SZ}_e$, is a sparse symmetric matrix whose off-diagonal elements different from zero are only those corresponding to pairs of adjacent vertices (i.e., edges):

$$[\mathbf{SZ}_e]_{ij} = \begin{cases} N_{i,e_{ij}} \cdot N_{j,e_{ij}} & \text{if } i \neq j \wedge e_{ij} \in \mathcal{E}(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases}$$

where the nonvanishing matrix elements are the product of the numbers $N_{i,e}$ and $N_{j,e}$ of vertices closer to each end of the edge e_{ij} and $\mathcal{E}(\mathcal{G})$ represents the set of edges of the graph.

The edge-Szeged matrix can be calculated from the path-Szeged matrix as

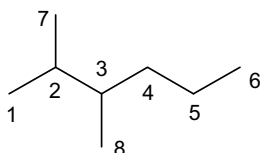
$$\mathbf{SZ}_e = \mathbf{SZ}_p \otimes \mathbf{A}$$

where the symbol $\otimes$ is the Hadamard matrix product and $\mathbf{A}$ the $\rightarrow$ *adjacency matrix*.

For any graph, both symmetric and unsymmetrical edge-Szeged matrices coincide with the symmetric and unsymmetrical edge-Cluj matrices, respectively. Moreover, the edge-Szeged matrix coincides with the edge-Wiener matrix for acyclic graphs. Therefore, as a consequence of this formal identity for acyclic graphs, the edge-Szeged matrix may be regarded as the extension of the edge-Wiener matrix to cycle-containing structures. Note that the path-Szeged matrix is different from the path-Wiener matrix for both acyclic and cycle-containing graphs.

Example S15

Unsymmetrical path-Szeged matrix and symmetrical edge- and path-Szeged matrices for 2,3-dimethylhexane. VS_i and CS_j indicate the matrix row and column sums, respectively.



	USZ_p								
	1	2	3	4	5	6	7	8	VS_i
1	0	1	1	3	3	5	1	3	17
2	7	0	3	3	5	5	7	3	33
3	5	5	0	5	5	6	5	7	38
4	5	3	3	0	6	6	5	3	31
5	3	3	2	2	0	7	3	3	23
6	3	2	2	1	1	0	3	2	14
7	1	1	1	3	3	5	0	3	17
8	5	1	1	1	5	5	5	0	23
CS_j	29	16	13	18	28	39	29	24	196

	$SZ_e = W_e$								
	1	2	3	4	5	6	7	8	VS_i
1	0	7	0	0	0	0	0	0	7
2	7	0	15	0	0	0	7	0	29
3	0	15	0	15	0	0	0	7	37
4	0	0	15	0	12	0	0	0	27
5	0	0	0	12	0	7	0	0	19
6	0	0	0	0	7	0	0	0	7
7	0	7	0	0	0	0	0	0	7
8	0	0	7	0	0	0	0	0	7
CS_j	7	29	37	27	19	7	7	7	140

Szeged index (SZ_e) = W = 70

	SZ_p								
	1	2	3	4	5	6	7	8	VS_i
1	0	7	5	15	9	15	1	15	67
2	7	0	15	9	15	10	7	3	66
3	5	15	0	15	10	12	5	7	69
4	15	9	15	0	12	6	15	3	75
5	9	15	10	12	0	7	9	15	77
6	15	10	12	6	7	0	15	10	75
7	1	7	5	15	9	15	0	15	67
8	15	3	7	3	15	10	15	0	68
CS_j	67	66	69	75	77	75	67	68	564

hyper-Szeged index (SZ_p) = 282

The **Szeged index** SZ_e was defined as an extension of the original Wiener index W to cycle-containing structures [Gutman, 1994a; Dobrynin and Gutman, 1994; Gutman and Klavžar, 1995; Khadikar, Deshpande *et al.*, 1995; Žerovnik, 1996, 1999]. While the Wiener index is the sum of the product of the number of vertices on each side of a bond, the Szeged index is defined as the sum of the product of the number of vertices closer to the atoms on each side of a bond; equidistant vertices are not counted.

The Szeged index can be calculated from the symmetric path-Szeged matrix SZ_p as the sum of the matrix entries corresponding to the pairs of adjacent vertices above the main diagonal, or by applying the $\rightarrow$ Wiener operator Wi to the symmetric edge-Szeged matrix SZ_e or from the unsymmetrical edge-Szeged matrix USZ_e applying the $\rightarrow$ orthogonal Wiener operator $Wi^\perp$:

$$SZ_e = Wi(SZ_e) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [SZ_e]_{ij} = Wi^\perp(USZ_e) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [USZ_e]_{ij} \cdot [USZ_e]_{ji}$$

For acyclic graphs, the Szeged index SZ_e is equal to the $\rightarrow$ Wiener index W ; for any graph, the Szeged index SZ_e is equal to the $\rightarrow$ Cluj-distance index CJD_e .

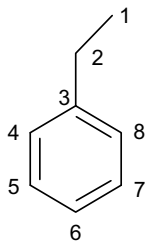
The **hyper-Szedged index** SZ_p is an extension of the Szedged index, which considers contributions from paths of any length and not only contributions from edges. It is calculated from the path-Szedged matrix as

$$SZ_p = Wi(SZ_p) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [SZ_p]_{ij} = Wi^\perp(USZ_p) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [USZ_p]_{ij} \cdot [USZ_p]_{ji}$$

For any graphs, the hyper-Szedged index SZ_p is different from the other hyper-indices.

Example S16

Unsymmetrical, symmetric edge- and path-Szedged matrices, together with Szedged index and hyper-Szedged index for ethylbenzene. VS_i and CS_j indicate the matrix row and column sums, respectively.



USZ _p									
	1	2	3	4	5	6	7	8	VS _i
1	0	1	1	2	2	3	2	2	13
2	7	0	2	2	4	3	4	2	24
3	6	6	0	5	4	5	4	5	35
4	6	3	3	0	5	4	5	2	28
5	4	4	2	3	0	5	2	3	23
6	5	3	3	2	3	0	3	2	21
7	4	4	2	3	2	5	0	3	23
8	6	3	3	2	5	4	5	0	28
CS _j	38	24	16	19	25	29	25	19	195

SZ _e = W _e									
	1	2	3	4	5	6	7	8	VS _i
1	0	7	0	0	0	0	0	0	7
2	7	0	12	0	0	0	0	0	19
3	0	12	0	15	0	0	0	15	42
4	0	0	15	0	15	0	0	0	30
5	0	0	0	15	0	15	0	0	30
6	0	0	0	0	15	0	15	0	30
7	0	0	0	0	0	15	0	15	30
8	0	0	15	0	0	0	15	0	30
CS _j	7	19	42	30	30	30	30	30	218

SZ _p									
	1	2	3	4	5	6	7	8	VS _i
1	0	7	6	12	8	15	8	12	68
2	7	0	12	6	16	9	16	6	72
3	6	12	0	15	8	15	8	15	79
4	12	6	15	0	15	8	15	4	75
5	8	16	8	15	0	15	4	15	81
6	15	9	15	8	15	0	15	8	85
7	8	16	8	15	4	15	0	15	81
8	12	6	15	4	15	8	15	0	75
CS _j	68	72	79	75	81	85	81	75	616

Szedged index (SZ_e) = W = 109

hyper-Szedged index (SZ_p) = 308

To make the Szedged index of compounds with odd member rings to be of comparable magnitude with that of compounds containing even member rings, Randić proposed to calculate the Szedged index by also taking into account contributions from atoms at the same distance from the atoms at both ends of a bond. This led to a novel molecular descriptor, which

was called **revised Wiener index** (RW), defined as the sum over all bonds of the product $N_i \times N_j$, where N_i and N_j are the number of atoms closer to atoms on each side of a bond with the additional rule that atoms at the same distance from atoms at both ends of a bond give a contribution of 0.5 [Randić, 2002b].

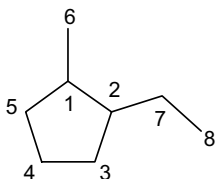
Moreover, two **tree-likeness indices** were defined as [Pisanski, personal communication]

$$\alpha = \frac{SZ}{RW} \quad \text{and} \quad \gamma = \frac{W}{RW} \quad 0 \leq \alpha, \gamma \leq 1$$

where SZ , W , and RW are the Szeged index, the $\rightarrow$ Wiener index, and the revised Wiener index, respectively. Both α and γ indices equal 1 for tree graphs, while they decrease in the presence of cycles.

Example S17

Calculation of revised Wiener index and Szeged index for 1-methyl-2-ethylcyclopentane.



Vertex pair (i, j)	N_i	N_j	$N_i \cdot N_j$
(1, 2)	3.5	4.5	15.75
(2, 3)	5.5	2.5	13.75
(3, 4)	5	3	15
(4, 5)	3.5	4.5	15.75
(1, 5)	5.5	2.5	13.75
(1, 6)	7	1	7
(2, 7)	6	2	12
(7, 8)	7	1	7

Revised Wiener index = 100

Vertex pair (i, j)	N_i	N_j	$N_i \cdot N_j$
(1, 2)	3	4	12
(2, 3)	5	2	10
(3, 4)	4	2	8
(4, 5)	2	3	6
(1, 5)	5	2	10
(1, 6)	7	1	7
(2, 7)	6	2	12
(7, 8)	7	1	7

Szeged index = 72

The tree-likeness indices are $\alpha = 0.72$ and $\gamma = 0.61$, with the Wiener index $W = 61$.

For vertex-weighted molecular graphs, a **weighted modified Wiener index** was defined as [Gutman and Žerovnik, 2002]

$${}^m W(w) = \sum_{b=1}^B (s_{i,b})^{-1} \cdot (s_{j,b})^{-1}$$

where $s_{i,b}$ is the sum of the weights of the vertices lying closer to v_i than to v_j and $s_{j,b}$ is the sum of the weights of the vertices lying closer to v_j than to v_i ; the summation goes over all the edges.

For vertex- and edge-weighted graphs, the Szeged index was defined as [Ivanciuc and Ivanciuc, 1999; Ivanciuc, 2000i]

$$SZ_e(w) = \sum_{i=1}^A w_i + \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot w_{ij} \cdot N_{i,(ij)} \cdot N_{j,(ij)}$$

where w is the $\rightarrow$ *weighting scheme* used to derive vertex parameters w_i and edge parameters w_{ij} ; a_{ij} are the elements of the adjacency matrix equal to one only for pairs of adjacent vertices; and $N_{i,(ij)}$ and $N_{j,(ij)}$ are the number of vertices closer to vertex v_i and vertex v_j , respectively.

A modified Szeged index, called **PI index (Padmakar–Ivan index)**, was defined as [Khadikar, Kale *et al.*, 2001; Khadikar and Karmarkar, 2001; Ashrafi and Manoochehrian, 2006; Klavžar, 2007]

$$PI = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot \left(N_{i,(ij)}^e + N_{j,(ij)}^e \right)$$

where the summation goes over all the pairs of vertices, but a_{ij} is equal to one only for pairs of adjacent vertices; $N_{i,(ij)}^e$ is the number of edges lying closer to the i th vertex than the j th vertex, and $N_{j,(ij)}^e$ the number of edges lying closer to the j th vertex than the i th vertex; equidistant edges are not counted.

Reciprocal Szeged matrices, denoted as SZ^{-1} , are matrices whose off-diagonal elements are the reciprocal of the corresponding elements of the Szeged matrices [Diudea, 1997a; Diudea, Pârv *et al.*, 1997b]:

$$[SZ^{-1}]_{ij} = \frac{1}{[SZ]_{ij}}$$

All elements equal to zero are left unchanged in the reciprocal Szeged matrices.

$\rightarrow$ *Harary Szeged indices* H_{SZ_e} and H_{SZ_p} are calculated from reciprocal symmetric edge- and path-Szeged matrices, respectively, by applying the Wiener operator.

The **Szeged difference matrix** SZ_{Δ} was proposed as the difference between the path-Szeged matrix and the edge-Szeged matrix [Diudea, Minailiuc *et al.*, 1997b; Ivanciuc, Ivanciuc *et al.*, 1997]:

$$SZ_{\Delta} = SZ_p - SZ_e$$

The **Szeged property matrices** SZ_{UP} are square unsymmetrical $A \times A$ weighted matrices derived from a vertex-weighted molecular graph. Each off-diagonal entry i – j is a function of a selected property of vertices $P_{i,(ij)}$ lying closer to the focused vertex v_i [Minailiuc, Katona *et al.*, 1998]:

$$[SZ_{UP}]_{ij} = P_{i,(ij)}$$

where

$$P_{i,(ij)} = f(P_v) \quad \text{where} \quad v|v \in V(G); d_{iv} < d_{jv}$$

$$f(P_v) = m \cdot \sum_v P_v \quad \text{or} \quad f(P_v) = \left(\prod_v P_v \right)^{1/N_{i,(ij)}}$$

where the properties are evaluated on all the $N_{i,(ij)}$ vertices v lying closer to the vertex i than j . $V(G)$ is the set of graph vertices and d is the topological distance. The two property functions are an additive function of the vertex properties and the geometric mean; the term m in the former is a weighting factor. $P_{i,(ij)}$ is considered a fragmental property since it collects the properties of the molecular fragment linked to the considered vertex.

The Szeged property matrix based on a unitary vertex property and the weighting factor m equal to one reduces to the classic unsymmetrical Szeged matrix **USZ**. Moreover, an interesting Szeged property matrix is obtained considering $m = 1$ with a generic vertex property P_v . Thus, for example, Szeged walk-property matrices $\mathbf{SZ}_U \mathbf{W}^k$ can be calculated using an $\rightarrow$ *atomic walk count* of different orders as the vertex property, the Szeged mass-property matrix $\mathbf{SZ}_U \mathbf{A}$ is obtained by weighting vertices by atomic masses and the Szeged electronegativity-property matrix $\mathbf{SZ}_U \mathbf{X}$ by using the $\rightarrow$ *Sanderson group electronegativity* calculated for each atom.

Normalized Szeged property matrices $\mathbf{SZ}_U \mathbf{NP}$ are a class of Szeged property matrices, defined using as the weighting factor m in the additive function, the reciprocal of the global molecular property $P(\mathcal{G})$:

$$f'(P_v) = \frac{\sum_v P_v}{P(\mathcal{G})}$$

$\rightarrow$ *Wiener-type indices*, called **Szeged fragmental indices**, are derived from Szeged property matrices applying the $\rightarrow$ *orthogonal Wiener operator* $Wi^\perp$ to both path- and edge-Szeged matrices:

$$\mathbf{SZP}_p \equiv Wi^\perp(\mathbf{SZ}_U \mathbf{P}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{SZ}_U \mathbf{P}_p]_{ij} \cdot [\mathbf{SZ}_U \mathbf{P}_p]_{ji}$$

and

$$\mathbf{SZP}_e \equiv Wi^\perp(\mathbf{SZ}_U \mathbf{P} \otimes \mathbf{A}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{SZ}_U \mathbf{P}_e]_{ij} \cdot [\mathbf{SZ}_U \mathbf{P}_e]_{ji}$$

where the indices $\mathbf{SZP}_e$ are derived from the unsymmetrical edge-Szeged matrices $\mathbf{USZP}_e$ calculated by the Hadamard matrix product between the path-Szeged property matrices $\mathbf{USZP}_p$ and the adjacency matrix $\mathbf{A}$.

📖 [Dobrynin and Gutman, 1996; Klavžar, Rajapaxi *et al.*, 1996; Chepoi and Klavžar, 1997; Das, Dömötör *et al.*, 1997; Diudea and Gutman, 1998; Gutman and Dobrynin, 1998; Mandloi, Sikarwar *et al.*, 2000; Ivanciuc, 2000b, 2000d, 2002b, 2000f; Jäntschi, Katona *et al.*, 2000; Agrawal and Khadikar, 2001; Ardelan, Katona *et al.*, 2001; Khadikar, Karmarkar *et al.*, 2002; Khadikar, Phadnis *et al.*, 2002; Khadikar, Agrawal *et al.*, 2002; Khadikar, Singh *et al.*, 2002, 2003; Randić, 2002b; Jaiswal and Khadikar, 2004; Khadikar, Joshi *et al.*, 2004; Khadikar, Sharma *et al.*, 2005a; Khadikar, Diudea *et al.*, 2006]

➤ **Szeged property matrices** $\rightarrow$ Szeged matrices.

T

■ TAE descriptor methodology ($\equiv$ Transferable Atom Equivalent descriptor methodology)

Based on the $\rightarrow$ AIM theory, TAE descriptor methodology encodes the distributions of electron density based on molecular properties, such as kinetic energy density, $\rightarrow$ molecular electrostatic potential, local average ionization potentials, $\rightarrow$ Fukui functions, electron density gradients, and second derivatives, in addition to the density itself [Breneman, Thompson *et al.*, 1995; Song, Breneman *et al.*, 2002; Breneman, Sundling *et al.*, 2003].

Namely, a TAE is an atomic electron density fragment bounded by interatomic zero-flux surfaces and an extended isodensity surface approximating the condensed-phase van der Waals surface. TAE fragments are composed of atomic charge density-derived properties (Table T1) that were precalculated from small molecules using *ab-initio* level approaches. These atomic fragments constitute the TAE library structured in a form that allows the atomic fragments of a new molecule to be rapidly retrieved. The RECON (RECONstruction) program reads atomic connectivities of the molecule and assigns the closest fragment match from the TAE library to each atom based on the atom type, hybridization, and structural environment.

Together with the surface electronic properties listed in Table T1, four integrated electronic properties are also included in the set of TAE descriptors: total energy, electron population, volume, and surface.

TAE histogram descriptors are produced by recording the distribution of these properties as surface histograms that quantify the molecular surface areas with specific value ranges of each property. In addition to histogram descriptors, property extrema, average values, and standard deviations of property distributions (in some cases, with separate σ values for positive and negative portions of the range) are included in the TAE descriptor set.

Table T1 TAE descriptors.

Symbol	Surface electronic property	Formula
EP	Electrostatic potential	$EP(r) \equiv v(r) = \sum_{a=1}^A \frac{Z_a}{ \mathbf{r}-\mathbf{R}_a } - \int \frac{\rho(\mathbf{r}')}{ \mathbf{r}-\mathbf{r}' } \cdot d\mathbf{r}'$
DRN	Electron density gradient normal to 0.002 e/au ³ electron density isosurface	$\nabla \rho \cdot \mathbf{n}$
G	Electronic kinetic energy density	$G(r) = -\frac{1}{2} \cdot (\nabla \psi^* \cdot \nabla \psi)$
K	Electronic kinetic energy density	$K(r) = -\frac{1}{2} \cdot (\psi^* \nabla^2 \psi + \psi \nabla^2 \psi^*)$

(Continued)

Table T1 (Continued)

Symbol	Surface electronic property	Formula
DGN	Gradient of the G electronic energy density normal to surface	$\nabla G \cdot \mathbf{n}$
DKN	Gradient of the K electronic energy density normal to surface	$\nabla K \cdot \mathbf{n}$
F	Fukui f^+ function scalar value	$f^+(r) = \rho_{\text{HOMO}}(r)$
L	Laplacian of the electronic density	$L(r) = -\nabla^2 \rho(r) = K(r) - G(r)$
BNP	Bare nuclear potential	$BNP(\mathbf{r}) \equiv v(\mathbf{r}) = \sum_{a=1}^A \frac{Z_a}{ \mathbf{r} - \mathbf{R}_a }$
PIP	Local average ionization potential	$PIP(\mathbf{r}) \equiv \bar{I}(\mathbf{r}) = \frac{\sum_i \rho_i(\mathbf{r}) \cdot \epsilon_i}{\rho(\mathbf{r})}$

$\rho(r)$ represents the electron density distribution.

PEST descriptors ($\equiv$ **Property-Encoded Surface Translator descriptors**) are hybrid shape/property descriptors encoding information about molecular shape, without any alignment procedure [Breneman, Sundling *et al.*, 2003]. Each TAE property encoded surface is subjected to Zauhar “Shape Signature” ray tracing approach [Zauhar, Moyna *et al.*, 2003] to generate PEST descriptors. A ray is initialized with a random location and direction within the molecular surface and reflected throughout inside the electron density isosurface until the molecular surface is adequately sampled. Molecular shape information is obtained by recording the ray path information including segment lengths, reflection angles, and property values at each point of incidence. Path information (segment length and point of incidence values) can be summarized into 2D histograms to obtain a surface shape profile. For a single electronic property, a 2D histogram having the distribution of distance (x-axis) versus the associated property value (y-axis) can give a characteristic distribution (z-axis) based on the overall shape and property value of the molecule. Such a 2D histogram can be created for every surface property. Each bin of the 2D histogram becomes a hybrid shape/property descriptor.

Surface property distribution can also be characterized by the use of discrete $\rightarrow$ *wavelet transform*. Then, **wavelet coefficient descriptors (WCD)**, produced through TAE reconstructions, are another set of descriptors encoding information about the 10 surface electronic properties defined in Table T1, each property being represented by a 1024-point distribution, retaining only 32 wavelet coefficients. Property distribution reconstructed from these 32 wavelet coefficients reproduces the original distribution to greater than 95% accuracy. Just for the TAE descriptors, wavelet coefficients of atomic property distributions can be simply summed, weighted by the atomic surface area, to give molecular wavelet representations [Breneman, Sundling *et al.*, 2003; Lavine, Davidson *et al.*, 2003].

PEST Autocorrelation Descriptors (or **PAD descriptors**) are spatial $\rightarrow$ *autocorrelation descriptors* defined on the basis of TAE and PEST descriptors [Breneman, Sundling *et al.*, 2003]. For each ray in PEST, the length of the ray and the product of the property values at starting and ending points are computed. The distribution is binned into 20 bins along the ray length and the autocorrelation values for each bin calculated. For 10 TAE properties, this yields a total of 200 PEST autocorrelation descriptors.

 [Tugcu, Ladiwala *et al.*, 2003; Oloff, Zhang *et al.*, 2006]

- **TAE histogram descriptors** → TAE descriptor methodology
- **Taft–Kutter–Hansch steric constants** → steric descriptors (⊙ Taft steric constant)
- **Taft–Lewis inductive constant** → electronic substituent constants (⊙ inductive electronic constants)
- **Taft polar constant** $\equiv$ Taft σ^* constant → electronic substituent constants (⊙ inductive electronic constants)
- **Taft resonance constants** → electronic substituent constants (⊙ resonance electronic constants)
- **Taft steric constant** → steric descriptors
- **Taft σ^* constant** → electronic substituent constants (⊙ inductive electronic constants)
- **TAGC-axis system** → biodescriptors (⊙ DNA sequences)

■ Taillander index

An empirical steric index, denoted as ΣD , for the substituted benzene rings and defined as the 6-term sum of the distances, given by the L Sterimol length parameters, between the six atoms bonded to the benzene carbon atoms, that is, the value of the external perimeter of the benzene ring [Taillander, Domard *et al.*, 1983; Ravanel, Taillander *et al.*, 1985]. It represents the perimeter of the efficacious section and describes the steric properties of aromatic compounds.

For example, for the unsubstituted benzene ring, the Taillander index is the sum of the six H...H distances; chlorobenzene perimeter is shown in Figure T1.

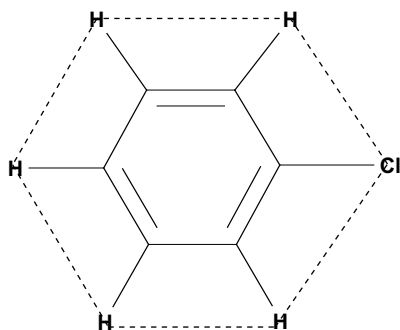


Figure T1 Chlorobenzene perimeter for the calculation of the Taillander index.

📖 [Argese, Bettiol *et al.*, 1999]

- **Tanimoto coefficient** → quantum similarity
- **Tanimoto distance** → similarity/diversity (⊙ Table S10)
- **Tanimoto similarity coefficients** → similarity/diversity (⊙ Table S9)
- **Tarko–Ivanciuc fitness function** → regression parameters
- **TAU indices** → ETA indices
- **TCAG-axis system** → biodescriptors (⊙ DNA sequences)
- **TDB-atom type descriptors** → autocorrelation descriptors (⊙ 3D-topological distance based descriptors)

- **TDB-electronic descriptors** → autocorrelation descriptors (⊙ 3D-topological distance based descriptors)
- **TDB-steric descriptors** → autocorrelation descriptors (⊙ 3D-topological distance based descriptors)

■ technological properties

Technological properties are properties of matters of technological importance [Katritzky, Maran *et al.*, 2000; Katritzky and Fara, 2005]. Together with all the → *physico-chemical properties* (such as, for example, critical temperature, vapor pressure, flash point, surface tension, and density) that are able to characterize any material constituted by pure species, other technological properties are those describing more specifically characteristics of materials, such as polymers, oil mixtures, and surfactants.

An important technological property of → *polymers* is the **glass transition temperature** (T_g), that is the temperature at which an amorphous polymer is transformed, in a reversible way, from a viscous or rubbery condition to a hard and relatively brittle one; in the vicinity of T_g , a polymer experiences a sudden increase in the rate of molecular motions and, as a result, undergoes a series of conformational transformations. Another important polymer technological property is the → *refractive index* of polymers, whose high values are usually related to highly conjugated, aromatic type, π -electron systems that bear heavy elements such as bromine or iodine.

An important technological property concerning oil mixtures is the **motor octane number** (MON), that is a measure of the autoignition resistance of gasoline and is defined as the number that gives the percentage, by volume, of iso-octane in a mixture of iso-octane and normal heptane, that would have the same antiknocking (or antidetonation) capacity as the fuel under consideration. For example, gasoline with the same detonation characteristics as a mixture of 95% iso-octane and 5% heptane would have an octane number of 95.

Surfactants represent a class of materials including wetting agents that lower the surface tension of a liquid, allowing easier spreading, and lower the interfacial tension between two liquids. A common property used to describe surfactants is the **critical micelle concentration** (CMC), which measures the ability of dissolved surfactants to reduce surface or interfacial tension [Katritzky, Maran *et al.*, 2000]. A low CMC indicates that the surfactant is thermodynamically favorable for the hydrophobic domain of the surfactant molecule to leave the aqueous solution, and this results in both an excess concentration at the interface and the formation of micelles. The ability to absorb at an interface and reduce interfacial tension is of great importance to many processes of technological interest, such as emulsification, foaming, wetting, solubilization, detergency, particle suspensions, and surface coatings.

Another property of surfactants is the **cloud point** of non-ionic surfactants. Below this temperature a single phase of molecular or micellar solution exists; above it, the surfactant has reduced water solubility, and a cloudy dispersion results [Bünz, Braun *et al.*, 1998; Katritzky, Maran *et al.*, 2000].

Some QSPR studies on the motor octane number are found in Refs. [Balaban and Motoc, 1979; Oberrrauch and Mazzanti, 1990; Pal, Purkayastha *et al.*, 1992; Balaban, Kier *et al.*, 1992b; Kirby, 1994; Pogliani, 1999a, 2000b; Estrada and Gutierrez, 2001; Randić and Pompe, 2001b; Hosoya, 2002; Pogliani, 2006b]. QSPR studies on the glass transition temperature are in Refs. [Joyce, Osguthorpe *et al.*, 1995; Katritzky, Sild *et al.*, 1998c; Carro, Campisi *et al.*, 2002; Kim, Kim *et al.*, 2002; Camacho-Zuñiga and Ruiz-Treviño, 2003; Yin, Shuai *et al.*, 2003; Afantitis, Melagraki *et al.*, 2005]. QSPR studies on CMC can be found in Refs. [Absalan, Hemmateenejad *et al.*, 2004; Jalali-Heravi and Konouz, 2005; Katritzky and Fara, 2005; Katritzky, Pacureanu *et al.*, 2007].

- **TEI** $\equiv$ *Topological Electronegativity Index* $\rightarrow$ spectral indices
- **terminal edges** $\rightarrow$ graph
- **terminal vertices** $\rightarrow$ graph
- **Testa lipophobic constant** $\equiv$ *interactive polar parameter* $\rightarrow$ lipophilicity descriptors
- **test set** $\rightarrow$ data set
- **TGD fingerprints** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **TGT fingerprints** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **Theil's index** $\rightarrow$ statistical indices ($\odot$ concentration indices)

■ Theoretical Linear Solvation Energy Relationships (TLSEs)

QSAR method based on the philosophy of the $\rightarrow$ *Linear Solvation Energy Relationships* whose empirically derived molecular descriptors are substituted by descriptors defined in the framework of $\rightarrow$ *computational chemistry* [Famini, Kassel *et al.*, 1991; Famini, Ashman *et al.*, 1992; Famini and Wilson, 1994a]. The TLSE descriptors were developed with the aim of optimally correlating with LSER descriptors and hence being as generally applicable to solute/solvent interactions as are the LSER descriptors.

The general TLSE equation for a given physico-chemical property ϕ is expressed as

$$\log \cdot \phi = b_0 + b_1 \cdot V^{vdw} + b_2 \cdot \pi_1 + b_3 \cdot \epsilon_A + b_4 \cdot q^+ + b_5 \cdot \epsilon_B + b_6 \cdot q^-$$

where b_0 to b_6 are regression coefficients estimated by multivariate regression analysis. V^{vdw} is the $\rightarrow$ *van der Waals volume*, in units of cubic Å, representing the $\rightarrow$ *cavity term*. π_1 is the **TLSE polarizability index** chosen to represent the $\rightarrow$ *dipolarity/polarizability term* and defined by the $\rightarrow$ *polarizability volume* divided by the molecular volume to get a size-independent parameter; it indicates the ease with which the electron cloud may be moved or polarized. ϵ_B and ϵ_A are, respectively, the **Covalent Hydrogen-Bond Basicity** (CHBB) and **Covalent Hydrogen-Bond Acidity** (CHBA), defined as

$$\epsilon_B = 0.30 - \frac{|\epsilon_{\text{HOMO}}(\text{solute}) - \epsilon_{\text{LUMO}}(\text{H}_2\text{O})|}{100} \quad \text{and} \quad \epsilon_A = 0.30 - \frac{|\epsilon_{\text{LUMO}}(\text{solute}) - \epsilon_{\text{HOMO}}(\text{H}_2\text{O})|}{100}$$

where ϵ_{LUMO} and ϵ_{HOMO} are the $\rightarrow$ *lowest unoccupied molecular orbital energy* and the $\rightarrow$ *highest occupied molecular orbital energy*, respectively.

q^- is the **Electrostatic Hydrogen-Bond Basicity** (EHBB) defined as the magnitude of the charge on the most negatively charged solute atom, and q^+ is the **Electrostatic Hydrogen-Bond Acidity** (EHBA) defined as the charge on the most positively charged solute hydrogen.

Note that the LSER hydrogen-bond parameters are splitted into covalent and electrostatic hydrogen-bond contributions.

The TLSE descriptors have been used to estimate a range of properties including retention indices [Donovan and Famini, 1996], gas phase acidity [Famini, Marquez *et al.*, 1993], toxicological indices and biological activities [Wilson and Famini, 1991; Sixt, Altschuh *et al.*, 1995; Famini, Loumbev *et al.*, 1998].

📖 [Famini and Penski, 1992; Cramer, Famini *et al.*, 1993; Famini and Wilson, 1993, 1994b; Headley, Starnes *et al.*, 1994; Lowrey, Cramer *et al.*, 1995; Lowrey and Famini, 1995; Chester, Famini *et al.*, 1996]

- **theoretical molecular descriptors** → molecular descriptors
- **therapeutic index** → biological activity indices (⊙ pharmacological indices)
- **therapeutic ratio** $\equiv$ *therapeutic index* → biological activity indices (⊙ pharmacological indices)
- **three-degree cyclic atom count** → ring descriptors
- **Three-Dimensional Holographic Vector of Atom Interacting Field descriptors** $\equiv$ *3D-HoVAIF descriptors* → 3D-VAIF descriptors
- **Three-Dimensional Vector of Atomic Interaction Field descriptors** $\equiv$ *3D-VAIF descriptors*
- **Threshold Toxicological Concern** → property filters (⊙ functional group filters)
- **TLSER polarizability index** → Theoretical Linear Solvation Energy Relationships
- **TMACC descriptors** → autocorrelation descriptors

■ TOMOCOMD descriptors

These are molecular descriptors calculated by the TOMOCOMD (*TOpological MOlecular COMputer Design*) program that was designed for molecular design and bioinformatic research. It includes four modules: CARDD (*Computed-Aided Rational Drug Design*), CAMPS (*Computed-Aided Modeling in Protein Science*), CANAR (*Computed-Aided Nucleic Acid Research*), and CABPD (*Computed-Aided Bio-Polymers Docking*).

TOMOCOMD descriptors are both topological and geometrical descriptors derived from the application of discrete mathematics and linear algebra theory to chemistry [Marrero-Ponce, 2003, 2004a; Marrero-Ponce, Cabrera Pérez *et al.*, 2003; Marrero-Ponce, Castillo-Garit *et al.*, 2004a; Marrero-Ponce, Marrero *et al.*, 2004]. Geometrical descriptors account for a trigonometric chirality correction factor to distinguish between enantiomers; moreover, specific macromolecular descriptors were developed for the structural characterization of proteins.

Topological descriptors calculated by TOMOCOMD are divided into **linear indices**, **bilinear indices**, and **quadratic indices**. They are distinguished into *atom-based indices* and *bond-based indices* [Marrero-Ponce, 2004b; Marrero-Ponce, Torrens *et al.*, 2006] depending on whether they are derived from the k th power of the → *adjacency matrix of a general graph* ${}^g\mathbf{A}^k$ or from the k th power of the → *edge adjacency matrix* $\mathbf{E}^k$.

The adjacency matrix of the general graph ${}^g\mathbf{A}$ is defined as

$$[{}^g\mathbf{A}]_{ij} = \begin{cases} m_{ij} & \text{if } (i,j) \in \mathcal{E}(G) \\ 1 & \text{if } i = j \wedge i \in \text{loop} \\ 0 & \text{otherwise} \end{cases}$$

where m_{ij} is the → *bond multiplicity* of the bond formed by the vertices v_i and v_j , $\mathcal{E}(G)$ is the set of the edges in the general graph, and the diagonal elements of the matrix account for the presence of loops on graph vertices. Loops are used to characterize atoms in aromatic rings having more than one canonical structure.

The edge adjacency matrix $\mathbf{E}$ is defined as

$$[\mathbf{E}]_{ij} = \begin{cases} 1 & \text{if } e_i \wedge e_j \text{ are adjacent bonds} \\ 0 & \text{otherwise} \end{cases}$$

where e_i and e_j are two edges in the molecular graph.

TOMOCOMD indices are further distinguished into *nonstochastic indices* and *stochastic indices* [Marrero-Ponce, Marrero *et al.*, 2006], depending on whether the graph-theoretical matrix they are derived from is or is not a → *stochastic matrix*. Namely, the k th order **stochastic adjacency**

matrix of a general graph ${}^g\mathbf{A}\mathbf{S}^k$ is a square matrix of dimension $(A \times A)$, A being the number of vertices in the molecular graph, defined as [Marrero-Ponce, Montero-Torres *et al.*, 2005]

$$[{}^g\mathbf{A}\mathbf{S}^k]_{ij} = \frac{[{}^g\mathbf{A}^k]_{ij}}{awc_i^{(k)}} \quad awc_i^{(k)} \equiv VS_i({}^g\mathbf{A}^k) = \sum_{j=1}^A [{}^g\mathbf{A}^k]_{ij}$$

where ${}^g\mathbf{A}^k$ is the k th power of the adjacency matrix of a general graph and $awc_i^{(k)}$ is the k th order $\rightarrow$ *atomic walk count* for the i th atom; VS is the $\rightarrow$ *vertex sum operator*. The i - j element $[{}^g\mathbf{A}^k]_{ij}$ of the k th power of the adjacency matrix of the general graph is the number of walks of length k between vertices v_i and v_j , which also accounts for the presence of multiple bonds and loops. Therefore, the atomic walk counts derived from this matrix constitute a family of $\rightarrow$ *weighted walk degrees*. The elements $[{}^g\mathbf{A}\mathbf{S}^k]_{ij}$ of the stochastic adjacency matrix are normalized walk counts.

The **stochastic edge adjacency matrix** of k th order $\mathbf{E}\mathbf{S}^k$ is a square matrix of dimension $(B \times B)$, B being the number of edges in the molecular graph, defined as [Marrero-Ponce, Khan *et al.*, 2007a]

$$[\mathbf{E}\mathbf{S}^k]_{ij} = \frac{[\mathbf{E}^k]_{ij}}{{}^k\epsilon_i} \quad {}^k\epsilon_i \equiv VS_i(\mathbf{E}^k) = \sum_{j=1}^B [\mathbf{E}^k]_{ij}$$

where $\mathbf{E}^k$ is the k th power of the edge adjacency matrix $\mathbf{E}$ and ${}^k\epsilon_i$ is the k th order $\rightarrow$ *edge degree*.

Both these normalized matrices (${}^g\mathbf{A}\mathbf{S}^k$ and $\mathbf{E}\mathbf{S}^k$) are called *stochastic* because a $\rightarrow$ *stochastic matrix* is a square matrix with nonnegative values and the property that the sum of the elements in each row (or column) is equal to 1. Moreover, these matrices, unlike the matrices they are derived from, are nonsymmetric.

Given a graph-theoretical square matrix $\mathbf{M}$ of dimension $N \times N$, three families of molecular descriptors are calculated from the k th power of the matrix $\mathbf{M}$. These are

(a) the total **linear indices**:

$$f_k(w_1) = \sum_{i=1}^N \sum_{j=1}^N [\mathbf{M}^k]_{ij} \cdot w_{1j} = \mathbf{u}^T \cdot \mathbf{M}^k \cdot \mathbf{w}_1$$

(b) the total **quadratic indices**:

$$q_k(w_1) = \sum_{i=1}^N \sum_{j=1}^N [\mathbf{M}^k]_{ij} \cdot w_{1i} \cdot w_{1j} = \mathbf{w}_1^T \cdot \mathbf{M}^k \cdot \mathbf{w}_1$$

(c) the total **bilinear indices**:

$$b_k(w_1, w_2) = \sum_{i=1}^N \sum_{j=1}^N [\mathbf{M}^k]_{ij} \cdot w_{1i} \cdot w_{2j} = \mathbf{w}_1^T \cdot \mathbf{M}^k \cdot \mathbf{w}_2$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are two N -dimensional column vectors collecting the weights for graph elements (i.e., vertices or edges) according to two different $\rightarrow$ *weighting schemes*; w_{1i} is the value of the weighting scheme w_1 for the i th graph element, while w_{2j} is the value of the weighting scheme w_2 for the j th graph element; $\mathbf{u}$ is a N -dimensional unitary column vector.

The weighting scheme w encodes information about atoms and/or bonds in the molecule. The weighting scheme for vertices is based on atomic properties such as $\rightarrow$ electronegativity, atomic mass, $\rightarrow$ electronic charge index (ECI), and so on. The weighting scheme for edges can be based on quantities characterizing directly bonds, such as bond distances or bond dipoles, or quantities derived from weights of those atoms involved in each bond, such as atomic mass, atomic electronegativity, surface area contribution of polar atoms, atomic charges, and so on.

To calculate the bilinear indices, the following parameter based on atomic weights was proposed to characterize each bond:

$$w_{ij} = \frac{w_i}{\delta_i^b} + \frac{w_j}{\delta_j^b}$$

where w_i and w_j here are the weights (e.g., atomic mass) for the atoms i and j forming the considered bond and δ_i^b and δ_j^b the corresponding $\rightarrow$ bond vertex degrees, which also account for bond multiplicity.

It should be noted that, since the stochastic adjacency matrices are not symmetric, the stochastic bilinear indices ${}^S b_k(w_1, w_2)$ derived from them are nonsymmetric. This means that two different k th order stochastic bilinear indices can be obtained from the same matrix for each pair of weighting schemes, namely ${}^S b_k(w_1, w_2) \neq {}^S b_k(w_2, w_1)$.

Local linear, bilinear, and quadratic indices are analogously defined to characterize a single molecular fragment instead of the whole molecule. They are derived from the k th order local graph-theoretical matrix $\mathbf{M}_L^k$ extracted from the corresponding k th order graph-theoretical matrix $\mathbf{M}^k$ by considering only those elements (either vertices or edges) belonging to the selected molecular fragment. This k th order local graph-theoretical matrix $\mathbf{M}_L^k$ is defined as

$$[\mathbf{M}_L^k]_{ij} = \begin{cases} [\mathbf{M}^k]_{ij} & \text{if both } i \text{ and } j \text{ are in the fragment} \\ \frac{1}{2} \cdot [\mathbf{M}^k]_{ij} & \text{if either } i \text{ or } j \text{ is in the fragment, but not both} \\ 0 & \text{otherwise} \end{cases}$$

where i and j can refer to vertices or edges, depending on the kind of molecular matrix.

The following relationships hold between total indices and the corresponding local indices:

$$f_k(w_1) = \sum_{l=1}^Z f_{kL}(w_1) \quad q_k(w_1) = \sum_{l=1}^Z q_{kL}(w_1) \quad b_k(w_1, w_2) = \sum_{l=1}^Z b_{kL}(w_1, w_2)$$

where Z is the number of molecular fragments into which the molecule has been partitioned; $f_{kL}(w_1)$, $q_{kL}(w_1)$, and $b_{kL}(w_1, w_2)$, are the linear index, the quadratic index, and the bilinear index for the L th fragment, respectively. Atom (linear, quadratic and bilinear) indices are a particular set of atom-based local (linear, quadratic, and bilinear) indices obtained when the molecule is partitioned into its constituent atoms. From atom indices, the k th order atom-type index for a given weighting scheme is calculated by summing up the k th order atom indices of all the atoms of the same type in the molecule. Analogously, bond indices and bond-type indices are calculated by partitioning the molecule into its constituents bonds.

All these descriptors can be calculated either considering or not considering hydrogen atoms.

Chiral TOMOCOMD descriptors [Marrero-Ponce, 2004b; Marrero-Ponce, González Díaz *et al.*, 2004; Castillo-Garit *et al.*, 2006, 2007; Castillo-Garit, Marrero-Ponce *et al.*, 2006, 2007] are calculated by using a weighting scheme w for vertices in the molecular graph, which encodes a

trigonometric 3D-chirality correction factor c_i defined as

$$c_i = \sin \left[\frac{\pi \cdot (\omega_i + 4 \cdot \Delta)}{2} \right]$$

where the variable ω , accounting for the spatial configuration of every atom in the molecule, and Δ are two parameters defined as the following: $\omega = +1$ and Δ is an odd number if the atom has *R*- or axial configuration or *E*-isomerism, $\omega = 0$ and Δ is an even number if the atom does not have a specific spatial configuration, and $\omega = -1$ and Δ is an odd number if the atom has *S*- or equatorial configuration or *Z*-isomerism.

The chirality correction factor c_i is added the chosen atomic property w :

$$w'_i = w_i + c_i$$

Protein TOMOCOMD descriptors [Marrero-Ponce, Marrero *et al.*, 2004; Marrero-Ponce, Castillo-Garit *et al.*, 2005a] are linear, bilinear, and quadratic indices calculated from a matrix encoding adjacencies of the α -Carbon atoms in the protein. Each α -Carbon atom is described by any property of the corresponding side-chain amino acid. Then, the vector w is here comprised of numeric values that represent a certain property of all the amino acids in the protein.

→ *Amino acid descriptors* can be → *z-scores*, side-chain → *isotropic surface area (ISA)*, → *electronic charge index (ECI)*, and so on.

If the protein is comprised of n amino acids, then the vector w of weights is n -dimensional. The protein adjacency matrix has off-diagonal elements equal to 1 if there is a covalent bond between two α -carbon atoms and zero otherwise; the diagonal elements are equal to 1 if the amino acid has a hydrogen-bond interaction between its side-chain and the main chain atom.

The formulas to calculate total, local amino acid, and amino acid-type indices are those previously defined for general TOMOCOMD descriptors.

📖 [Marrero-Ponce, Castillo-Garit *et al.*, 2004b, 2005; Marrero-Ponce and Castillo-Garit, 2005; *et al.*, 2005b; Marrero-Ponce, Huesca-Guillén *et al.*, 2005; Marrero-Ponce, Medina-Marrero *et al.*, 2005; Montero-Torres, Vega *et al.*, 2005; Casañola-Martín, Khan *et al.*, 2006; Montero-Torres, García Sánchez *et al.*, 2006; Alvarez-Ginarte, Marrero-Ponce *et al.*, 2007; Casañola-Martín, Marrero-Ponce *et al.*, 2007b; Marrero-Ponce, Khan *et al.*, 2007b]

- **ToPD fingerprints** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **Topochemically Arrived Unique indices** ≡ *TAU indices*
- **topochemical indices** → graph invariants
- **topochromatic vector** → DARC/PELCO analysis
- **topoelectric indices** → topoelectric matrices

■ topoelectric matrices (TEM)

These are square symmetric matrices 3×3 calculated from the powers of the → *adjacency matrix* A and a standardized *properties matrix* defined in terms of → *atomic properties* derived from the → *ionization potential* IP and → *electron affinity* EA of the atoms, that is, the → *Mulliken electronegativity* and the equilibrium charge [Borodina, Filimonov *et al.*, 1998].

A topoelectric matrix of order m is calculated as

$$\mathbf{TEM}^m = \frac{\hat{\mathbf{P}}^T \cdot \mathbf{A}^m \cdot \hat{\mathbf{P}}}{A} \quad m = 1, 2, \dots, K$$

where A is the number of molecule atoms and K is the maximum considered power of adjacency matrix.

The **properties matrix** $\hat{\mathbf{P}}$ used to define $\mathbf{TEM}$ is an $A \times 3$ matrix, where each i th row is a vector $\langle 1; \hat{p}_i; \hat{q}_i \rangle$, $\hat{p}_i$ and $\hat{q}_i$ being the autoscaled Mulliken electronegativity and the autoscaled equilibrium charge (approximated by a quadratic function) of the i th atom, respectively:

$$\begin{aligned} \hat{p}_i &= \frac{p_i - \bar{p}}{\sigma_p} & \bar{p} &= \frac{\sum_{i=1}^A p_i}{A} & \sigma_p^2 &= \frac{\sum_{i=1}^A (p_i - \bar{p})^2}{A} \\ \hat{q}_i &= \frac{q_i - \bar{q}}{\sigma_q} & \bar{q} &= \frac{\sum_{i=1}^A q_i}{A} & \sigma_q^2 &= \frac{\sum_{i=1}^A (q_i - \bar{q})^2}{A} \end{aligned}$$

where A is the number of molecule atoms and p_i and q_i are defined as

$$p_i \equiv \chi_i^{\text{MU}} = \frac{\text{IP}_i + \text{EA}_i}{2} \quad q_i = -\frac{p_i}{\text{IP}_i - \text{EA}_i} = -\frac{1}{2} \cdot \frac{\text{IP}_i + \text{EA}_i}{\text{IP}_i - \text{EA}_i}$$

where χ^{MU} is the Mulliken electronegativity and IP and EA the ionization potential and electron affinity of the atom.

Each molecule is characterized by the ordered sequence of topoelectric matrices defined for increasing powers of the adjacency matrix:

$$\{\mathbf{TEM}^1; \mathbf{TEM}^2; \dots; \mathbf{TEM}^K\}$$

For each topoelectric matrix $\mathbf{TEM}^m$, a set of **topoelectric indices** is derived, these indices being all elements of each matrix; the first element $[\mathbf{TEM}^m]_{11}$ and the elements below the main diagonal are not considered, thus resulting in five descriptors for each matrix, $TEM_1^m, \dots, TEM_5^m$. If K matrices are calculated for each molecule, a total of $5 \times K$ topoelectric indices are used to characterize molecules.

From topoelectric indices, a measure of similarity between two compounds s and t was also proposed as

$$s_{st} = \frac{1}{1 + \frac{1}{5 \cdot K} \cdot \sum_{m=1}^K \sum_{j=1}^5 [TEM_j^m(s) - TEM_j^m(t)]^2}$$

where K is the number of considered matrices.

- **topographic distance** → molecular geometry
- **topographic distance-detour distance combined matrix** → matrices of molecules (© Table M3)

- **topographic distance/detour distance quotient matrix** → matrices of molecules (☉ Table M2)
- **topographic distance-resistance distance combined matrix** → matrices of molecules (☉ Table M3)
- **topographic distance/resistance distance quotient matrix** → matrices of molecules (☉ Table M2)
- **topographic distance-topological distance combined matrix** → molecular geometry
- **topographic distance/topological distance quotient matrix** → molecular geometry
- **topographic electronic descriptors** → charge descriptors
- **topographic indices** → molecular geometry
- **topographic matrix** → molecular geometry
- **topological atomic charge** → self-returning walk counts
- **topological atomic valencies** → self-returning walk counts
- **topological atom pairs** → substructure descriptors
- **topological binding property pairs** → substructure descriptors (☉ pharmacophore-based descriptors)
- **topological bond index** $\equiv$ *molecular path count* → path counts
- **topological bond order** → bond order indices (☉ graphical bond order)
- **topological bond orders** → self-returning walk counts

■ **topological charge indices** ($\equiv$ *charge-transfer indices*)

Topological charge indices were proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule [Gálvez, García *et al.*, 1994; Gálvez, García-Domenech *et al.*, 1995].

Let **M** be the matrix obtained by multiplying the → *adjacency matrix* **A** by → *the reciprocal square distance matrix* **D**⁻², that is,

$$\mathbf{M} = \mathbf{A} \cdot \mathbf{D}^{-2}$$

To avoid division by zero, the diagonal entries of the distance matrix remain the same; the obtained matrix **M**, called **Gálvez matrix**, is a square unsymmetric matrix $A \times A$, A being the number of atoms in the molecule.

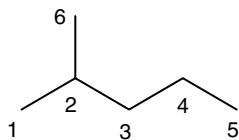
An unsymmetric **charge term matrix**, denoted as **CT**, is derived from the matrix **M**; its terms CT_{ij} for each pair of vertices v_i and v_j are defined as

$$[CT]_{ij} \equiv CT_{ij} = \begin{cases} \delta_i & \text{if } i = j \\ m_{ij} - m_{ji} & \text{if } i \neq j \end{cases}$$

where m_{ij} are the elements of the matrix **M** and δ_i is the → *vertex degree* of the i th atom. The diagonal entries of the matrix **CT** represent the topological valence of the atoms and the off-diagonal entries CT_{ij} represent a measure of the net charge transferred from the atom j to the atom i ; if CT_{ij} is negative, atom i will transfer net charge to atom j . To take into account also the heteroatoms, the diagonal entries of the adjacency matrix can be substituted by the Pauling's → *atom electronegativity* (taking as the reference value 2 for the chlorine atom) or simply the → *valence vertex degree* or any other atomic property.

Example T1

Calculation of the topological charge indices for 2-methylpentane.



Adjacency matrix A							reciprocal square distance matrix D ⁻²						
Atom	1	2	3	4	5	6	Atom	1	2	3	4	5	6
1	0	1	0	0	0	0	1	0	1	0.250	0.111	0.063	0.250
2	1	0	1	0	0	1	2	1	0	1	0.250	0.111	1
3	0	1	0	1	0	0	3	0.250	1	0	1	0.250	0.250
4	0	0	1	0	1	0	4	0.111	0.250	1	0	1	0.111
5	0	0	0	1	0	0	5	0.063	0.111	0.250	1	0	0.063
6	0	1	0	0	0	0	6	0.250	1	0.250	0.111	0.063	0

Galvex matrix M						
Atom	1	2	3	4	5	6
1	1	0	1	0.250	0.111	1
2	0.500	3	0.500	1.222	0.375	0.500
3	1.111	0.250	2	0.250	1.111	1.111
4	0.313	1.111	0.250	2	0.250	0.313
5	0.111	0.250	1	0	1	0.111
6	1	0	1	0.250	0.111	1

Charge-transfer matrix CT						
Atom	1	2	3	4	5	6
1	1	-0.500	-0.111	-0.063	0	0
2	0.500	3	0.250	0.111	0.123	0.500
3	0.111	-0.250	2	0	0.111	0.111
4	0.063	-0.111	0	2	0.250	0.063
5	0	-0.123	-0.111	-0.250	1	0
6	0	-0.500	-0.111	-0.063	0	1

$$G_1 = |CT_{12}| + |CT_{26}| + |CT_{23}| + |CT_{34}| + |CT_{45}| = 0.5 + 0.5 + 0.25 + 0 + 0.25 = 1.5$$

$$G_2 = |CT_{16}| + |CT_{13}| + |CT_{63}| + |CT_{24}| + |CT_{35}| = 0 + 0.111 + 0.111 + 0.111 + 0.111 = 0.444$$

$$G_3 = |CT_{14}| + |CT_{64}| + |CT_{25}| = 0.063 + 0.063 + 0.123 = 0.249$$

$$G_4 = |CT_{15}| + |CT_{65}| = 0$$

$$J = J_1 + J_2 + J_3 + J_4 = 1.5/5 + 0.444/5 + 0.248/5 + 0 = 0.438$$

For each path of length k , a topological charge index G_k is defined as

$$G_k = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A |CT_{ij}| \cdot \delta(d_{ij}; k)$$

where d_{ij} is the topological distance between i th and j th atoms; $\delta(d_{ij}; k)$ is a Kronecker delta function equal to 1 if $d_{ij} = k$, zero otherwise.

Therefore, G_k is the half-sum of all charge terms CT_{ij} corresponding to pair of vertices with topological distance $d_{ij} = k$ and would represent the total charge transfer between atoms placed at topological distance k . The maximum number of G_k terms is equal to the $\rightarrow$ topological diameter D . The G_1 index is closely related to the $\rightarrow$ molecular branching.

A mean topological charge index J_k is defined as

$$J_k = \frac{G_k}{A-1}$$

where the denominator $A - 1$ is the number of edges in an acyclic molecule.

Moreover, a global topological charge index J is defined as

$$J = \sum_{k=1}^5 J_k$$

where the superior limit equal to 5 was proposed by the Authors to obtain $\rightarrow$ uniform-length descriptors such as $\{G_1, G_2, \dots, G_5; J_1, J_2, \dots, J_5; J\}$. In any case, G_k (and consequently J_k and J) values are set at zero for k values greater than the diameter of the molecule.

Note. It must be observed that the diagonal elements of the charge-transfer matrix **CT**, which correspond to the vertex degrees, are not actually exploited to generate the topological charge indices.

Exploiting the charge transfer information encoded by the matrix **CT**, the algebraic semisum charge-transfer index, denoted as μ_{alg} , was proposed as a measure of the molecular $\rightarrow$ dipole moment calculated as the half-sum of the CT_{ij} elements corresponding to pairs of bonded vertices [Torrens, 2001, 2004, 2005]:

$$\mu_{alg} = \frac{1}{2} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot CT_{ij} = \frac{1}{2} \cdot \sum_b C_b$$

where a_{ij} are the elements of the $\rightarrow$ adjacency matrix that are equal to 1 for pairs of adjacent vertices, and zero otherwise; CT_{ij} are the elements of the charge term matrix; C_b is the CT_{ij} index for vertices v_i and v_j forming the bond b . Moreover, each edge dipole moment μ_b can be evaluated from the corresponding term C_b as

$$\mu_b = \frac{1}{2} \cdot C_b$$

Then, each edge dipole can be associated with a vector μ_b in space, which has magnitude μ_b and lies along the edge b with direction from vertex v_j to v_i . The molecular dipole moment vector μ is obtained from the vector sum of the edge dipole moments as

$$\mu = \sum_b \mu_b$$

The vector semisum charge-transfer index, denoted as μ_{vec} , is defined as the norm of μ :

$$\mu_{vec} = \|\mu\| = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2}$$

where μ_x , μ_y and μ_z are the three dipole components.

Valence charge-transfer indices were also defined to take into account the presence of heteroatoms in the molecule by using an $\rightarrow$ *augmented adjacency matrix* ${}^a\mathbf{A}(\chi)$ based on atomic electronegativities χ [Torrens, 2003b]:

$$[{}^a\mathbf{A}(\chi)]_{ij} = \begin{cases} a_{ij} & i \neq j \\ 2.2 \cdot (\chi_i - \chi_C) & i = j \end{cases}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* $\mathbf{A}$ and χ_C the electronegativity of the carbon atom.

From the augmented adjacency matrix ${}^a\mathbf{A}(\chi)$, the corresponding Gálvez matrix $\mathbf{M}^V$ is calculated as

$$\mathbf{M}^V = {}^a\mathbf{A}(\chi) \cdot \mathbf{D}^{-2}$$

where $\mathbf{D}^{-2}$ is the $\rightarrow$ *reciprocal square distance matrix*.

From the $\mathbf{M}^V$ matrix, by analogy with the procedure outlined for the topological charge indices, the $\mathbf{CT}^V$ matrix is calculated and the **valence topological charge-transfer index** G_k^V , the **valence mean topological charge-transfer index** J_k^V , the **valence global topological charge-transfer index** J^V , the **valence algebraic semisum charge-transfer index** μ_{alg}^V , and the **valence vector semisum charge-transfer index** μ_{vec}^V were proposed.

Moreover, to account for the significant decrease of the polarity in the hydrogen-bond formation capacity of the sp^3 -oxygen atoms that are directly linked to a sp^2 -carbon atom (like in esters, aromatic ethers and furans), the main diagonal terms of the ${}^a\mathbf{A}(\chi)$ matrix were modified as [Torrens, 2005]

$$[{}^a\mathbf{A}(\chi)]_{ii} = 1.1 \cdot (\chi_i - \chi_C)$$

The use of this correction was extended to all sp^3 -X ($-X-$), for X equal to O, Si, P, S, Ge, As, Se, Sn, Sb, Te, Pb, Bi, and Po.

📖 [Gálvez, Garcia *et al.*, 1995, 1996; Rios-Santamarina, García-Domenech *et al.*, 1998; Torrens, 2003a]

- **topological chirality descriptors** $\rightarrow$ chirality descriptors
- **topological complexity indices** $\equiv$ *Bonchev topological complexity indices* $\rightarrow$ molecular complexity
- **topological diameter** $\rightarrow$ distance matrix
- **topological diameter from edge eccentricity** $\rightarrow$ edge distance matrix
- **topological distance** $\rightarrow$ distance matrix
- **Topological Distance Connectivity Indices** $\rightarrow$ distance matrix
- **topological distance-detour distance combined matrix** $\equiv$ *distance-detour combined matrix* $\rightarrow$ detour matrix
- **topological distance/detour distance quotient matrix** $\equiv$ *distance/detour quotient matrix* $\rightarrow$ detour matrix
- **topological distance-geometric distance combined matrix** $\rightarrow$ molecular geometry
- **topological distance/geometric distance quotient matrix** $\rightarrow$ molecular geometry
- **Topological Distance Measure Connectivity Indices** $\equiv$ *Topological Distance Connectivity Indices* $\rightarrow$ distance matrix
- **topological distance-resistance distance combined matrix** $\rightarrow$ matrices of molecules (⊙ Table M3)

- **topological distance/resistance distance quotient matrix** $\equiv$ distance/resistance quotient matrix $\rightarrow$ resistance matrix
- **topological distance-topographic distance combined matrix** $\rightarrow$ molecular geometry
- **topological distance/topographic distance quotient matrix** $\rightarrow$ molecular geometry
- **topological edge distance** $\rightarrow$ edge distance matrix
- **topological electronegativity index** $\rightarrow$ spectral indices

■ **topological feature maps** ($\equiv$ feature maps)

Feature maps are two-dimensional $\rightarrow$ *self-organizing maps* of the molecular surface [Gasteiger *et al.*, 1994a, 1994b; Gasteiger, Li *et al.*, 1994a, 1994b; Gasteiger and Li, 1994; Anzali, Barnickel *et al.*, 1996]. Any property of the surface such as the $\rightarrow$ *molecular electrostatic potential* (MEP) can be projected into the map and visualized after scaling the property values into selected colors.

To generate a feature map, first, a number (e.g., 20000) of points are randomly sampled on the molecular surface. These surface points, each described by the Cartesian coordinates x , y , and z , are entered into the network for training. Therefore, each neuron of the network is defined by a vector of three weights, corresponding to the surface point coordinates. The neuron weights contain hence information about the shape of the considered molecular surface.

After the network is trained, all surface points are projected into the map and each neuron is assigned the average value of the surface property of those points that excite it. Alternatively, the minimum or the maximum value of the property can be assigned to the neuron (Figure T2).



Figure T2 Feature map of a molecule.

In addition to the MEP, another common property used to generate feature maps is the atom surface assignment (ASA) obtained by cutting the molecular surface into parts that are assigned to the nearest atoms or pharmacophore points.

Feature maps encode information about molecular surface resulting from heteroatoms, conformation, and chirality. They are mainly used to find similarities in a series of molecules and highlight features responsible for biological activity (Figure T3).

The **Comparative Molecular Surface Analysis** (CoMSA) is a 3D-QSAR approach that makes use of the topological feature maps combined with PLS method to quantitatively predict

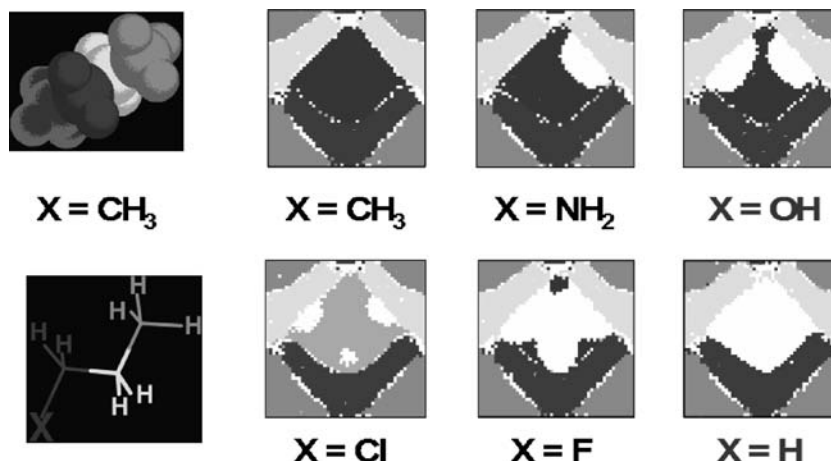


Figure T3 Feature maps of monosubstituted propane with six different substituent groups.

molecular properties [Polanski and Walczak, 2000]. Like CoMFA and related methods, a preliminary superimposition of all the molecules being studied is required. Unlike CoMFA that is based on a set of discrete points, CoMSA exploits average property values calculated for certain areas of the molecular surface.

While the topological feature map of a molecule is obtained by training a neural network with the surface points of the molecule, in CoMSA a comparative feature map is generated first by training the neural network with the surface points of a reference molecule (e.g., the template) in order to produce the template map. Then, the comparative map of each molecule in the data set is obtained by projecting its surface points into the template map. Points are assigned to a neuron only if their distance from the neuron is within a selected winning distance (MD). Neurons corresponding to those parts of the surface of the template molecule that have no counterpart on the surface of the compared molecule will not become excited and thus will be empty. These empty neurons indicate the surface regions of the compared molecule that differ from the corresponding regions of the reference molecule.

Finally, the average value of the surface property of the molecule is assigned to each excited neuron in the comparative map. Two surfaces can be compared with different tolerance depending on the value of the neuron winning distance; different MD thresholds result in a series of different maps. The comparative feature maps can be then described by various global descriptors to be used for QSAR modeling.

A technique similar to CoMSA was previously proposed by Barlow [Barlow, 1995].

 [Anzali, Barnickel *et al.*, 1997; Polanski, 1997; Anzali *et al.*, 1998a, 1998b; Anzali, Gasteiger *et al.*, 1998a, 1998b; Polanski *et al.*, 1998, 2002; Polanski, Gasteiger *et al.*, 1998, 2002; Gasteiger, 2003a; Gasteiger, 2003c; Polanski, 2003; Polanski and Gieleciak, 2003; Polanski, Bak *et al.*, 2004; Gieleciak and Polanski, 2007]

- **topological Hammett function** → combined descriptors
- **topological index of hydrophobicity** → lipophilicity descriptors
- **topological indices** → graph invariants
- **topological information content** $\equiv$ *vertex orbital information content* → orbital information indices

■ topological information indices

These are $\rightarrow$ *graph invariants* that view the $\rightarrow$ *molecular graph* as a source of different probability distributions to which information theory definitions can be applied. They can be considered a quantitative measure of the lack of structural homogeneity or the diversity of a graph, in this way being related to symmetry associated with structure. The information content of a graph is not unique, depending on the equivalence relation defined on the graph.

Several information indices, usually calculated as $\rightarrow$ *total information content* and $\rightarrow$ *mean information content*, are based on partitioning graph elements or matrix elements in $\rightarrow$ *equivalence classes* according to two basic criteria:

- *equality criterion*: elements are considered equivalent if their values are equal;
- *magnitude criterion*: each element is considered as an equivalence class whose cardinality, that is, number of elements, is equal to the magnitude of the element.

The symbol G is usually used to denote the number of equivalence classes and the symbol n_g to denote the cardinality of the set of elements in the g th class.

Collected below are the information indices calculated on the most important $\rightarrow$ *graph-theoretical matrices* such as $\rightarrow$ *adjacency matrix* $\mathbf{A}$, $\rightarrow$ *distance matrix* $\mathbf{D}$, $\rightarrow$ *edge distance matrix* ${}^E\mathbf{D}$, $\rightarrow$ *edge adjacency matrix* $\mathbf{E}$, $\rightarrow$ *vertex-cycle incidence matrix* ${}^{\text{VC}}\mathbf{I}$, $\rightarrow$ *edge-cycle incidence matrix* ${}^{\text{EC}}\mathbf{I}$ [Bonchev, Mekenyan *et al.*, 1981c; Bonchev and Trinajstić, 1982; Bonchev, 1983].

Other specific information indices are $\rightarrow$ *orbital information indices*, $\rightarrow$ *information connectivity indices*, $\rightarrow$ *indices of neighborhood symmetry*, $\rightarrow$ *information layer index*, $\rightarrow$ *chromatic information index*, $\rightarrow$ *Bonchev centric information indices*, information indices from $\rightarrow$ *incidence matrices* and $\rightarrow$ *Hosoya Z index*. Moreover, information indices are the $\rightarrow$ *Shannon Entropy Descriptors* (SHED) and some comprised in the $\rightarrow$ *GETAWAY descriptors*.

Information indices on the adjacency matrix $\mathbf{A}$ are listed below.

• **total information content on the adjacency equality** (${}^V I_{adj}^E$)

Derived from the adjacency matrix $\mathbf{A}$ it is defined as

$${}^V I_{adj}^E = A^2 \cdot \log_2 A^2 - 2B \cdot \log_2 2B - (A^2 - 2B) \cdot \log_2 (A^2 - 2B)$$

where A is the number of graph vertices and B the number of graph edges. Moreover, the entries of the adjacency matrix equal to 1 are $2B$, thus the entries equal to zero are $A^2 - 2B$; in particular, for acyclic graphs the total number of entries equal to 1 is $2(A - 1)$ and the number of entries equal to zero is $A^2 - 2(A - 1)$; for cyclic graphs they are $2A$ and $A^2 - 2A$, respectively.

This index is constant for a given number of atoms A , is insensitive to any kind of branching and can distinguish only between trees and different cyclic structures.

• **mean information content on the adjacency equality** (${}^V \bar{I}_{adj}^E$)

This is calculated by dividing the total information content on the adjacency matrix elements equality by the total number A^2 of adjacency matrix elements as

$${}^V \bar{I}_{adj}^E = -\frac{2B}{A^2} \cdot \log_2 \frac{2B}{A^2} - \left(1 - \frac{2B}{A^2}\right) \cdot \log_2 \left(1 - \frac{2B}{A^2}\right)$$

It is based on the probability of a randomly selected entry to signify or not signify the adjacency.

- **total information content on the adjacency magnitude** (${}^V I_{adj}^M$)

This is derived from the partition of the adjacency matrix elements according to their magnitude and is a trivial quantity as zero entries do not contribute to the $\rightarrow$ *total adjacency index* A_V , that is, the sum of all adjacency matrix elements, corresponding to twice the number B of graph edges:

$${}^V I_{adj}^M = A_V \cdot \log_2 A_V - 2B \cdot 1 \cdot \log_2 1 = A_V \cdot \log_2 A_V$$

where $2B$ is the number of adjacency matrix elements equal to 1.

- **mean information content on the adjacency magnitude** (${}^V \bar{I}_{adj}^M$)

This is calculated by dividing the total information content on the adjacency magnitude by the total adjacency index A_V :

$${}^V \bar{I}_{adj}^M = -2B \cdot \frac{1}{A_V} \cdot \log_2 \frac{1}{A_V} = 1 + \log_2 B$$

where B is the number of graph edges and $2B$ the number of adjacency matrix elements equal to 1.

- **mean information content on the vertex degree equality** (${}^V \bar{I}_{adj,deg}^E$)

Derived from the adjacency matrix A and based on the partition of vertices according to $\rightarrow$ *vertex degree equality*, it is defined as

$${}^V \bar{I}_{adj,deg}^E = - \sum_{g=1}^G \frac{{}^g F}{A} \cdot \log_2 \frac{{}^g F}{A}$$

where ${}^g F$ is the $\rightarrow$ *vertex degree count*, that is, the number of vertices with vertex degree equal to g , A is the number of graph vertices, and G the maximum vertex degree value, that is, the molecule $\rightarrow$ *eccentricity*.

- **mean information content on the vertex degree magnitude** (${}^V \bar{I}_{adj,deg}^M$) ($\equiv$ *degree complexity*)

Based on the partition of vertices according to the vertex degree magnitude, it is defined as

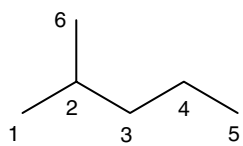
$${}^V \bar{I}_{adj,deg}^M \equiv I^d = - \sum_{i=1}^A \frac{\delta_i}{A_V} \cdot \log_2 \frac{\delta_i}{A_V} = - \sum_{g=1}^G {}^g F \cdot \frac{g}{A_V} \cdot \log_2 \frac{g}{A_V}$$

where A is the number of graph vertices, δ_i is the vertex degree of the i th atom, ${}^g F$ is the $\rightarrow$ *vertex degree count*, that is, the number of vertices with vertex degree equal to g , A_V is the $\rightarrow$ *total adjacency index*, and G the maximum vertex degree value.

This index was proposed as a measure of $\rightarrow$ *molecular complexity* together with some other information indices derived from the $\rightarrow$ *distance matrix* [Raychaudhury, Ray *et al.*, 1984].

Example T2

Calculation of some information indices derived from the adjacency matrix **A** for the H-depleted molecular graph of 2-methylpentane. δ is the vertex degree.



$$\mathbf{A} = \begin{array}{c|cccccc|c} & \text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 & \delta_i \\ \hline 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 2 & 2 & 1 & 0 & 1 & 0 & 0 & 0 & 3 \\ 3 & 3 & 0 & 1 & 0 & 1 & 0 & 0 & 2 \\ 4 & 4 & 0 & 0 & 1 & 0 & 1 & 1 & 2 \\ 5 & 5 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 6 & 6 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}$$

$$A = 6 \quad B = 5$$

$$\text{The vertex degree counts are } {}^1F = 3 \quad {}^2F = 2 \quad {}^3F = 1$$

$$V_{adj}^E = 6^2 \times \log_2 6^2 - 2 \times 5 \times \log_2 (2 \times 5) - (6^2 - 2 \times 5) \times \log_2 (6^2 - 2 \times 5) = 30.687$$

$$V_{adj}^E = -\frac{2 \times 5}{6^2} \times \log_2 \frac{2 \times 5}{6^2} - \left(1 - \frac{2 \times 5}{6^2}\right) \times \log_2 \left(1 - \frac{2 \times 5}{6^2}\right) = 0.852$$

$$V_{adj}^M = (2 \times 5) \times \log_2 (2 \times 5) = 33.219$$

$$V_{adj}^M = 1 + \log_2 5 = 3.322$$

$$V_{adj,deg}^E = -\left[\left(\frac{3}{6} \times \log_2 \frac{3}{6}\right) + \left(\frac{2}{6} \times \log_2 \frac{2}{6}\right) + \left(\frac{1}{6} \times \log_2 \frac{1}{6}\right)\right] = 1.459$$

$$V_{adj,deg}^M = -\left[\left(3 \times \frac{1}{2 \times 5} \times \log_2 \frac{1}{2 \times 5}\right) + \left(2 \times \frac{2}{2 \times 5} \times \log_2 \frac{2}{2 \times 5}\right) + \left(1 \times \frac{3}{2 \times 5} \times \log_2 \frac{3}{2 \times 5}\right)\right] = 2.446$$

Information indices on the edge adjacency matrix **E** are listed below.

- **total information content on the edge adjacency equality** (${}^E I_{adj}^E$)

Derived from the $\rightarrow$ edge adjacency matrix, it is defined as

$${}^E I_{adj}^E = B^2 \cdot \log_2 B^2 - 2N_2 \cdot \log_2 2N_2 - (B^2 - 2N_2) \cdot \log_2 (B^2 - 2N_2)$$

where B is the number of graph edges and N_2 the $\rightarrow$ connection number, that is, the number of second order paths in the $\rightarrow$ molecular graph. Moreover, the entries of the edge adjacency matrix equal to 1 are $2N_2$, thus the entries equal to zero are $B^2 - 2N_2$.

- **mean information content on the edge adjacency equality** (${}^E \bar{I}_{adj}^E$)

This is calculated by dividing the total information content of the edge adjacency equality by the total number B^2 of edge adjacency matrix elements:

$${}^E \bar{I}_{adj}^E = -\frac{2N_2}{B^2} \cdot \log_2 \frac{2N_2}{B^2} - \left(1 - \frac{2N_2}{B^2}\right) \cdot \log_2 \left(1 - \frac{2N_2}{B^2}\right)$$

where B is the number of graph vertices and N_2 the connection number. It is based on the probability of a randomly selected entry to signify or not signify the edge adjacency.

- **total information content on the edge adjacency magnitude** ($^E I_{adj}^M$)

This is derived from the partition of edge adjacency matrix elements according to their magnitude and is a trivial quantity because the zero entries do not contribute to the $\rightarrow$ *total edge adjacency index* A_E ($\rightarrow$ *Platt number*), that is, the sum of all elements of the edge adjacency matrix, corresponding to twice the connection number N_2 :

$$^E I_{adj}^M = A_E \cdot \log_2 A_E - 2N_2 \cdot 1 \cdot \log_2 1 = A_E \cdot \log_2 A_E$$

- **mean information content on the edge adjacency magnitude** ($^E \bar{I}_{adj}^M$)

This is calculated by dividing the total information content on the edge adjacency magnitude by the total edge adjacency index A_E :

$$^E \bar{I}_{adj}^M = -2N_2 \cdot \frac{1}{A_E} \cdot \log_2 \frac{1}{A_E} = 1 + \log_2 N_2$$

where N_2 is the connection number and $2N_2$ is the number of edge adjacency matrix elements equal to 1.

- **mean information content on the edge degree equality** ($^E \bar{I}_{adj,deg}^M$)

Derived from the edge adjacency matrix and based on the partition of edges according to the equality of their edge degrees, it is defined as

$$^E \bar{I}_{adj,deg}^E = - \sum_{g=1}^G \frac{{}^g F_E}{B} \cdot \log_2 \frac{{}^g F_E}{B}$$

where ${}^g F_E$ is the $\rightarrow$ *edge degree count*, that is, the number of edges with $\rightarrow$ *edge degree* equal to g , B is the number of graph edges, and G is the maximum edge degree value.

- **mean information content on the edge degree magnitude** ($^E \bar{I}_{adj,deg}^M$)

Derived from the edge adjacency matrix and based on the partition of edges according to the magnitude of their edge degrees, it is defined as

$$^E \bar{I}_{adj,deg}^M = - \sum_{b=1}^B \frac{\epsilon_b}{A_E} \cdot \log_2 \frac{\epsilon_b}{A_E} = - \sum_{g=1}^G {}^g F_E \cdot \frac{g}{A_E} \cdot \log_2 \frac{g}{A_E}$$

where ϵ_b is the edge degree of the b th edge, A_E is the total edge adjacency index, B is the number of graph edges, and ${}^g F_E$ is the edge degree count, that is, the number of edges with edge degree ϵ equal to g .

Information indices on the distance matrix D are listed below.

- **total information content on the distance equality** ($^V I_D^E$)

Based on the equality of topological distances in the graph, it is defined as [Bonchev and Trinajstić, 1978]

$$^V I_D^E \equiv I_D^E = \frac{A(A-1)}{2} \cdot \log_2 \frac{A(A-1)}{2} - \sum_{g=1}^G {}^g f \cdot \log_2 {}^g f$$

where A is the number of graph vertices, ${}^g f$ is the number of distances equal to g in the triangular D submatrix (i.e., the $\rightarrow$ *graph distance count*), and G is the maximum distance value, that is, the $\rightarrow$ *topological diameter* D .

- **mean information content on the distance equality** (${}^V\bar{I}_D^E$)

Obtained by dividing the total information content of distance equality by the total number $A(A-1)/2$ of distances in the graph, it is calculated as

$${}^V\bar{I}_D^E \equiv \bar{I}_D^E = - \sum_{g=1}^G \frac{2 \cdot {}^g f}{A(A-1)} \cdot \log_2 \frac{2 \cdot {}^g f}{A(A-1)}$$

where ${}^g f$ is the number of distances equal to g in the triangular $\mathbf{D}$ submatrix, and G is the maximum distance value, that is, the topological diameter D .

- **total information content on the distance magnitude** (${}^V I_D^M$)

The information content on the distribution of distances according to their magnitude is defined as

$${}^V I_D^M \equiv I_D^M = W \cdot \log_2 W - \sum_{g=1}^G {}^g f \cdot g \cdot \log_2 g$$

where W is the $\rightarrow$ *Wiener index*, that is, the total sum of distances in the graph, ${}^g f$ is the number of distances equal to g in the graph, and G is the maximum distance value, that is, the topological diameter D .

- **mean information content on the distance magnitude** (${}^V\bar{I}_D^M$)

This is calculated by dividing the total information content of the distance magnitude by the Wiener index W ; it is therefore also called **information Wiener index** [Bonchev and Trinajstić, 1977]:

$${}^V\bar{I}_D^M \equiv \bar{I}_D^M = - \sum_{g=1}^G {}^g f \cdot \frac{g}{W} \cdot \log_2 \frac{g}{W}$$

where ${}^g f$ is the number of distances equal to g in the graph, and G is the maximum distance value, that is, the topological diameter D .

- **mean information content on the distance degree equality** (${}^V\bar{I}_{D, \text{deg}}^E$)

The mean information content of the partition of vertex distance degrees according to their equality is defined as:

$${}^V\bar{I}_{D, \text{deg}}^E \equiv \bar{I}_{D, \text{deg}}^E = - \sum_{g=1}^G \frac{n_g}{A} \cdot \log_2 \frac{n_g}{A}$$

where n_g is the cardinality of the g th set of vertices having equal $\rightarrow$ *vertex distance degree* σ , G is the number of equivalence classes and A the number of graph vertices.

- **mean information content on the distance degree magnitude** (${}^V\bar{I}_{D, \text{deg}}^M$)

The mean information content of the partition of vertex distance degrees according to their magnitude is defined as

$${}^V\bar{I}_{D, \text{deg}}^M \equiv \bar{I}_{D, \text{deg}}^M \equiv H_2 = - \sum_{i=1}^A \frac{\sigma_i}{2W} \cdot \log_2 \frac{\sigma_i}{2W} = - \sum_{g=1}^G n_g \cdot \frac{\sigma_g}{2W} \cdot \log_2 \frac{\sigma_g}{2W}$$

where W is the $\rightarrow$ Wiener index (and $2W$ the $\rightarrow$ Rouvray index), σ_i is the vertex distance degree of the i th atom, n_g is the number of vertices having equal vertex distance degrees in the g th class, σ_g is the vertex distance degree of the vertices in the g th class, and G is the number of equivalence classes. This index was denoted as H_2 by Skorobogatov [Skorobogatov, Konstantinova *et al.*, 1991].

- **autometricity index (H_1)**

The autometricity index is defined as [Skorobogatov, Konstantinova *et al.*, 1991]:

$$H_1 = - \sum_{g=1}^G \frac{n_g}{A} \cdot \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices in the g th class of autometricity. The **autometricity** class is the set of vertices with the same $\rightarrow$ vertex distance code, that is, the vector of vertex distance counts ${}^g f_i$, ${}^g f_i$ being the number of vertices at distance g from the i th vertex.

A group of local vertex invariants and corresponding molecular graph invariants derived from the distance matrix were proposed as quantities related to the $\rightarrow$ molecular complexity [Raychaudhury, Ray *et al.*, 1984, 1992, 1993c]: these are *vertex complexity*, *vertex distance complexity*, *normalized vertex distance complexity*, *relative vertex distance complexity*, *mean extended local information on distances*, *extended local information on distances*, *Balaban-like information indices*, *graph vertex complexity*, and *graph distance complexity*.

- **vertex complexity (v_i^c)**

Derived from the distance matrix, it is a local vertex invariant defined as [Raychaudhury, Ray *et al.*, 1984]

$$v_i^c = - \sum_{g=0}^{\eta_i} \frac{{}^g f_i}{A} \cdot \log_2 \frac{{}^g f_i}{A} = - \sum_{g=0}^{\eta_i} p_g \cdot \log_2 p_g = \frac{1}{A} \cdot \log_2 A - \sum_{g=1}^{\eta_i} p_g \cdot \log_2 p_g$$

where ${}^g f_i$ is the number of distances from the vertex v_i equal to g , that is, the $\rightarrow$ vertex distance counts. ${}^0 f_i$ is always equal to 1, that is, there is only one distance equal to zero; η_i is the $\rightarrow$ atom eccentricity and A the number of graph vertices; p_g is the probability of randomly selecting a distance from vertex v_i equal to g .

- **vertex distance complexity ($\bar{v}_i^d$) ($\equiv$ mean local information on distances, u_i)**

Derived from the distance matrix, it is a local vertex invariant calculated on the $\rightarrow$ H -filled molecular graph and defined as the mean information content of a vertex [Raychaudhury, Ray *et al.*, 1984; Klopman and Raychaudhury, 1988; Klopman, Raychaudhury *et al.*, 1988]:

$$\bar{v}_i^d \equiv u_i \equiv H_{D(i)} = - \sum_{j=1}^A \frac{d_{ij}}{\sigma_i} \cdot \log_2 \frac{d_{ij}}{\sigma_i} = - \sum_{g=1}^{\eta_i} {}^g f_i \cdot \frac{g}{\sigma_i} \cdot \log_2 \frac{g}{\sigma_i}$$

where A is the number of vertices, g spans all of the different distances from the vertex v_i , ${}^g f_i$ is the number of distances from the vertex v_i equal to g , σ_i is the i th $\rightarrow$ vertex distance degree and η_i is the i th $\rightarrow$ atom eccentricity.

The mean local information on distances u_i was originally proposed by Balaban for $\rightarrow$ H -depleted molecular graphs [Balaban and Balaban, 1991; Ivanciuc, Balaban *et al.*, 1993a].

The same quantity, calculated on H-depleted molecular graphs, was also called **vertex information distance index** and denoted as $H_{D(i)}$. The **information distance index** was then defined as [Konstantinova and Paleev, 1990]

$$H_D = \sum_{i=1}^A H_{D(i)}$$

Moreover, another molecular topological index derived from these local invariants was defined by applying the $\rightarrow$ Ivanciuc–Balaban operator and simply called **U index**:

$$U = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (u_i \cdot u_j)^{-1/2}$$

where a_{ij} are the elements of the $\rightarrow$ adjacency matrix which are equal to 1 only for adjacent vertices and zero otherwise, and u_i and u_j are the LOVIs relative to the vertices v_i and v_j ; B and C are the number of edges and the $\rightarrow$ cyclomatic number, respectively. $\rightarrow$ U-like indices were also proposed as an extension of the U index to any $\rightarrow$ graph-theoretical matrix.

A similar approach was also applied to the $\rightarrow$ cardinality layer matrix **LC** of a molecular graph whose elements are the frequencies of the different distances from each vertex, $lc_{ig} = {}^g f_i$. The **information layer index** was then defined as the mean information content as [Konstantinova, 1996, 2006]

$$H_{LC} = \sum_{i=1}^A H_{LC(i)} = - \sum_{i=1}^A \sum_{g=0}^{\eta_i} \frac{lc_{ig}}{A} \cdot \log_2 \frac{lc_{ig}}{A} = - \sum_{i=1}^A \sum_{g=0}^{\eta_i} \frac{{}^g f_i}{A} \cdot \log_2 \frac{{}^g f_i}{A}$$

where A is the number of graph vertices, η_i is the $\rightarrow$ atom eccentricity, that is, the maximum distance from a vertex, and lc_{ig} is the number of vertices located at distance g from the focused i th vertex, which is equal to the number ${}^g f_i$ of distances g from the i th vertex. $H_{LC(i)}$ is called **vertex information layer index** of the i th vertex.

- **normalized vertex distance complexity** ($\tilde{v}_i^d$)

This is derived from the vertex distance complexity, by dividing it by its maximum value [Klopman and Raychaudhury, 1990]:

$$\tilde{v}_i^d = \frac{\bar{v}_i^d}{\log_2 \sigma_i}$$

where σ_i is the $\rightarrow$ vertex distance degree. This index is independent of the molecular size.

- **relative vertex distance complexity** (v_i) ($\equiv$ local information on distances)

Derived from the vertex distance complexity u_i ($\equiv \bar{v}_i^d$), it is defined as the difference between the maximum information content of a vertex and its mean information content [Balaban and Balaban, 1991; Ivanciuc, Balaban *et al.*, 1993a]:

$$v_i = \sigma_i \cdot \log_2 \sigma_i - u_i$$

where σ_i is the $\rightarrow$ vertex distance degree. The corresponding molecular topological index was calculated applying the $\rightarrow$ Ivanciuc–Balaban operator; it is simply called **V index**:

$$V = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (v_i \cdot v_j)^{-1/2}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* that are equal to 1 for adjacent vertices, and zero otherwise; v_i and v_j are the LOVIs relative to the vertices i and j ; B and C are the number of edges and the cyclomatic number, respectively. $\rightarrow$ *V-like indices* were also proposed as an extension of the V index to any graph-theoretical matrix.

- **mean extended local information on distances (γ_i)**

Derived from the partition of the distances from a vertex v_i according to their magnitude, it is defined as the total information content [Balaban and Balaban, 1991; Ivanciuc, Balaban *et al.*, 1993a]:

$$\gamma_i = \sum_{g=1}^{\eta_i} {}^g f_i \cdot g \cdot \log_2 g$$

where g runs over all of the different distances from the vertex i , ${}^g f_i$ is the number of distances from the vertex i equal to g , and η_i is the i th atom eccentricity. The corresponding molecular topological index was calculated applying the Ivanciuc–Balaban operator; it is simply called **Y index**:

$$Y = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\gamma_i \cdot \gamma_j)^{-1/2}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* that are equal to 1 for adjacent vertices, and zero otherwise; γ_i and γ_j are the LOVIs relative to the vertices i and j ; B and C are the number of edges and the cyclomatic number, respectively. $\rightarrow$ *Y-like indices* were also proposed as an extension of the Y index to any graph-theoretical matrix.

- **extended local information on distances (x_i)**

This is defined as the total information content of a vertex [Balaban and Balaban, 1991; Ivanciuc, Balaban *et al.*, 1993a]:

$$x_i = \sigma_i \cdot \log_2 \sigma_i - \gamma_i$$

where σ_i is the vertex distance degree and γ_i is the mean extended local information on distances of the i th vertex. The corresponding molecular topological index was calculated by applying the $\rightarrow$ *Ivanciuc–Balaban operator*; it is simply called **X index**:

$$X = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (x_i \cdot x_j)^{-1/2}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix*, which are equal to 1 for adjacent vertices, and zero otherwise; x_i and x_j are the LOVIs relative to the vertices i and j ; B and C are the number of edges and the cyclomatic number, respectively. $\rightarrow$ *X-like indices* were also proposed as an extension of the X index to any graph-theoretical matrix.

• **Balaban-like information indices** ($\equiv$ information-theory operators)

By analogy with U , V , X , and Y indices, based on the information content on distances, U -like indices $U(\mathbf{M}, w)$, V -like indices $V(\mathbf{M}, w)$, X -like indices $X(\mathbf{M}, w)$, and Y -like indices $Y(\mathbf{M}, w)$ were defined in terms of the information content on the elements of any $\rightarrow$ graph-theoretical matrix $\mathbf{M}(w)$, representing a vertex- and/or edge-weighted molecular graph and calculated by the $\rightarrow$ weighting scheme w [Ivanciuc and Balaban, 1999a; Ivanciuc, Ivanciuc *et al.*, 1999a].

Balaban-like information indices are calculated by replacing vertex distance degrees of the $\rightarrow$ Balaban distance connectivity index J with different local invariants that measure the information content of the matrix elements associated with the respective vertex, defined as

$$\begin{aligned} u_i(\mathbf{M}, w) &= - \sum_{j=1}^A \frac{|[\mathbf{M}(w)]_{ij}|}{VS_i(|\mathbf{M}(w)|)} \cdot \log_2 \frac{|[\mathbf{M}(w)]_{ij}|}{VS_i(|\mathbf{M}(w)|)} \\ v_i(\mathbf{M}, w) &= VS_i(|\mathbf{M}(w)|) \cdot \log_2 VS_i(|\mathbf{M}(w)|) - u_i(\mathbf{M}, w) \\ y_i(\mathbf{M}, w) &= \sum_{j=1}^A |[\mathbf{M}(w)]_{ij}| \cdot \log_2 |[\mathbf{M}(w)]_{ij}| \\ x_i(\mathbf{M}, w) &= VS_i(|\mathbf{M}(w)|) \cdot \log_2 VS_i(|\mathbf{M}(w)|) - y_i(\mathbf{M}, w) \end{aligned}$$

where A is the number of graph vertices, VS_i , which stands for $\rightarrow$ vertex sum operator, is the sum of the elements of the i th row of the matrix $\mathbf{M}(w)$ and the summations go over the nonzero matrix elements. Moreover, because the logarithm is defined only for positive arguments, absolute values of matrix elements need to be used.

Since some elements of matrix $\mathbf{M}$ may have values lower than one, giving negative values of the local invariants, the corresponding U -like indices $U(\mathbf{M}, w)$, V -like indices $V(\mathbf{M}, w)$, X -like indices $X(\mathbf{M}, w)$, and Y -like indices $Y(\mathbf{M}, w)$ were then defined by applying a modified Balaban-like formula:

$$\begin{aligned} U(\mathbf{M}, w) &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot f(u_i(\mathbf{M}, w), u_j(\mathbf{M}, w)) \\ V(\mathbf{M}, w) &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot f(v_i(\mathbf{M}, w), v_j(\mathbf{M}, w)) \\ X(\mathbf{M}, w) &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot f(x_i(\mathbf{M}, w), x_j(\mathbf{M}, w)) \\ Y(\mathbf{M}, w) &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot f(y_i(\mathbf{M}, w), y_j(\mathbf{M}, w)) \end{aligned}$$

where a_{ij} are the elements of the $\rightarrow$ adjacency matrix that are equal to 1 only for pairs of adjacent vertices, and the subscripts i and j refer to the vertices v_i and v_j ; B and C are the number of edges and the $\rightarrow$ cyclomatic number, respectively. The function f is defined as

$$f(u_i, u_j) = \begin{cases} (u_i \cdot u_j)^{-1/2} & \text{if } u_i \cdot u_j > 0 \\ -(|u_i \cdot u_j|)^{-1/2} & \text{if } u_i \cdot u_j < 0 \end{cases}$$

These molecular descriptors were derived from several graph-theoretical matrices such as $\rightarrow$ reciprocal distance matrix [Ivanciuc and Balaban, 1999c], $\rightarrow$ reverse Wiener matrix [Ivanciuc,

Ivanciuc *et al.*, 2002b], $\rightarrow$ *Ivanciuc weighted distance matrices* [Ivanciuc, Ivanciuc *et al.*, 1999a], $\rightarrow$ *complementary distance matrix* [Ivanciuc, Ivanciuc *et al.*, 2000a], $\rightarrow$ *geometry matrix* [Ivanciuc and Balaban, 1999b], and $\rightarrow$ *reciprocal geometry matrix* [Ivanciuc and Balaban, 1999b].

- **graph vertex complexity (H_V)**

Derived from the $\rightarrow$ *distance matrix*, it is defined as molecular average vertex complexity [Raychaudhury, Ray *et al.*, 1984]:

$$H_V = \frac{1}{A} \cdot \sum_{i=1}^A v_i^c$$

where v_i^c is the $\rightarrow$ *vertex complexity* and A is the number of graph vertices.

- **graph distance complexity (H_D)**

Derived from the distance matrix, it is a graph invariant defined as [Raychaudhury, Ray *et al.*, 1984]

$$H_D = \sum_{i=1}^A \frac{\sigma_i}{I_{\text{ROUV}}} \cdot v_i^d = \sum_{i=1}^A \frac{\sigma_i}{2W} \cdot v_i^d$$

where v_i^d is the $\rightarrow$ *vertex distance complexity*, σ_i is the i th $\rightarrow$ *vertex distance degree*, I_{ROUV} is the $\rightarrow$ *Rouvray index* and W the $\rightarrow$ *Wiener index*.

Information indices on the edge distance matrix ${}^E D$ are listed below.

- **total information content on the edge distance equality (${}^E I_D^E$)**

Based on the equality or inequality of edge distances in the graph it is defined as

$${}^E I_D^E = \frac{B \cdot (B-1)}{2} \cdot \log_2 \frac{B \cdot (B-1)}{2} - \sum_{g=1}^G {}^g f \cdot \log_2 {}^g f$$

where B is the number of edges, ${}^g f$ is the number of edge distances equal to g in the graph, and G is the maximum edge distance value.

- **mean information content on the edge distance equality (${}^E \bar{I}_D^E$)**

Obtained dividing the total information content of the edge distance equality by the total number $B \times (B-1)/2$ of edge distances in the graph it is calculated as

$${}^E \bar{I}_D^E = - \sum_{g=1}^G \frac{2 \cdot {}^g f}{B \cdot (B-1)} \cdot \log_2 \frac{2 \cdot {}^g f}{B \cdot (B-1)}$$

where B is the number of edges, ${}^g f$ is the number of edge distances equal to g in the graph, and G is the maximum edge distance value.

- **total information content on the edge distance magnitude (${}^E I_D^M$)**

The information content of the distribution of edge distances according to their magnitude is defined as

$${}^E I_D^M = {}^E W \cdot \log_2 {}^E W - \sum_{g=1}^G {}^g f \cdot g \cdot \log_2 g$$

where ${}^E W$ is the $\rightarrow$ *edge Wiener index*, that is, the half of the $\rightarrow$ *total edge distance* D_E , ${}^E f$ is the number of edge distances equal to g in the graph, and G is the maximum edge distance value.

- **mean information content on the edge distance magnitude** (${}^E \bar{I}_D^M$)

It is calculated dividing the total information content of the edge distance magnitude by the $\rightarrow$ *edge Wiener index* ${}^E W$:

$${}^E \bar{I}_D^M = - \sum_{g=1}^G {}^E f \cdot \frac{g}{{}^E W} \cdot \log_2 \frac{g}{{}^E W}$$

where ${}^E f$ is the number of edge distances equal to g in the graph and G is the maximum edge distance value.

- **mean information content on the edge distance degree equality** (${}^E \bar{I}_{D, \deg}^E$)

The mean information content on the partition of $\rightarrow$ *edge distance degrees* ${}^E \sigma$ according to their equality is defined as

$${}^E \bar{I}_{D, \deg}^E = - \sum_{g=1}^G \frac{n_g}{B} \cdot \log_2 \frac{n_g}{B}$$

where n_g is the cardinality of the g th set of vertices having an equal edge distance degree ${}^E \sigma$, G is the number of different edge distance degree values and B is the graph vertices.

- **mean information content on the edge distance degree magnitude** (${}^E \bar{I}_{D, \deg}^M$)

The mean information content of the partition of edge distance degrees according to their magnitude is defined as

$${}^E \bar{I}_{D, \deg}^M = - \sum_{b=1}^B \frac{{}^E \sigma_b}{D_E} \cdot \log_2 \frac{{}^E \sigma_b}{D_E} = - \sum_{g=1}^G n_g \cdot \frac{{}^E \sigma_g}{D_E} \cdot \log_2 \frac{{}^E \sigma_g}{D_E}$$

where D_E is the $\rightarrow$ *total edge distance* and ${}^E \sigma_b$ is the $\rightarrow$ *edge distance degree* of the b th edge, n_g is the number of edges having equal edge distance degree in the g th class, ${}^E \sigma_g$ is the edge distance degree of the edges in the g th class, and G is the number of equivalence classes.

Information indices on the vertex-cycle incidence matrix ${}^{VC} I$ are listed below.

- **total information content on the vertex-cycle matrix elements equality** (${}^V I_{cyc}^E$)

This is derived from the $\rightarrow$ *vertex-cycle incidence matrix* and is based on the partition of matrix elements according to their equalities:

$${}^V I_{cyc}^E = A \cdot C^+ \log_2 A \cdot C^+ - n_1 \log_2 n_1 - n_0 \log_2 n_0$$

where A and C^+ are the number of graph vertices and the $\rightarrow$ *cyclicity*, that is, the total number of rings in the graph, respectively; n_1 and n_0 are the number of matrix entries equal to 1 and the number of entries equal to zero, respectively.

- **mean information content on the vertex-cycle matrix elements equality** (${}^V\bar{I}_{cyc}^E$)

This is calculated by dividing the total information content on the vertex-cycle matrix elements equality by the total number of matrix elements $A \cdot C^+$ as

$${}^V\bar{I}_{cyc}^E = -\frac{n_1}{A \cdot C^+} \log_2 \frac{n_1}{A \cdot C^+} - \frac{n_0}{A \cdot C^+} \log_2 \frac{n_0}{A \cdot C^+}$$

where A is the number of graph vertices, C^+ the number of rings, n_1 and n_0 are the number of matrix entries equal to 1 and the number of entries equal to zero, respectively. It is based on the probability of a randomly selected entry to signify or not that a vertex belongs to a given cycle.

- **total information content on the vertex-cycle matrix elements magnitude** (${}^V I_{cyc}^M$)

This is derived from the $\rightarrow$ vertex-cycle incidence matrix and is based on the magnitude of matrix elements:

$${}^V I_{cyc}^M = C_{VC} \cdot \log_2 C_{VC} - 1 \cdot \log_2 1 = C_{VC} \cdot \log_2 C_{VC}$$

where C_{VC} is the $\rightarrow$ total vertex cyclicity, that is, the total sum of matrix elements. Since the zero-entries do not contribute to the total vertex cyclicity, the information index ${}^V I_{cyc}^M$ is simply a logarithmic function of the number of matrix entries equal to 1.

- **mean information content on the vertex-cycle matrix elements magnitude** (${}^V\bar{I}_{cyc}^M$)

This is calculated by dividing the total information content on the vertex-cycle matrix elements magnitude by the $\rightarrow$ total vertex cyclicity C_{VC} as

$${}^V\bar{I}_{cyc}^M = -C_{VC} \cdot \frac{1}{C_{VC}} \log_2 \frac{1}{C_{VC}} = \log_2 C_{VC}$$

This information index is simply the total vertex cyclicity expressed in bits, that is, information units.

- **mean information content on the vertex cyclic degree equality** (${}^V\bar{I}_{cyc,deg}^E$)

Derived from the $\rightarrow$ vertex-cycle incidence matrix and based on the partition of vertices according to the equality of their cyclic degrees, it is defined as

$${}^V\bar{I}_{cyc,deg}^E = -\sum_{g=1}^G \frac{n_g}{A} \cdot \log_2 \frac{n_g}{A}$$

where n_g is the number of vertices having the same $\rightarrow$ vertex cyclic degree γ^v , G is the number of different degree values, and A is the number of graph vertices.

- **mean information content on the vertex cyclic degree magnitude** (${}^V\bar{I}_{cyc,deg}^M$)

Derived from the $\rightarrow$ vertex-cycle incidence matrix and based on the magnitude of cyclic degrees of the vertices, it is defined as

$${}^V\bar{I}_{cyc,deg}^M = -\sum_{i=1}^A \frac{\gamma_i^v}{C_{VC}} \cdot \log_2 \frac{\gamma_i^v}{C_{VC}}$$

where γ_i^v is the vertex cyclic degree of the i th vertex, C_{VC} is the total vertex cyclicity, that is, the sum of all cyclic degrees, and A is the number of graph vertices.

Information indices on the edge-cycle incidence matrix ${}^{EC}I$ are listed below.

- **total information content on the edge-cycle matrix elements equality** (${}^E I_{cyc}^E$)

This is derived from the $\rightarrow$ *edge-cycle incidence matrix* and is based on the partition of matrix elements according to their equalities:

$${}^E I_{cyc}^E = (B \cdot C^+) \cdot \log_2(B \cdot C^+) - n_1 \cdot \log_2 n_1 - n_0 \cdot \log_2 n_0$$

where B and C^+ are the number of graph edges and the cyclicity, that is, the total number of rings in the graph, respectively; n_1 and n_0 are the number of matrix entries equal to 1 and the number of entries equal to zero, respectively.

- **mean information content on the edge-cycle matrix elements equality** (${}^E \bar{I}_{cyc}^E$)

This is calculated by dividing the total information content on the edge-cycle matrix elements equality by the total number of matrix elements BC^+ :

$${}^E \bar{I}_{cyc}^E = -\frac{n_1}{B \cdot C^+} \cdot \log_2 \frac{n_1}{B \cdot C^+} - \frac{n_0}{B \cdot C^+} \cdot \log_2 \frac{n_0}{B \cdot C^+}$$

where B is the number of graph edges, C^+ the number of rings, n_1 and n_0 are the number of matrix entries equal to 1 and the number of entries equal to zero, respectively. It is based on the probability of a randomly selected entry to signify or not signify that an edge belongs to a given cycle.

- **total information content on the edge-cycle matrix elements magnitude** (${}^E I_{cyc}^M$)

This is derived from the $\rightarrow$ *edge-cycle incidence matrix* and is based on the magnitude of matrix elements:

$${}^E I_{cyc}^M = C_{EC} \cdot \log_2 C_{EC} - 1 \cdot \log_2 1 = C_{EC} \cdot \log_2 C_{EC}$$

where C_{EC} is the $\rightarrow$ *total edge cyclicity*, that is, the total sum of matrix elements. Since the zero-entries do not contribute to the total edge cyclicity, the information index ${}^E I_{cyc}^M$ is simply a logarithmic function of the number of matrix entries equal to 1.

- **mean information content on the edge-cycle matrix elements magnitude** (${}^E \bar{I}_{cyc}^M$)

This is calculated by dividing the total information content on the edge-cycle matrix elements magnitude by the $\rightarrow$ *total edge cyclicity* C_{EC} :

$${}^E \bar{I}_{cyc}^M = -\left(C_{EC} \cdot \frac{1}{C_{EC}}\right) \cdot \log_2 \frac{1}{C_{EC}} = \log_2 C_{EC}$$

This information index is simply the total edge cyclicity expressed in bits, that is, information units.

- **mean information content on the edge cyclic degree equality** (${}^E \bar{I}_{cyc,deg}^E$)

Derived from the $\rightarrow$ *edge-cycle incidence matrix* and based on the partition of edges according to the equality of their cyclic degrees, it is defined as

$${}^E\bar{I}_{\text{cyc,deg}}^E = - \sum_{g=1}^G \frac{n_g}{B} \cdot \log_2 \frac{n_g}{B}$$

where n_g is the number of edges having the same $\rightarrow$ *edge cyclic degree* γ^e , G is the number of different degree values, and B is the number of graph edges.

• **mean information content on the edge cyclic degree magnitude** (${}^E\bar{I}_{\text{cyc,deg}}^M$)

Derived from the $\rightarrow$ *edge-cycle incidence matrix* and based on the magnitude of cyclic degrees of the edges, it is defined as

$${}^E\bar{I}_{\text{cyc,deg}}^M = - \sum_{i=1}^B \frac{\gamma_i^e}{C_{\text{EC}}} \cdot \log_2 \frac{\gamma_i^e}{C_{\text{EC}}}$$

where γ_i^e is the $\rightarrow$ *edge cyclic degree* of the i th edge, C_{EC} is the $\rightarrow$ *total edge cyclicity*, that is, the sum of all cyclic degrees, and B is the number of graph edges.

📖 [Gutman and Trinajstić, 1973a; Bonchev and Trinajstić, 1977; Bonchev, Balaban *et al.*, 1980; Mekenyan, Bonchev *et al.*, 1980; Cvetković and Gutman, 1985; Kunz, 1986; Nikolić, Medicsaric *et al.*, 1993; Ivanciuc and Balaban, 1999c; Ivanciuc, Ivanciuc *et al.*, 2000d; Konstantinova, Skorobogatov *et al.*, 2003; Konstantinova and Vidyuk, 2003; Agrawal, Srivastava *et al.*, 2004; Bonchev, 2005; Konstantinova, 2006; Jelfs, Ertl *et al.*, 2007]

- **topological lipophilicity potential** $\rightarrow$ molecular interaction fields (☉ hydrophobic fields)
- **Topological MAXimum Cross-Correlation descriptors** $\equiv$ *TMACC descriptors* $\rightarrow$ autocorrelation descriptors

■ **Topological Polar Surface Area (TPSA)**

The topological polar surface area (TPSA) is calculated according to the model proposed by Ertl [Ertl, Rohde *et al.*, 2000, 2001; Ertl, 2008], based on a $\rightarrow$ *group contribution method*.

The TPSA of a molecule is determined by the summation of tabulated surface contributions of polar atom types (Table T2):

$$\text{TPSA} = \sum_i N_i \cdot G_i$$

where the summation runs over the defined types of polar atoms, N_i is the frequency of the i th atom type in the molecule, and G_i is its surface contribution. The surface contributions were calculated by least-squares fitting of the TPSA-based atom types to the single conformer 3D PSA of a training set consisting of 34 810 drug-like molecules taken from the World Drug Index database. The statistical parameters of the model are $R^2 = 0.982$ and $s = 7.83$.

DRAGON software calculates two topological polar surface area descriptors, namely TPSA (NO) and TPSA(tot), the first being derived only from polar fragments with nitrogen and oxygen and the second from polar fragments with nitrogen and oxygen plus “slightly polar” fragments containing phosphorus and sulfur.

Table T2 Surface contributions of polar atom types.

No.	Atom type	PSA contributions	No.	Atom type	PSA contributions
1	[N](-*)(-*)-*	3.24	23	[nH](:*)(:)*	15.79
2	[N](-*)=*	12.36	24	[n ⁺](:*)(:)*	4.10
3	[N]#*	23.79	25	[n ⁺](-*)(:*)(:)*	3.88
4	[N](-*)(=*) ^{*a}	11.68	26	[nH ⁺](:*)(:)*	14.14
5	[N](=*)# ^{*b}	13.60	27	[O](-*)-*	9.23
6	[N]1(-*)-*-*-1 ^c	3.01	28	[O]1-***-1 ^c	12.53
7	[NH](-*)-*	12.03	29	[O]=*	17.07
8	[NH]1-*-*-1 ^c	21.94	30	[OH]-*	20.23
9	[NH]=*	23.85	31	[O-]-*	23.06
10	[NH2]-*	26.02	32	[o](:*)(:)*	13.14
11	[N ⁺](-*)(-*)(-*)-*	0.00	33	[S](-*)-*	25.30
12	[N ⁺](-*)(-*)=*	3.01	34	[S]=*	32.09
13	[N ⁺](-*)# ^{*d}	4.36	35	[S](-*)(-*)=*	19.21
14	[NH ⁺](-*)(-*)-*	4.44	36	[S](-*)(-*)(=*)=*	8.38
15	[NH ⁺](-*)=*	13.97	37	[SH]-*	38.80
16	[NH2 ⁺](-*)-*	16.61	38	[s](:*)(:)*	28.24
17	[NH2 ⁺]=*	25.59	39	[s](=*)(:*)(:)*	21.70
18	[NH3 ⁺]-*	27.64	40	[P](-*)(-*)(-*)-*	13.59
19	[n](:*)(:)*	12.89	41	[P](-*)=*	34.14
20	[n](:*)(:*)(:)*	4.41	42	[P](-*)(-*)(-*)=*	9.81
21	[n](-*)(:*)(:)*	4.93	43	[PH](-*)(-*)=*	23.47
22	[n](=*)(:*)(:)* ^e	8.39			

An asterisk (*) stands for any nonhydrogen atom, – for a single bond, = for a double bond, # for a triple bond, : for an aromatic bond; atomic symbol in lowercase means that the atom is part of an aromatic system.

^aAs in nitro group.

^bMiddle nitrogen in azide group.

^cAtom in a three-membered ring.^dNitrogen in isocyano group.^eAs in pyridine *N*-oxide.

- **topological radius** → distance matrix
- **topological radius from edge eccentricity** → edge distance matrix
- **topological representation** → molecular descriptors
- **topological resonance energy** → delocalization degree indices
- **topological state** → weighted matrices ($\odot$ weighted distance matrices)
- **topological state matrix** → weighted matrices ($\odot$ weighted distance matrices)
- **topological steric effect index** → steric descriptors ($\odot$ Taft steric constant)
- **topological substructural molecular design** → edge adjacency matrix
- **topological torsion descriptor** → substructure descriptors
- **topostructural indices** → graph invariants
- **TOPP descriptors** $\equiv$ *Triplet Of Pharmacophoric Points descriptors* → substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **TOPS-MODE** $\equiv$ *TOPological Substructural MOlecular DEsign* → edge adjacency matrix
- **torsion angles** → molecular geometry

- **TOSS-MODE** $\equiv$ *TO*pological *S*ub*S*tructural *M*olecular *D*esign $\rightarrow$ edge adjacency matrix
- **total absolute atomic charge** $\rightarrow$ charge descriptors
- **total adjacency index** $\rightarrow$ adjacency matrix
- **total bond resistance deficit index** $\rightarrow$ resistance matrix
- **total charge weighted negative surface area** $\rightarrow$ charged partial surface area descriptors
- **total charge weighted positive surface area** $\rightarrow$ charged partial surface area descriptors
- **total connection orbital information content** $\rightarrow$ orbital information indices
- **total edge adjacency index** $\rightarrow$ edge adjacency matrix
- **total edge cyclicity** $\rightarrow$ incidence matrices ($\odot$ cycle matrices)
- **total edge distance** $\rightarrow$ edge distance matrix
- **total edge orbital information content** $\rightarrow$ orbital information indices
- **total electrophilic superdelocalizability** $\rightarrow$ quantum-chemical descriptors ($\odot$ electrophilic superdelocalizability)
- **total excess bond conductance index** $\rightarrow$ resistance matrix
- **total hydrophobic surface area** $\rightarrow$ charged partial surface area descriptors
- **total information content** $\rightarrow$ information content
- **total information content on the adjacency equality** $\rightarrow$ topological information indices
- **total information content on the adjacency magnitude** $\rightarrow$ topological information indices
- **total information content on the distance equality** $\rightarrow$ topological information indices
- **total information content on the distance magnitude** $\rightarrow$ topological information indices
- **total information content on the edge adjacency equality** $\rightarrow$ topological information indices
- **total information content on the edge adjacency magnitude** $\rightarrow$ topological information indices
- **total information content on the edge-cycle matrix elements equality** $\rightarrow$ topological information indices
- **total information content on the edge-cycle matrix elements magnitude** $\rightarrow$ topological information indices
- **total information content on the edge distance equality** $\rightarrow$ topological information indices
- **total information content on the edge distance magnitude** $\rightarrow$ topological information indices
- **total information content on the incidence matrix** $\rightarrow$ incidence matrices ($\odot$ vertex-edge incidence matrix)
- **total information content on the leverage equality** $\rightarrow$ GETAWAY descriptors
- **total information content on the vertex-cycle matrix elements equality** $\rightarrow$ topological information indices
- **total information content on the vertex-cycle matrix elements magnitude** $\rightarrow$ topological information indices
- **total information index on atomic composition** $\rightarrow$ atomic composition indices
- **total interaction energy fields** $\rightarrow$ molecular interaction fields
- **total molecular surface area** $\equiv$ *van der Waals surface area* $\rightarrow$ molecular surface ($\odot$ van der Waals molecular surface)
- **total negative charge** $\rightarrow$ charge descriptors
- **total nucleophilic superdelocalizability** $\rightarrow$ quantum-chemical descriptors ($\odot$ nucleophilic superdelocalizability)
- **total path count** $\rightarrow$ path counts
- **total path number** $\equiv$ *total path count* $\rightarrow$ path counts

- **Total Pharmacophore Diversity fingerprints** $\equiv$ *ToPD fingerprints* $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **total polar surface area** $\rightarrow$ charged partial surface area descriptors
- **total positive charge** $\rightarrow$ charge descriptors
- **total radical superdelocalizability** $\rightarrow$ quantum-chemical descriptors ($\odot$ radical superdelocalizability)
- **total ranking methods** $\rightarrow$ chemometrics ($\odot$ ranking methods)
- **total self-atom polarizability** $\rightarrow$ electric polarization descriptors ($\odot$ atom-atom polarizability)
- **total sequence count** $\rightarrow$ sequence matrices
- **total softness** $\rightarrow$ quantum-chemical descriptors ($\odot$ softness indices)
- **total solvent-accessible surface area** $\equiv$ *solvent-accessible surface area* $\rightarrow$ molecular surface ($\odot$ solvent-accessible molecular surface)
- **total square atomic charge** $\rightarrow$ charge descriptors ($\odot$ total absolute atomic charge)
- **total structure connectivity index** $\rightarrow$ connectivity indices
- **total subgraph count** $\rightarrow$ molecular graph
- **total sum of squares** $\rightarrow$ regression parameters
- **total sum operator** $\rightarrow$ algebraic operators
- **total topological state index** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **total topological information content** $\rightarrow$ orbital information indices
- **total valence topological state index** $\rightarrow$ weighted matrices ($\odot$ weighted distance matrices)
- **total vertex cyclicity** $\rightarrow$ incidence matrices ($\odot$ cycle matrices)
- **total vertex distance** $\equiv$ *Rouvray index* $\rightarrow$ distance matrix
- **total walk count** $\rightarrow$ walk counts
- **toxicological indices** $\rightarrow$ biological activity indices
- **toxicophore** $\rightarrow$ drug design
- **trace** $\rightarrow$ algebraic operators
- **trail** $\rightarrow$ graph
- **training/evaluation splitting** $\rightarrow$ validation techniques
- **training set** $\rightarrow$ data set
- **Transferable Atom Equivalent descriptors** $\equiv$ *TAE descriptors*
- **transformations of molecular descriptors** $\rightarrow$ molecular descriptors
- **transition matrices** $\equiv$ *stochastic matrices* $\rightarrow$ algebraic operators
- **translational invariance** $\rightarrow$ molecular descriptors ($\odot$ invariance properties of molecular descriptors)
- **transmission coefficient** $\rightarrow$ electronic substituent constants
- **transposition of a matrix** $\rightarrow$ algebraic operators
- **Tratch–Stankevitch–Zefirov-type indices** $\equiv$ *generalized expanded Wiener numbers* $\rightarrow$ expanded distance matrices
- **tree** $\rightarrow$ graph
- **tree-likeness indices** $\rightarrow$ Szeged matrices
- **triangular descriptors** $\rightarrow$ substructure descriptors
- **TRI descriptors** $\rightarrow$ molecular descriptors ($\odot$ invariance properties of molecular descriptors)
- **triple bond count** $\rightarrow$ multiple bond descriptors
- **triplet descriptors** $\equiv$ *triangular descriptors* $\rightarrow$ substructure descriptors

- **triplet indices** $\equiv$ *triplet topological indices* $\rightarrow$ MPR approach
- **Triplets Of Pharmacophoric Points descriptors** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **triplet topological indices** $\rightarrow$ MPR approach
- **true negative rate** $\rightarrow$ classification parameters
- **true positive rate** $\rightarrow$ classification parameters

■ TSAR descriptors

These are the molecular descriptors calculated by the TSAR software [TSAR – Oxford Molecular Ltd., 1999]. They include: molecular surface area and volume, moments of inertia, ellipsoidal volume, $\rightarrow$ *Verloop parameters*, dipole moments, lipole moments, molecular weight, $\rightarrow$ *Wiener index*, $\rightarrow$ *Molecular Connectivity Indices*, molecular shape indices, $\rightarrow$ *electrotopological state indices*, $\log P$, number of defined atoms (carbons, oxygens, etc.), and number of defined functional groups (methyl, hydroxyl, etc.). Most of these descriptors can be calculated for both the entire molecule and substituents. There are also descriptors, which can be calculated for substituents only, such as the bond dipole and bond lipole descriptors. Due to specific definitions and numbering of substituents, these descriptors are capable of distinguishing between stereoisomers. To this end, each molecule is described as a template with a defined number of substituents attached to this template by a single bond. A single hydrogen atom may also serve as a substituent. All substituents are numbered according to their positions in molecules.

Applications of the TSAR descriptors reported in the literature are in Refs. [Horwell, Howson *et al.*, 1995; Benigni, Gallo *et al.*, 1999; Rodrigues, Lopes *et al.*, 2000; Klocker, Wailzer *et al.*, 2002a, 2002b; Patel, Schultz *et al.*, 2002; Schultz, Cronin *et al.*, 2002; Kovatcheva, Buchbauer *et al.*, 2003; Netzeva, Aptula *et al.*, 2003; Schefzik, Kibbey *et al.*, 2004; Aptula, Roberts *et al.*, 2005a; Schultz, Netzeva *et al.*, 2005]

- **T-scale** $\rightarrow$ biodescriptors ($\odot$ amino acid descriptors)
- **TSEI** $\equiv$ *Topological Steric Effect Index* $\rightarrow$ steric descriptors ($\odot$ Taft steric constant)
- **Turko–Ivanciuc fitness function** $\rightarrow$ regression parameters
- **Tversky similarity coefficient** $\rightarrow$ similarity/diversity
- **two-degree cyclic atom count** $\rightarrow$ ring descriptors

U

- **U index** → topological information indices (⊙ vertex distance complexity)
- **U-like indices** → topological information indices (⊙ Balaban-like information indices)
- **unbiased constants** → electronic substituent constants
- **uncertainty in the prediction** → regression parameters
- **uniform-length descriptors** → vectorial descriptors
- **uninformative variable elimination by PLS** → variable subset selection
- **unipolarity** → distance matrix
- **unique atomic code** ≡ *canonical numbering*
- **unique atomic ordering** ≡ *canonical numbering*
- **unit matrix** → algebraic operators
- **UNITY fingerprints** → substructure descriptors (⊙ fingerprints)
- **Unsat index** → multiple bond descriptors
- **Unsat-p index** → multiple bond descriptors
- **unsaturation index** → multiple bond descriptors
- **unsaturation number** → multiple bond descriptors
- **unsaturated bonds** → multiple bond descriptors
- **unsymmetrical interaction geodesic matrices** → weighted matrices (⊙ weighted distance matrices)
- **unsymmetrical interaction graph matrices** → weighted matrices (⊙ weighted distance matrices)
- **unusual vertices** → walk counts
- **unusual walks** → walk counts
- **Ursu-Diudea chirality** → chirality descriptors

V

- **VAA indices** → spectral indices (⊙ eigenvalues of the adjacency matrix)
- **VAD indices** → spectral indices (⊙ eigenvalues of the distance matrix)
- **valence algebraic semisum charge-transfer index** → topological charge indices
- **valence charge-transfer indices** → topological charge indices
- **valence connectivity indices** → connectivity indices
- **valence DP indices** → indices of differences of path lengths
- **valence electron descriptor** → vertex degree
- **Valence Electron Mobile environment** → ETA indices
- **valence global topological charge-transfer index** → topological charge indices
- **valence mean topological charge-transfer index** → topological charge indices
- **valence molecular connectivity indices** → connectivity indices
- **valence shell** → path counts
- **valence shell counts** → path counts
- **valence state indicator** → electrotopological state indices
- **valence state indicator** → vertex degree
- **valence topological charge-transfer index** → topological charge indices
- **valence topological state** → weighted matrices (⊙ weighted distance matrices)
- **valence topological state matrix** → weighted matrices (⊙ weighted distance matrices)
- **valence vector semisum charge-transfer index** → topological charge indices
- **valence vertex degree** → vertex degree
- **valence-weighted Schultz distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **valence Zagreb indices** → Zagreb indices
- **valency of a vertex** → vertex degree
- **valency index** → quantum-chemical descriptors (⊙ electron density)

■ validation techniques

Validation techniques constitute a fundamental tool for the assessment of the validity of models obtained from a → *data set* by multivariate regression and classification methods. Validation techniques are used to check the prediction power of the models, that is, to give a measure of their capability to perform reliable predictions of the modeled response for new cases for which the response is unknown [Diaconis and Efron, 1983; Myers, 1986; Cramer III, Bunce *et al.*, 1988; Rawlings, 1988; Golbraikh and Tropsha, 2002c].

The main goal of validation techniques is to select (to find) the model(s) with the best predictive ability to allow a real use of the model(s) for future predictions. It should be noted that, when more than one final model is selected by using $\rightarrow$ *variable selection* techniques able to produce a set of possibly reliable models, $\rightarrow$ *consensus analysis* allows the contemporary use of all the selected models.

Validation techniques are closely related to the concept of $\rightarrow$ *applicability domain* that consists in the evaluation of the reliability of a QSAR/QSPR model for the prediction of the modeled response for a new chemical.

A necessary condition for the validity of a regression model is that the multiple correlation coefficient R^2 is as close as possible to one and the standard error of the estimate s small. However, this condition (fitting ability) is not sufficient for model validity as the models give a closer fit (smaller s and larger R^2) the larger the number of parameters and variables in the models. Moreover, unfortunately, these parameters are not necessarily related to the capability of the model of making reliable predictions on future data.

Other problems for the validity of the models arise when models, often with only few variables, are obtained by using procedures based on $\rightarrow$ *variable selection* [Allen, 1971]. When a set with a large number of descriptors to select from is available, simple models can be found with apparently good fitting properties due to **chance correlation**, that is, collinearity without predictive ability [Topliss and Edwards, 1979; Wold and Dunn III, 1983; Clark and Cramer III, 1993].

To avoid models with chance correlation, a check with different validation procedures must be adopted, such as, for example, cross-validation, y-scrambling and QUICK rule.

A general validation procedure [Wold, 1991] would be the deletion of some objects before the selection of the variables: applying the variable selection procedure and then predicting the responses for excluded objects. The whole procedure, including variable selection, is then repeated a number of times, depending on the adopted specific validation technique.

In the best situation, a fairly and representative validation set of compounds (external validation set), for which predicted response values can be compared with actual ones, should be available. However, for the obvious reasons of time and cost, adequate validation sets are rarely available. As Swante Wold said "Without a real validation set, a simulated one may be better than nothing." [Wold, 1991].

With the aim of achieving a better understanding of the relationships between response and predictors, the interpretability, simplicity and comparability of a model can always add useful information about its validity.

A number of statistical techniques have been proposed to simulate the predictive ability of a model. The most popular validation techniques are listed below.

- **cross-validation**

This is the most common validation technique based on the generation of a number of modified data sets created by deleting, in each case, one or a small group of objects from the data in such a way that each object is taken away once and only once [Efron, 1983; Osten, 1988].

For each reduced data set, the model is calculated and responses for the deleted objects are predicted from the model. The squared differences between the true response and the predicted response for each object left out are added to *PRESS* ($\rightarrow$ *prediction error sum of squares*). From the final *PRESS*, the Q^2 (or R_{cv}^2) and *RMSEP* ($\rightarrow$ *root mean square error in prediction*) values are usually calculated [Cruciani, Baroni *et al.*, 1992].

The simplest and most general cross-validation procedure is the **leave-one-out technique (LOO technique)**, where each object is taken away, one at a time. In this case, given n objects, n reduced models have to be calculated. This technique is particularly important as this deletion scheme is unique and the predictive ability of the different models can be compared accurately. However, in several cases, the predictive ability obtained is too optimistic, particularly when the number of objects is quite large. This is due to a too small perturbation of the data when only one object is left out.

When the number of objects is not too small, more realistic predictive abilities are obtained deleting more than one object at each step. To apply this cross-validation procedure, called **leave-more-out technique (LMO technique)**, the number of cancellation groups is defined by the user, that is, the number of blocks the data are divided into and, at each step, all the objects belonging to a block are left out from the calculation of the model.

The number of cancellation groups G ranges from 2 to n (in the latter case, leave-more-out coincides with the leave-one-out technique). For example, given 60 objects ($n = 60$), for 2, 3, 5, 10 cancellation groups G , at each time n/G objects are left in the evaluation sets, that is, 30, 20, 12, and 6 objects, respectively.

Rules for selecting the group of objects for the evaluation set at each step must be adopted in such a way that each object is left out only one time.

- **training/evaluation set splitting**

A validation technique based on the splitting of the data set into a training set and an evaluation set. The model is calculated from the training set and the predictive power is checked on the evaluation set. The splitting is performed by randomly selecting the objects belonging to the two sets. As the results are strongly dependent on the splitting of the data, this technique is better used by repeating the splitting several hundred of times and averaging the predictive capabilities, that is, using the **repeated evaluation set technique** [Boggia, Forina *et al.*, 1997].

The **single evaluation set technique** can be used reliably only if the splitting is performed by partitioning the objects by a well-stated criterion, such as a criterion based on experimental design or cluster analysis or other deterministic approaches.

- **bootstrap**

By this validation technique, the original size of the data set (n) is preserved for the training set, by the selection of n objects with repetition; in this way, the training set contains some repeated objects and the evaluation set is constituted by the objects left out [Efron, 1982, 1987; Wehrens, Putter *et al.*, 2000]. The model is calculated on the training set and responses are predicted on the evaluation set. All the squared differences between the true response and the predicted response of the objects of the evaluation set are collected in *PRESS* ($\rightarrow$ *prediction error sum of squares*). This procedure of building training sets and evaluation sets is repeated thousands of times, *PRESS* are summed up and the average predictive power is calculated.

- **external validation**

A validation technique where, together with the training and evaluation sets, an additional external set is created to perform a further check on the predictive capabilities of a model obtained from a training set and with predictive power optimized by an evaluation set. When the number of objects is large enough, the use of an external data set for a further model validation is strongly suggested.

- **y-scrambling** ($\equiv$ *y-randomization test*)

This validation technique is adopted to check models for the presence of chance correlation, that is, models where the independent variables are randomly correlated to the response variable [Lindgren, Hansen *et al.*, 1996; Clark and Fox, 2004]. The test is performed by calculating the quality of the model (usually R^2 or, better, Q^2) randomly modifying the sequence of the response vector y , that is, by assigning to each object a response randomly selected from the true responses. If the original model has no chance correlation, there is a significant difference in the quality of the original model and that associated with models obtained with random responses. The procedure is repeated several hundred times. Variants of y-scrambling were proposed by Clark and Fox [Clark and Fox, 2004; Miller, 1990a] and Rücker *et al.* [Rücker, Rücker *et al.*, 2007].

- **lateral validation**

A technique that refers to the method of validating a new model, that is, obtained from a new data set, by comparing it with other models previously obtained for the same response [Kim, 1995b]. The similarity of the regression coefficients and the equality of their signs support the reliability of the models. The new model can also be based on different descriptors, but with the same physical meaning.

- **QUIK rule**

A rule based on the $\rightarrow$ *multivariate K correlation index* which compares the multivariate correlation index K_X of the X-block of the predictor variables with the multivariate correlation index K_{XY} obtained by the augmented X-block matrix by adding the column of the response variable [Todeschini, Consonni *et al.*, 1998]. Only regression models having multivariate correlation K_{XY} greater than multivariate correlation K_X can fulfill the QUIK rule, a necessary condition for the model validity, that is,

$$K_{XY} > K_X$$

This constraint is included among the criteria proposed by the $\rightarrow$ *RQK statistic* to obtain predictive models and prevent to taken into account models with collinearity but without prediction power, that is, chance correlation.

 Additional references are collected in the thematic bibliography (see Introduction).

■ **van der Waals excluded volume method**

A 3D-QSAR technique proposed to overcome the drawbacks of the $\rightarrow$ *grid-based QSAR techniques* such as $\rightarrow$ *CoMFA* related to the superimposition of the molecules. The descriptors produced by this method do not require $\rightarrow$ *alignment rules* and are not significantly affected by the orientation of the molecules [Tominaga and Fujiwara, 1997b; Tominaga, 1998b].

The method is based on the use of a $\rightarrow$ *probe* given by the excluded volume of two spheres with different radii and an identical center corresponding to the $\rightarrow$ *barycenter* of the molecule. The probe is layered like an onion, each layer being the excluded volume. The first sphere, that is, the component of the probe, is an atom (e.g., iodine atom with van der Waals radius of 2.05 Å, carbon atom with van der Waals radius of 1.52 Å, or hydrogen atom with van der Waals radius of 1.08 Å) whose volume defines the first layer, then 60 atoms of the same type (e.g., iodine atoms) construct the second sphere whose surface is like a fullerene and which shares the same center as the first probe. The excluded volume between the first and second spheres defines the second layer. In the same way, the subsequent spheres and layers are also defined.

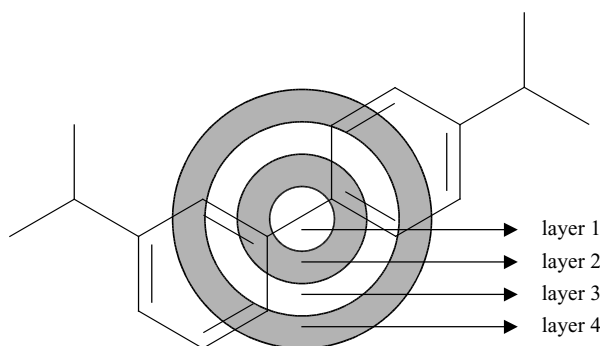


Figure V1 Layers defining the sequence of excluded volumes.

Finally, the molecular descriptors, called **EV_{WHOLE} descriptors**, are calculated as the excluded van der Waals volume between the molecule and probe layer. They represent the expansion of molecular volume in 3D space. In addition, molecular descriptors, called **EV_{TYPE} descriptors**, were also proposed as the excluded volume between a specific type of atom of the molecule and probe layer; 15 atom-types were defined for classifying molecule atoms. Using 21 layers in the probe, there is a total of 336 molecular descriptors considering both categories.

- **van der Waals interaction fields** $\equiv$ *steric interaction fields* $\rightarrow$ molecular interaction fields
- **van der Waals molecular surface** $\rightarrow$ molecular surface
- **van der Waals radius** $\rightarrow$ volume descriptors ($\odot$ van der Waals volume)
- **van der Waals volume** $\rightarrow$ volume descriptors
- **vapor pressure** $\rightarrow$ physico-chemical properties
- **variable** $\rightarrow$ data set
- **variable augmented graph-theoretical matrix** $\rightarrow$ variable descriptors
- **variable Balaban index** $\rightarrow$ variable descriptors
- **variable connectivity indices** $\rightarrow$ variable descriptors

■ variable descriptors

Variable descriptors are local and graph invariants containing adjustable parameters whose values are optimized to improve the statistical quality of a given regression model. Sometimes also called **flexible descriptors** or **optimal descriptors**, their flexibility in modeling is useful to obtain good models; however, due to the increased number of parameters needing to be optimized, they require more intensive validation procedures to generate predictive models.

These descriptors are called “variable” because their values are not fixed for a molecule but change depending on the training set and the property to be modeled.

There are different approaches to generate variable molecular descriptors.

Generalized topological indices are calculated by using common formulas of topological indices, where the exponent, if any, is allowed to differ from the standard value (e.g., $-1/2$ in the Randić connectivity index). Examples of these are the $\rightarrow$ *variable Zagreb indices*, $\rightarrow$ *generalized connectivity indices*, and $\rightarrow$ *generalized Wiener indices*.

Generalized topological indices represent a special class of variable descriptors since each value of the variable exponent generates a different topological index; however, this value is usually *a priori* chosen and not optimized during the modeling phase.

By following the approach proposed by Randić [Randić, 1991b; Randić and Basak, 2000a], variable topological indices can be easily obtained from the $\rightarrow$ *augmented adjacency matrix* or $\rightarrow$ *augmented distance matrix*, where the main diagonal elements, usually representing fixed physico-chemical properties of atoms, are replaced with variable parameters $x, y, z, \dots$ for different atom-types. A **variable augmented graph-theoretical matrix** $\mathbf{M}(x)$ is defined as

$$[\mathbf{M}(x)]_{ij} = \begin{cases} [\mathbf{M}]_{ij} & \text{if } i \neq j \\ x_i & \text{if } i = j \end{cases}$$

where $\mathbf{M}$ is any graph theoretical matrix with diagonal elements equal to zero and x_i is the variable representing the i th atom; $[\mathbf{M}]_{ij}$ are the off-diagonal elements of the matrix $\mathbf{M}$. On the matrix diagonal, there are as many different variables ($x, y, z, \dots$) as the different atom-types in the molecule.

The row sum of the variable augmented matrix is a local vertex invariant containing the variable x as

$$VS_i(\mathbf{M}(x)) = \sum_{j=1}^A [\mathbf{M}(x)]_{ij} = VS_i(\mathbf{M}) + x_i$$

where VS_i stands for the $\rightarrow$ *vertex sum operator* of the i th vertex and A is the number of vertices in the graph. Then, the row sum of any variable augmented matrix is simply the vertex sum of the corresponding matrix $\mathbf{M}$ plus the variable parameter x characterizing the considered atom.

From variable augmented matrices, several molecular descriptors, containing variable parameters, are calculated by applying the common matrix operators [Randić, Mills *et al.*, 2000; Randić and Pompe, 2001a; Randić, 2001e; Randić, Plavšić *et al.*, 2001]. The optimal parameter values are searched for to reach the best regression model quality.

This approach was applied to derive **variable connectivity indices**, denoted as ${}^m\chi^f$ and calculated by using the formula of $\rightarrow$ *Kier–Hall connectivity indices*, where the classic $\rightarrow$ *vertex degrees* are replaced by the row sums of the variable augmented adjacency matrix [Randić, Mills *et al.*, 2000; Randić and Pompe, 2001b; Randić, 2001e]:

$${}^m\chi_t^f = \sum_{k=1}^K \left(\prod_{i=1}^n (\delta_i + x_i) \right)_k^{-1/2}$$

where δ_i is the vertex degree of the i th atom, x_i is the variable parameter used to characterize the atom-type of the i th atom; k runs over all of the m th order subgraphs constituted by n vertices and m edges; K is the total number of m th order subgraphs present in the molecular graph; the subscript “ t ” refers to the type of $\rightarrow$ *molecular subgraph* and *ch* is for chain or ring, *pc* for path-cluster, *c* for cluster, and *p* for path (that can also be omitted).

Some applications of variable connectivity indices found in literature are [Randić, Basak *et al.*, 2001; Randić and Basak, 2001c; Randić, Plavšić *et al.*, 2001; Randić and Pompe, 2001a; Zefirov and Palyulin, 2001; Kezele, Klasinc *et al.*, 2002; Li, Hu *et al.*, 2004; Peng, Fang *et al.*, 2004; Pompe, Veber *et al.*, 2004; Randić, Pompe *et al.*, 2004; Zhong, He *et al.*, 2004; Janežič, Lučić *et al.*, 2006; Pompe and Randić, 2007].

With the purpose to obtain variable connectivity indices represented by positive real numbers, the atomic weights x and y have to be varied in such a way that the row sums of the augmented adjacency matrix remain positive ($\delta_i^f = \delta_i + x_i > 0$); in this way, the influence of individual atoms and bonds is in the range from zero to plus infinite. However, as a consequence of the fact that atom or bond contributions in variable connectivity indices are always positive unless

suitably modified, they cannot model such cases in which there is a negative region of influence, which is named *anticonnectivity region* [Pompe, 2005; Pompe and Randić, 2006]. To overcome this drawback, the **anticonnectivity indices** were defined by considering also negative subgraph contributions; these variable indices are formally defined as

$${}^m\chi_t^f = \sum_{k=1}^K \left(\pm \prod_{i=1}^n (\delta_i + x_i) \right)_k^{-1/2}$$

Whether an atom or a bond or, in general, a subgraph comprising of m bonds, makes positive or negative contributions is found during the optimization procedure.

The **variable Balaban index**, denoted as J^* , is calculated by analogy with the $\rightarrow$ *Balaban distance connectivity index* from the row sums of the variable augmented distance matrix as [Randić and Pompe, 2001b]

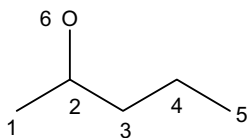
$$J^* = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot [(\sigma_i + x_i) \cdot (\sigma_j + x_j)]^{-1/2}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to 1 for pairs of adjacent vertices, and zero otherwise; B is the number of edges in the molecular graph, C is the $\rightarrow$ *cyclomatic number*; σ_i and σ_j are the $\rightarrow$ *distance degree* of vertices v_i and v_j ; x_i and x_j are the variable parameters for the atom-types of the two vertices v_i and v_j .

Other variable descriptors were derived from different variable augmented graph-theoretical matrices, [Randić and Pompe, 2001b].

Example V1

Variable connectivity index, variable Balaban index, and variable Zagreb indices for 2-pentanol. ${}^a\mathbf{A}(x, y)$ and ${}^a\mathbf{D}(x, y)$ are the variable augmented adjacency matrix and the variable augmented distance matrix, respectively. VS_i indicates the matrix row sums; x and y are the variable parameters for carbon and oxygen atom, respectively.



Augmented adjacency matrix

$${}^a\mathbf{A}(x, y) = \begin{array}{c|cccccc|c} \text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 & VS_i \\ \hline 1 & x & 1 & 0 & 0 & 0 & 0 & 1+x \\ 2 & 1 & x & 1 & 0 & 0 & 1 & 3+x \\ 3 & 0 & 0 & x & 1 & 0 & 0 & 2+x \\ 4 & 0 & 0 & 1 & x & 1 & 0 & 2+x \\ 5 & 0 & 0 & 0 & 1 & x & 0 & 1+x \\ 6 & 0 & 1 & 0 & 0 & 0 & y & 1+y \end{array}$$

Augmented distance matrix

$${}^a\mathbf{D}(x, y) = \begin{array}{c|cccccc|c} \text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 & VS_j \\ \hline 1 & x & 1 & 2 & 3 & 4 & 2 & 12+x \\ 2 & 1 & x & 1 & 2 & 3 & 1 & 8+x \\ 3 & 2 & 1 & x & 1 & 2 & 2 & 8+x \\ 4 & 3 & 2 & 1 & x & 1 & 3 & 10+x \\ 5 & 4 & 3 & 2 & 1 & x & 4 & 14+x \\ 6 & 2 & 1 & 2 & 3 & 4 & y & 21+y \end{array}$$

The variable connectivity index is then defined as

$${}^1\chi^f = \frac{1}{\sqrt{(1+x) \cdot (3+x)}} + \frac{1}{\sqrt{(3+x) \cdot (2+x)}} + \frac{1}{\sqrt{(2+x) \cdot (2+x)}} + \frac{1}{\sqrt{(2+x) \cdot (1+x)}} + \frac{1}{\sqrt{(3+x) \cdot (1+\gamma)}}$$

The variable Balaban index is then defined as

$$J^* = \frac{5}{0+1} \cdot \left[\frac{1}{\sqrt{(12+x) \cdot (8+x)}} + \frac{1}{\sqrt{(8+x) \cdot (8+x)}} + \frac{1}{\sqrt{(8+x) \cdot (10+x)}} + \frac{1}{\sqrt{(10+x) \cdot (14+x)}} + \frac{1}{\sqrt{(8+x) \cdot (12+\gamma)}} \right]$$

The first variable Zagreb index is then defined as

$$M_1^f = 2 \cdot (1+x)^2 + (3+x)^2 + 2 \cdot (2+x)^2 + (1+\gamma)^2$$

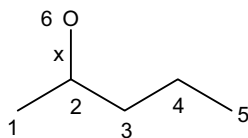
The second variable Zagreb index is then defined as

$$M_2^f = (1+x) \cdot (3+x) + (3+x) \cdot (2+x) + (2+x) \cdot (2+x) + (2+x) \cdot (1+x) + (3+x) \cdot (1+\gamma)$$

Other variable topological indices are **variable path counts**, which are $\rightarrow$ *weighted path counts* obtained by weighting the graph edges involving heteroatoms with one or more variable parameters [Randić and Basak, 1999, 2000b; Randić and Pompe, 1999; Amić, Basak *et al.*, 2002].

Example V2

Variable path counts of 2-pentanol. The weight of the edge representing C—O bond is taken to be x , while the weights of the edges representing C—C bonds are equal to 1. Each path involving the C—O bond gives a contribution of x to the path count.



Atom	${}^1p_i$	${}^2p_i$	${}^3p_i$	${}^4p_i$
1	1	$1+x$	1	1
2	$2+x$	1	1	0
3	2	$2+x$	0	0
4	2	1	$1+x$	0
5	1	1	1	$1+x$
6	x	$2x$	x	x
Variable path counts	$4+x$	$3+2x$	$2+x$	$1+x$

The **External Factor Variable Connectivity Indices** (EFVCI) are variable descriptors in which the atomic attribute is divided into two parts: the innate part and the external part or perturbation term [Hu, Liang *et al.*, 2003b, 2004]. The innate part is defined in terms of the number of valence electrons, while the perturbation term by reciprocal square distances and a variable parameter x . The local vertex invariant relative to the i th atom is calculated as

$$\gamma_i = Z_i^v + VS_i(\mathbf{D}^{-2}) \cdot x$$

where Z^v is the number of valence electrons and VS_i is the i th row sum of the $\rightarrow$ *reciprocal square distance matrix* $\mathbf{D}^{-2}$. Then, the External Factor Variable Connectivity Indices are calculated by using the variable local vertex invariants γ_i in place of the classic vertex degree δ_i in the formula of the $\rightarrow$ *Kier–Hall connectivity indices*.

To the class of variable descriptors also belong the $\rightarrow$ *semiempirical molecular connectivity terms* proposed by Pogliani [Pogliani, 1997a] that are $\rightarrow$ *combined descriptors* that also account for a variable parameter.

Other variable descriptors are generated by a systematic approach proposed by Toropov in 1999 and initially called **Correlation Weights of the Local Invariants of Molecular Graphs** (CWLIMG) [Mercader, Castro *et al.*, 2000; Toropova and Toropov, 2000, 2001b; Krenkel, Castro *et al.*, 2001a, 2001b; Peruzzo, Marino *et al.*, 2001] and later renamed **Optimization of Correlation Weights of Local Invariants** (OCWLI) [Toropov and Toropova, 2002a].

By this approach, a family of flexible molecular descriptors based on different mathematical functions obtained from the molecular graph G is defined. The basic idea is to vary the **correlation weights** CW of the different atom-types under consideration, aimed at obtaining an as high as possible correlation coefficient between experimental and calculated values of a selected molecular property. Atom-types are defined using chemical information and $\rightarrow$ *local vertex invariants*. The general form of the molecular descriptor $\mathcal{D}$ in terms of a number of selected atom-types is

$$\mathcal{D} = f\{\text{CW}(\mathcal{L}_1), \text{CW}(\mathcal{L}_2), \dots\}$$

where $\mathcal{L}$ is an atomic property or local vertex invariant used to define the different atom-types present in the molecules of the data set; for instance, atoms can be distinguished on the basis of their chemical element or the different values of any local vertex invariant, such as the $\rightarrow$ *vertex degree*, the $\rightarrow$ *extended connectivity* and the $\rightarrow$ *Nearest Neighboring Code*. $\text{CW}(\mathcal{L})$ represents the contribution of each atom-type to the molecular descriptor value and f is any function that relates the different atom-type contributions to the whole molecular descriptor. There are as many correlation weights as the different atom-types defined by the selected local invariants and the molecules present in the data set. The correlation weights are estimated by an optimization procedure (e.g., Monte Carlo technique) applied to the training set molecules to yield the best correlation with the studied property.

An example of molecular descriptors defined in terms of correlation weights is [Krenkel, Castro *et al.*, 2001a]

$$\mathcal{D} = f\{\text{CW}(\delta), \text{CW}(\text{EC}^1)\}$$

where the different atom-types are defined by the different values of the $\rightarrow$ *vertex degree* δ and the $\rightarrow$ *extended connectivity* EC^1 assumed by the molecules in the training set. The function f is one of the common functions used for the calculation of several topological molecular

descriptors $\mathcal{D}$ from local vertex invariants:

$$\begin{aligned}\mathcal{D} &= \sum_{i=1}^A [\text{CW}(\delta_i) + \text{CW}(\text{EC}_i^1)] \\ \mathcal{D} &= \sum_{i=1}^A [\text{CW}(\delta_i) \cdot \text{CW}(\text{EC}_i^1)] \\ \mathcal{D} &= \prod_{i=1}^A [\text{CW}(\delta_i) + \text{CW}(\text{EC}_i^1)] \\ \mathcal{D} &= \prod_{i=1}^A [\text{CW}(\delta_i) \cdot \text{CW}(\text{EC}_i^1)] \\ \mathcal{D} &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot [\text{CW}(\delta_i) \cdot \text{CW}(\text{EC}_i^1) + \text{CW}(\delta_j) \cdot \text{CW}(\text{EC}_j^1)]\end{aligned}$$

where A is the number of graph vertices and a_{ij} are the elements of the $\rightarrow$ adjacency matrix equal to 1 for pairs of adjacent vertices, and zero otherwise.

This approach can be applied both to hydrogen-filled molecular graph and $\rightarrow$ Graph of Atomic Orbitals (GAO). In the first case, types of vertices are the different chemical elements, while in the GAO, types of vertices are the types of atomic orbitals.

Applications of OCWLI approach are [Krenkel, Castro *et al.*, 2001a; Toropov and Toropova, 2001a, 2001b, 2001c, 2002, 2002a, 2002b, 2002c, 2003, 2004; Toropova and Toropov, 2001, Duchowicz, Castro *et al.*, 2002; Toropov and Schultz, 2003; Toropov, Nesterov *et al.*, 2003a, 2003b; Toropov, Duchowicz *et al.*, 2003; Toropov, Toropova *et al.*, 2003, 2005; Duchowicz, Castro *et al.*, 2004, 2005; Toropov and Roy, 2004, 2005; Toropov and Benfenati, 2004a, 2004b; Costescu, Moldovan *et al.*, 2006; Toropov, Rasulev *et al.*, 2007].

The GTI-simplex approach is a general strategy to search for optimized quantitative-structure property relationship models based on variable topological indices, called **Estrada Generalized Topological Indices (GTI)**, and the down hill simplex optimization procedure [Estrada, 2001, 2003b, 2004c, 2005b; Estrada and Gutierrez, 2001; Matamala and Estrada, 2005a, 2005b, 2007]. The main objective of this approach is to obtain the best optimized molecular descriptors for each property under study. The family of GTI-simplex descriptors is comprised of $\rightarrow$ autocorrelation descriptors defined by the following general form:

$$\text{GTI} = \sum_{k=1}^D C_k(x_0, p_0) \cdot \eta^{(k)}$$

where the summation goes over the different topological distances in the graph, D being the $\rightarrow$ topological diameter, and accounts for the contributions $\eta^{(k)}$ of pairs of vertices located at the same topological distance k . Each contribution $\eta^{(k)}$ is scaled by two real parameters x_0 and p_0 through the $C_k(x_0, p_0)$ coefficient defined as

$$C_k(x_0, p_0) = k^{p_0} \cdot x_0^{p_0(k-1)}$$

By definition, the C_k coefficient is equal to 1 for any pair of adjacent vertices ($k = 1$), regardless of the parameter values. Note that the coefficients C_k are the elements of the so-called **generalized molecular-graph matrix Γ** , which is a square symmetric $A \times A$ matrix, defined as

[Estrada, 2003b, 2004c]

$$[\Gamma(x_0, p_0)]_{ij} \equiv g_{ij}(x_0, p_0) = \begin{cases} 1 & \text{if } d_{ij} = 1 \\ [d_{ij} \cdot x_0^{(d_{ij}-1)}]^{p_0} & \text{if } i \neq j \wedge d_{ij} > 1 \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} is the topological distance between vertices v_i and v_j . This matrix was also defined in terms of interatomic geometric distances.

The term $\eta^{(k)}$ defines the contribution of all those interactions due to the pairs of vertices at distance k in the graph as

$$\eta^{(k)} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \langle i, j \rangle \cdot [{}^k\mathbf{B}]_{ij}$$

where A is the number of vertices in the graph, ${}^k\mathbf{B}$ is the $\rightarrow$ *geodesic matrix* of order k , whose elements are equal to 1 for pairs of vertices at distance k , and 0 otherwise. The term $\langle i, j \rangle$ is the ‘geodesic-bracket’ term encoding information about the molecular shape on the basis of a connectivity-like formula as

$$\langle i, j \rangle = \frac{1}{2} \cdot (u_i \cdot v_j + v_i \cdot u_j)$$

where u and v are two functions of the variable parameters x and p and can be considered as generalized vertex degrees defined as

$$u_i(x_1, p_1, \mathbf{w}) = \left[w_i + \delta_i + \sum_{k=2}^D k \cdot x_1^{k-1} \cdot {}^kf_i \right]^{p_1} \quad v_i(x_2, p_2, \mathbf{s}) = \left[s_i + \delta_i + \sum_{k=2}^D k \cdot x_2^{k-1} \cdot {}^kf_i \right]^{p_2}$$

where δ_i is the simple $\rightarrow$ *vertex degree* of the i th vertex and ${}^kf_i$ is its $\rightarrow$ *vertex distance count*, that is, the number of vertices at distance k from the i th vertex. The scalars $x_0, x_1, x_2, p_0, p_1, p_2, \mathbf{w}$ and $\mathbf{s}$ define a $(2A + 6)$ -dimensional real space of parameters; $\mathbf{w}$ and $\mathbf{s}$ are two A -dimensional vectors collecting atomic properties. The first six parameters x_0, x_1, x_2, p_0, p_1 , and p_2 are free parameters to be optimized, whereas the parameters $\mathbf{w}$ and $\mathbf{s}$ are predefined quantities used to distinguish among the different atom-types. For each combination of the possible values of these parameters, a different topological index is obtained for a molecule. It has to be noted that several of the well-known topological indices can be calculated by the *GTI* formula by settling specific combinations of the parameters; for instance, for $\mathbf{w} = (0, 0, \dots, 0)$ and $\mathbf{s} = (0, 0, \dots, 0)$, the index *GTI* reduces to the $\rightarrow$ *Wiener index* when $x_0 = 1, x_1 = \text{any}, x_2 = \text{any}, p_0 = 1, p_1 = 0, p_2 = 0$, while *GTI* coincides with the $\rightarrow$ *Randić connectivity index* when $x_0 = 0, x_1 = 0, x_2 = 0, p_0 = 1, p_1 = -1/2, p_2 = -1/2$.

- **variable importance for projection** $\equiv$ *VIP score* $\rightarrow$ variable selection
- **variable inflation factor** $\rightarrow$ variable reduction
- **variable path counts** $\rightarrow$ variable descriptors

■ **variable reduction** ($\equiv$ *feature reduction*)

Variable reduction consists in the selection of a subset of variables able to preserve the essential information contained in the whole $\rightarrow$ *data set*, but eliminating redundancy, too highly intercorrelated variables, and so on.

Variable reduction differs from $\rightarrow$ *variable selection* in the fact that the subset of descriptor variables is selected independently of the response variable of interest.

The most common methods for variable reduction are listed below.

- **constant and near-constant variables**

A preliminary approach to variable reduction consisting in the elimination of all the variables taking the same value for all the objects in the data set. Near-constant variables, that is, variables that assume the same value except in one or very few cases, would also be excluded. A good measure for evaluating near-constant variables is the $\rightarrow$ *standardized Shannon's entropy*: the entropy of a variable with one different value over 10 objects is 0.141, over 20 objects is 0.066, with two different values over 100 objects is 0.024.

- **pair correlation cutoff selection**

This method is based on the *correlation matrix* and on a selected correlation cutoff value (e.g., $r^* = 0.95$). For each pair of variables with a correlation value greater than the cutoff value, one of the two correlated variables is arbitrarily eliminated. The variable that has to be eliminated is chosen arbitrarily or by selecting the variable that shows the largest correlation sum with all the other variables in the variable set.

- **variable inflation factor (VIF)**

The variable inflation factor of the j th variable is evaluated by the $\rightarrow$ *coefficient of determination* R_j^2 obtained by considering the j th variable as the response variable and excluding it from the dependent variables. It is defined as

$$VIF_j = \frac{1}{1 - R_j^2}$$

Variables showing a *VIF* value greater than a predefined cutoff value (often 5, 10) are excluded.

- **cluster analysis feature selection**

Methods of $\rightarrow$ *cluster analysis* are applied to the variables, on the so-called *Q-mode data matrix*, that is, on the transposed data matrix. Once cluster analysis has been performed, one (or two) variable(s) for each cluster is retained as representative of all the variables within that cluster. Which and how many are the retained/excluded variables depends on the chosen cluster analysis method.

- **principal component analysis feature selection**

Visual inspection of the significant loading plots obtained by $\rightarrow$ *Principal Component Analysis* can be a nonquantitative but useful tool to select the most relevant variables to preserve the most important information contained in the original data [Jolliffe, 1986; Jackson, 1991].

More quantitative approaches based on PCA have also been defined, based on the largest loadings in absolute value of the eigenvectors, usually obtained from the correlation matrix.

Jolliffe techniques of variable reduction [Jolliffe, 1972; Jolliffe, 1973] exploit the association of the original variables with the eigenvectors (PCs), usually obtained from the correlation matrix. The criterion of these techniques is to keep as much variance of the data in the subset of selected variables. The Jolliffe technique B2 associates one variable with each of the last $p - M$ eigenvectors and deletes those $p - M$ variables with the largest coefficients in the $p - M$

eigenvectors. M is the number of selected significant components (e.g. with eigenvalues larger than 0.7).

The Jolliffe technique B1 is the iterative version of technique B2, that is, after deletion of $p - M$ variables, a PCA is newly performed on the remaining M variables, and a further set of variables is deleted; then the procedure is repeated until the last eigenvalue assumes some significant value.

The Jolliffe technique B3 calculates the sum of squares of the correlation coefficients in the last $p - M$ PCs for each variable:

$$\max_j \left(\sum_{m=M+1}^p \lambda_m \cdot \ell_{jm}^2 \right)$$

where M is the number of significant components, p the total number of variables, λ_m the eigenvalue and ℓ_{jm}^2 the squared loading of the j th variable in the m th eigenvector. The variables discarded are those that maximize this sum of squares.

The Jolliffe technique B4 is similar to the B2, but the procedure is applied starting from the M eigenvectors corresponding to the largest eigenvalues and retaining the variables showing the largest absolute loadings.

McCabe techniques of variable reduction are based on the calculation of the conditional covariance (or correlation) matrix of the excluded variables {McCabe, 1984 1410 /id}. This matrix represents the residual information left in the variable that are not selected, after the effect of the most relevant variables has been removed. It is a square symmetric matrix of order $q = p - k$, where p is the total number of variables and k is the number of retained variables, and is derived from the covariance (correlation) matrix of the retained variables $\mathbf{S}_R$ (of size $k \times k$), the covariance (correlation) matrix of the deleted variables $\mathbf{S}_D$ (of size $q \times q$), and the cross-covariance (correlation) matrix between the two sets of variables $\mathbf{S}_{RD}$ (of size $k \times q$):

$$\mathbf{S}_{D,R} = \mathbf{S}_D - \mathbf{S}_{RD}^T \cdot \mathbf{S}_R^{-1} \cdot \mathbf{S}_{RD}$$

From the conditional correlation matrix $\mathbf{S}_{D,R}$ the q eigenvalues are calculated. Finally, the set of k variables, called *principal variables*, satisfying one of the four criteria are retained:

$$\begin{aligned} \text{(a)} \quad & \min_{\{k\}} \left(\sum_{m=1}^q \lambda_m \right) \\ \text{(b)} \quad & \min_{\{k\}} \left(\prod_{m=1}^q \lambda_m \right) \\ \text{(c)} \quad & \min_{\{k\}} \left(\sum_{m=1}^q \lambda_m^2 \right) \\ \text{(d)} \quad & \max_{\{k\}} \left(\sum_{m=1}^{\min(k,q)} \rho_m^2 \right) \end{aligned}$$

where λ_m are the eigenvalues of the conditional covariance (or correlation) matrix of the set of the q deleted variables, given the values of the k selected variables; ρ_m are the canonical

correlations between the set of q deleted variables and the set of k selected variables; the summation runs over the minimum between k and q .

• K correlation analysis

A method consisting in the iterative procedure of the elimination of one variable at a time, based on the $\rightarrow$ *multivariate K correlation index* [Todeschini, 1997; Todeschini, Consonni *et al.*, 1998]. All the variables are removed one at a time and the K multivariate correlation $K_{p/j}$ ($j = 1, p$) of the set of $p - 1$ variables is calculated. The j th variable associated with the minimum $K_{p/j}$ correlation is removed from the set of p variables (i.e., the variable that is maximally correlated with all the other) and the procedure is repeated on the remaining variables. The elimination procedure ends when the minimum $K_{p/j}$ is greater than the correlation K of the whole set of the remaining variables or when a standardized K correlation value, called **K inflation factor (KIF)**, is less than 0.6–0.5.

■ Variable Selection (VS) ($\equiv$ Feature Selection, Variable Subset Selection, VSS)

The aim of the variable subset selection is to reach optimal model complexity in predicting a response variable by a reduced set of independent variables [Hocking, 1976; Topliss and Edwards, 1979; Miller, 1990a; Wikel and Dow, 1993; Tetko, Villa *et al.*, 1996]. Regression (and classification) models based on optimal subsets of a few predictor variables have several advantages. In effect, simple models show more stable statistical properties, can be more easily interpreted, and can give higher predictive power. On the other hand, the major drawback of the variable subset selection procedures consists in the possibility of selecting model variables having $\rightarrow$ *chance correlation*. Thus, to avoid chance correlation, particular attention must be paid to $\rightarrow$ *validation techniques*.

Within the framework of the variable subset selection techniques, a vector $\mathbf{I}^*$ is usually defined, which collects binary variables indicating the presence of the j th variable ($I_j^* = 1$) or its absence ($I_j^* = 0$) in the final model. This p -dimensional vector $\mathbf{I}^*$ is usually obtained by validation. The actual model dimension k ($k \leq p$) is the sum of all the entries equal to 1 in the $\mathbf{I}^*$ vector.

The methods for variable subset selection can be divided into two main categories: methods which work on the original p variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$, *direct variable subset selection*, where the best binary vector $\mathbf{I}^*$ is evaluated by considering the relationships among the $\mathbf{x}_j$ variables and a response variable $\mathbf{y}$, that is,

$$\mathbf{I}^* = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p; \mathbf{y})$$

and methods which work on linear combinations $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_M$ of the original variables, *indirect variable subset selection*, where the best binary vector $\mathbf{I}^*$ is evaluated by considering the weights of the linear combinations $\mathbf{t}_m$ and the response variable $\mathbf{y}$, that is,

$$\mathbf{I}^* = f[(w_{11}, w_{21}, \dots, w_{p1}), (w_{12}, w_{22}, \dots, w_{p2}), \dots, (w_{1M}, w_{2M}, \dots, w_{pM}); \mathbf{y}]$$

where M is the number of significant linear combinations and w_{jm} the weights of the j th variable in the m th component. These new variables can be obtained by $\rightarrow$ *Principal Component Analysis* (PCA), Partial Least Square regression (PLS), or other related techniques. Among these, variable selection of wavelet coefficients calculated by $\rightarrow$ *wavelet transforms* of the original variables was also proposed both for classification and regression problems [Jouan-Rimbaud, Walczak *et al.*, 1997; Alsberg, Woodward *et al.*, 1998; Depczynski, Jetter *et al.*, 1999a; Wold, Trygg *et al.*, 2001; Cocchi, Corbellini *et al.*, 2005].

The most common methods for variable selection are listed below.

- **all possible models**

It is the simplest method based on an exhaustive examination of all the possible models of k variables (i.e., the model size) obtained by a set of p variables, where k is a parameter between 1 and a user-defined value (using ordinary least square regression, the maximum theoretical value is limited by the number n of objects in the data set). The best model(s) is evaluated by validation or by parameters depending on the model degrees of freedom.

As the procedure consists in the evaluation of the quality of all the models with one x variable (i.e., p univariate models), of all the models with two variables (i.e., $p \times (p - 1)$ bivariate models), until all the possible models with L variables, the greatest disadvantage of this method is the extraordinary increase in the required computer time when p and k are quite large. The total number t of models is given by the relationship:

$$t = \sum_{k=1}^L \frac{p!}{k! \cdot (p-k)!} \leq 2^p - 1$$

where L is the maximum user-defined model size. For example, with 32 total variables ($p = 32$) and $1 \leq k \leq 8$ ($L = 8$), $t = 1 \times 10^7$. The maximum number of possible models is $2^p - 1 = 4\,294\,967\,295 \approx 4 \times 10^9$. The main advantage of this method is the exhaustive search for the best model in the model space.

- **response-variable correlation cutoff**

Variable selection is performed by checking the $\rightarrow$ *coefficient of determination* R^2 or the corresponding cross-validated quantity Q^2 from univariate regression models $y = b_0 + b_1 x_j$, the selection being made separately for each j th variable of the p variables $x_1, x_2, \dots, x_p$.

Variables showing R^2 or Q^2 values lower than a cut-off value, that is, evidently uncorrelated with the y response, are definitively excluded from the searching for the best models. Acceptable cutoff values for R^2 are usually between 0.05 and 0.01 and for Q^2 less than zero.

This procedure can be used to perform a preliminary screening of the variables, while other techniques are used on the remaining variables, that is, genetic algorithm variable subset selection or searching for all subset models. A disadvantage of this procedure is that a variable with a low correlation with the response, but correlated with residuals and thus able to give a contribution in the final model, is excluded.

- **stepwise regression methods (SWR)**

Commonly used regression methods proposed to evaluate only a small number of subsets by either adding or deleting variables one at a time according to a specific criterion [Draper and Smith, 1998].

Forward selection (FS-SWR) is a technique starting with no variables in the model and adding one variable at a time until either all variables are entered or a stopping criterion is satisfied.

The variable considered for inclusion at any step is the one yielding the largest single degree of freedom F -ratio among the variables eligible for inclusion and this value is larger than a fixed value F_{in} . At each step, the j th variable is added to a k -size model if

$$F_j = \max_j \left(\frac{RSS_k - RSS_{k+j}}{s_{k+j}^2} \right) > F_{in}$$

when RSS is the $\rightarrow$ residual sum of squares and s^2 the $\rightarrow$ mean square error. The subscript $k + j$ refers to quantities computed when the j th variable is added to the current k variables already in the model.

Backward elimination (BE-SWR) is a stepwise technique starting with a model in which all the variables are included and then deleting one variable at a time. At any step, the variable with the smallest F -ratio is eliminated if this F -ratio does not exceed a specified value F_{out} . Then, at each step, the j th variable is eliminated from a k -size model if

$$F_j = \min_j \left(\frac{RSS_{k-j} - RSS_k}{s_k^2} \right) < F_{out}$$

when RSS is the $\rightarrow$ residual sum of squares and s^2 the $\rightarrow$ mean square error.

The subscript $k - j$ refers to quantities computed when the j th variable is excluded from the current k variables in the model. Obviously, when the total number of variables is greater than the number of objects, this technique cannot be applied in this form.

The most popular stepwise technique combines the two previous approaches (FW and BE) and is called *Elimination-Selection* (ES-SWR) [Efroymson, 1960]. It is basically a forward selection but at each step (when the number of model variables is greater than two) the possibility of deleting a variable as in the BE approach is considered.

The two major drawbacks of the stepwise procedures are that none of them ensure that the “best” subset of a given size is found and, perhaps more critical, it is not uncommon that the first variable included in FS becomes unnecessary in the presence of other variables.

To avoid some drawbacks of the stepwise approaches, the *i-fold stepwise variable selection* was later proposed [Lučić, Trinajstić *et al.*, 1999]. This technique is based on the descriptor orthogonalization and, at each subsequent step, adds the set of the best i descriptors.

📖 [Whitley, Ford *et al.*, 2000]

• Genetic Algorithm–Variable Subset Selection (GA–VSS)

Variable selection is performed by using *Genetic Algorithms* (GA), based on the evolution of a population of models. In genetic algorithm terminology, the binary vector $\mathbf{I}$ is called *chromosome*, which is a p -dimensional vector where each position (a *gene*) corresponds to a variable (1 if included in the model, 0 otherwise). Each chromosome represents a model with a subset of variables [Goldberg, 1989; Leardi, Boggia *et al.*, 1992; Leardi, 1994; Luke, 1994; Leardi, 2001].

The statistical parameter to be optimized must be defined (e.g., maximizing Q^2 by a leave-one-out validation procedure), together with the model population size P (for example, $P = 100$) and the maximum number L of allowed variables in a model (for example, $L = 5$); the minimum number of allowed variables is usually assumed equal to 1. Moreover, a *crossover probability* p_C (usually high, for example, $p_C > 0.9$) and a *mutation probability* p_M (usually small, for example, $p_M < 0.1$) must also be defined by the user.

Once the leading parameters are defined, the genetic algorithm evolution starts based on three main steps:

(1) Random initialization of the population

The model population is initially built by random models with a number of variables between 1 and L , and the models are ordered with respect to the selected statistical parameter – the quality of the model (the best model is in first place, the worst model at position P). This step is performed only once.

(2) Crossover step

From the population, pairs of models are selected (randomly or with a probability proportional to their quality). Then, for each pair of models the common characteristics are preserved (i.e., variables excluded in both models remain excluded, variables included in both models remain included). For variables included in a model and excluded from the other, a random number is tried and compared with the crossover probability p_C : if the random number is lower than the crossover probability, the excluded variable is included in the model and vice versa. Finally, the statistical parameter for the new model is calculated: if the parameter value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank; otherwise, it is not longer considered. This procedure is repeated for several pairs (for example 100 times).

(3) Mutation step

For each model present in the population (i.e., each chromosome), p random numbers are tried and one at a time each is compared with the defined mutation probability p_M : each gene remains unchanged if the corresponding random number exceeds the mutation probability, otherwise, it is changed from zero to one or vice versa. Low values of p_M allow only a few mutations, thus obtaining new chromosomes not too different from the generating chromosome.

Once the mutated model is obtained, the statistical parameter for the model is calculated: if the parameter value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank; otherwise, it is not longer considered. This procedure is repeated for all the chromosomes (i.e., P times).

(4) Stop conditions

The second and third steps are repeated until some stop condition is encountered (e.g., a user-defined maximum number of iterations) or the process is arbitrarily ended.

An important characteristic of the GA-VSS method is that a single model is not necessarily obtained but the result usually is a population of acceptable models; this characteristic, sometimes considered a disadvantage, provides an opportunity to make an evaluation of the relationships with the response from different points of view. A theoretical disadvantage is that the absolute best model could be not present in the final population. However, after a careful selection of the best models, $\rightarrow$ *consensus analysis* can be performed contemporarily using the selected models and estimating the response as weighted average of the responses of the single models.

📖 [Mestres and Scuseria, 1995; Devillers, 1996a; Hopfinger and Patel, 1996; Judson, 1996; Kemsley, 1998; Wehrens, Pretsch *et al.*, 1999; Xue and Bajorath, 2000; Cho and Hermsmeier, 2002; Schmidt and Heilmann, 2002; Barakat, Jiang *et al.*, 2004; Pednekar, Kelkar *et al.*, 2004; Schefzik and Bradley, 2004; Todeschini, 2004; Esteban-Diez, González-Sáiz *et al.*, 2006b; Kang, Choi *et al.*, 2007; Mandal, Johnson *et al.*, 2007]

• Genetic Function Approximation (GFA)

This is a variable subset selection algorithm [Rogers and Hopfinger, 1994; Rogers, 1995] that combines genetic algorithms [Holland, 1975] with Friedman's Multivariate Adaptive Regression Splines (MARS) algorithm [Friedman, 1988]. MARS uses splines as basis functions to partition data space as it builds its regression models. The searching for spline-based regression models is improved using a genetic algorithm rather than the original incremental approach.

GFA algorithm is derived from the G/SPLINES algorithm [Rogers, 1991, 1992], it automatically selects which variables are to be used in its basis functions, and determines the appropriate number of basis functions.

The initial models are generated by randomly selecting a number of variables from the training set, building basis functions from these variables using user-specified basis function types, and then building the models from random sequences of these basis functions. Improved models are then constructed by performing the genetic crossover operation to recombine the terms of the better performing models.

Model performances are evaluated by using **Friedman's lack-of-fit function (LOF)** [Friedman, 1988], defined as

$$\text{LOF} = \frac{LSE}{\left(1 - \frac{c + d \cdot p}{n}\right)^2}$$

where LSE is the least-squares error, c the number of basis functions, p the number of variables, n the number of objects, and d a user-defined smoothing parameter.

Using the information contained in the binary vector $\mathbf{I}^*$ collecting the selected variables, it is possible to calculate the regression coefficients by ordinary least squares regression.

📖 [Shi, Fan *et al.*, 1998]

• **M**utation and **S**election Uncover Models (MUSEUM)

A variable selection approach based on genetic algorithms, and, more specifically, on an evolutionary algorithm, mutation being the preferred genetic operation [Kubinyi, 1994a, 1994b, 1996].

After a random generation of a model population, random mutations of one or a few variables are tried. If better models are obtained after a fixed number of trials, this procedure is repeated on the new generation of models, otherwise random mutations of several variables become allowed. Also in this case, if better models are obtained after a fixed number of trials, the first procedure is repeated on the new generation of models, otherwise systematic addition and elimination of the variables of the population models is performed. If better models are obtained, the procedure restarts from the first step, otherwise, if all the variables are checked, the procedure ends and the variables of the final population models are checked for their statistical significance and eventually eliminated. MUSEUM approach evaluates the quality of the models maximizing the $\rightarrow$ *Kubinyi fitness function*.

• **s**equential **s**earch

A variable selection method proposed to decrease the huge number of models that must be evaluated by all subset model searching [Miller, 1990a].

The statistical parameter to optimize (e.g., maximizing Q^2 by a leave-one-out validation procedure) must be selected and the maximum number L of allowed variables in a model (e.g., $L = 5$). The minimum number of allowed variables is usually taken equal to 1. For each size, a small group of models is selected by a preliminary evaluation of the statistical parameter to optimize.

For each model, the sequential search is based on the following procedure:

- a) Each variable present in the model is excluded, one at a time, and is replaced by all the excluded variables, one at a time. For each replacement, the value of the statistical parameter

to optimize is calculated and stored. If k is the number of variables in the model ($1 \leq k \leq L$) and p the total number of available variables, this step involves $k \times (p - k)$ model calculations.

- b) After all the substitutions are performed, if the best optimized parameter obtained during the replacement procedure is better than the previous value, the old model is substituted by the best new model. In this case, the procedure is iteratively repeated for the new model. Otherwise, if the best optimized parameter is worse than the previous value, the procedure ends and a new model among the initially selected models is taken into consideration.

• Cluster Significance Analysis (CSA)

This is a method proposed for determining which molecular descriptors of a set of compounds are associated with a biological response. The active compounds are expected to be similar to each other with respect to these relevant descriptors and so will cluster together in the space defined by the corresponding descriptors.

This approach, originally proposed for binary response variables [McFarland and Gans, 1986], was extended to the quantitative biological responses y , scaled between zero and one [Rose and Wood, 1998] and then called **Generalized Cluster Significance Analysis (GCSA)**.

Let X be a data matrix of n rows (i.e., the compounds) and p columns (i.e., the descriptors) and y the vector of the n biological responses.

The mean squared distance MSD_j was proposed to measure the tightness of the cluster of active compounds with respect to each j th molecular descriptor:

$$MSD_j = \frac{\sum_{s=1}^{n-1} \sum_{t=s+1}^n y_s \cdot y_t \cdot (x_{sj} - x_{tj})^2}{n \cdot (n-1)}$$

where n is the number of compounds, y_s and y_t the scaled biological responses of compounds s and t , x_{sj} and x_{tj} the j th descriptor values of the two compounds. A small MSD value indicates that the considered descriptor has a good capability to characterize compounds with the same biological activity.

The MSD calculated as above is proportional to that calculated as

$$MSD_j = \sum_{i=1}^n y_i \cdot (x_{ij} - \bar{x}_j^W)^2$$

where the weighted mean is calculated as

$$\bar{x}_j^W = \frac{\sum_{i=1}^n y_i \cdot x_{ij}}{\sum_{i=1}^n y_i}$$

To reach a statistical evaluation of the clustering capability of each descriptor, a test for significance is performed using a random permutation of the responses and using the permuted values to recalculate MSD values; this calculation is repeated N times (e.g., $N = 100\,000$). Then, for any given descriptor, the number c_j of times giving a value less than or equal to MSD_j is used to obtain the significance level ("p-value") and the standard error s of this estimate:

$$p_j = \frac{c_j}{N} \quad s_j = \sqrt{\frac{p_j(1-p_j)}{N}}$$

The best descriptor is chosen based on the minimum p-value.

If some descriptors are being considered together, the corresponding *MSD* random values are added together, as are the corresponding actual *MSD* values, before the count is taken.

Therefore, the selection of the best subset model can be performed by forward stepwise selection starting from the variable with the lowest *p*-value (the current model); next each of the variables not yet included in the current model is added to it in turn, producing a set of candidates with corresponding *p*-values. The candidate model with the lowest *p*-value is selected and the process is repeated on the new current model.

Moreover, a further implementation was proposed by calculating the conditional probability of candidate models.

📖 [McFarland and Gans, 1990a, 1990b, 1994, 1995; Ordorica, Velazquez *et al.*, 1993; Bayada, Hemersma *et al.*, 1999]

A number of variable selection techniques were also suggested for the Partial Least Squares (PLS) regression method [Lindgren, Geladi *et al.*, 1994; Höskuldsson, 2001]. The different strategies for **PLS-based variable selection** are usually based on a rotation of the standard solution by a manipulation of the PLS weight vector **w** or the regression coefficient vector **b** of the PLS closed form.

Generally, in PLS approach, variable selection is accomplished by calculating the **VIP score** (or **variable influence on projection**) and excluding all variables with a *VIP* score below 1 and, subsequently, keeping only the variable providing an increase in the predictive ability of the model [Wold, 1995; Lindgren, Hansen *et al.*, 1996; Chong and Jun, 2005]. The *VIP* value is a weighted sum of squares of the PLS weights *w*, taking into account the amount of explained *y* variance of each PLS dimension according to

$$VIP_j = \sqrt{\frac{p}{\sum_{m=1}^M SS(b_m \cdot t_m)} \cdot \sum_{m=1}^M w_{mj}^2 \cdot SS(b_m \cdot t_m)}$$

where *p* is the number of variables, *w_{mj}* the PLS weight of the *j*th variable for the *m*th latent variable, *SS(b_mt_m)* is the percentage of *y* explained by the *m*th latent variable.

The square sum of all *VIP* scores is equal to the number of model variables.

Other specific PLS-based variable selection methods were proposed and presented below.

Intermediate Least Squares regression (ILS) is an extension of the Partial Least Squares (PLS) algorithm that calculates the optimal variable subset model as the intermediate to PLS and stepwise regression, by two parameters whose values are estimated by cross-validation [Frank, 1987]. The first parameter is the number of optimal latent variables and the second is the number of elements in the weight vector **w** set to zero. This last parameter (ALIM) controls the number of selected variables by acting on the weight vector of each *m*th latent variable as the following:

$$\max_j(w_{jm}) \rightarrow w_{jm} = 1 \quad j = 1, p$$

that is, setting at one the *p*-ALIM largest weights, and setting at zero the remaining ALIM weights (thus excluding the corresponding variables). By modifying the number of elements to be set at zero in each weight vector, a whole range of models can be calculated between PLS models (ALIM = 0) and stepwise models (ALIM = *p* − 1).

Developments of PLS-based variable selection are the **Interactive Variable Selection for PLS** (IVS-PLS) [Lindgren, Geladi *et al.*, 1994, 1995], where two methods for manipulation of the weight vector $\mathbf{w}$, called *inside-out* and *outside-in* thresholding, are iteratively used, and **Selective PLS** proposed as a tool for multivariate design and able to separate the variables into a small number of orthogonal groups [Kettaneh-Wold, MacGregor *et al.*, 1994]. Another method, called **Uniformative Variable Elimination by PLS** (UVE-PLS), was proposed based on an analysis of the variance of the regression coefficients obtained by PLS on autoscaled or centered data [Centner, Massart *et al.*, 1996]. The fitness to enter the j th variable in the model is evaluated by the function:

$$c_j = \frac{|b_j|}{s(b_j)}$$

where b and $s(b)$ denote the regression coefficients and their standard deviations, respectively; these are estimated by the variability of the coefficients in the leave-one-out procedure. Only the variables with a c value greater than a cutoff value are retained in the model. The cutoff value of c is estimated in such a way to exclude from the model all the random variables added to the original variables. The values of the random generated variables range between 0 and 10^{-10} , thus preserving the coefficient variability but negligibly influencing the model.

Other techniques for variable selection were proposed within the framework of the $\rightarrow$ *grid-based QSAR techniques* (often for $\rightarrow$ CoMFA approach).

Among these methods, **GOLPE (Generating Optimal Linear PLS Estimations)** is a variable selection method for selecting by $\rightarrow$ *experimental design* a limited number of $\rightarrow$ *interaction energy values*, aimed at obtaining the best predictive PLS models.

The GOLPE procedure begins with the calculation of the PLS model using all variables, followed by variable preselection according to a D-optimal design in the loading space. The D-optimal design works on the original variables described by their loadings in the principal component space, the dimensionality of which is selected according to the usual cross-validation criteria for PLS. The number of variables to be selected is user specified, but it is suggested to keep not less than a half of the variables each time, in an iterative manner [Baroni, Costantino *et al.*, 1993a, 1993b]. Several variable subsets showing D-optimality are tried and the subset with the best prediction ability is retained. Variable effect and its significance on the prediction ability of the PLS model is checked. Moreover, to avoid $\rightarrow$ *chance correlation*, dummy variables are also introduced in the design matrix. For this kind of constrained problem, D-optimal designs are more efficient than fractional factorial designs (FFD designs), as initially proposed for variable selection [Baroni, Clementi *et al.*, 1992].

Various **grid region selection methods** have been recently proposed with the aim of significantly reducing the number of field variables and their mutual correlation.

Cross-validated R^2 guided region selection (Q^2 -GRS) divides the grid box into several small boxes, and CoMFA is separately performed for each box [Cho and Tropsha, 1995, 1998]. The box associated with Q^2 values greater than a specified threshold value is selected for further analysis. The *genetic algorithm-based PLS* (GA-PLS) has been proposed as a statistical tool able to select field variable combinations by the genetic algorithm using cross-validated R^2 value of the PLS model [Leardi, Boggia *et al.*, 1992; Leardi, 1994; Hasegawa, Miyashita *et al.*, 1997; Hasegawa and Funatsu, 1998; Kimura, Hasegawa *et al.*, 1998]. In this case, only a few number of significant variables are extracted. Furthermore, the *GA-based region selection method* (GA-RGS) uses domains of variables instead of single field variables [Norinder, 1996b; Pastor, Cruciani *et al.*,

1997; Hasegawa, Kimura *et al.*, 1999]; *GOLPE-Guided region selection* uses the Voronoi polyhedra for the region representation [Cruciani, Pastor *et al.*, 1997; Cruciani, Clementi *et al.*, 1998].

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **variable subset selection** $\equiv$ *variable selection*
- **variable Wiener indices** $\equiv$ *generalized Wiener indices* $\rightarrow$ Wiener index
- **variable Zagreb indices** $\rightarrow$ Zagreb indices
- **variance** $\rightarrow$ statistical indices ($\odot$ indices of dispersion)
- **variation** $\rightarrow$ distance matrix
- **VDI** $\equiv$ *vertex degree-distance index* $\rightarrow$ Cao–Yuan indices
- **$v^m d^n$ matrices** $\rightarrow$ distance-degree matrices
- **VEA indices** $\rightarrow$ spectral indices ($\odot$ eigenvalues of the adjacency matrix)

■ vectorial descriptors

Any structural information recorded into a numerical ordered string that provides a unique pattern to describe a molecule. It consists of bins, each bin being a specific molecular feature or the spectral region where an instrumental signal is recorded or, simply, a molecular descriptor.

Each bin of the string is associated with a number that can be (a) a real number (often normalized between 0 and 1), representing the intensity of an instrumental signal (IR, mass, NMR, etc.) or value of molecular descriptors, such as $\rightarrow$ *RDF descriptors*, $\rightarrow$ *3D-MoRSE descriptors*, $\rightarrow$ *WHIM descriptors*; (b) an integer number, representing the number of occurrences of the molecular feature expressed by the bin; these vectors are sometimes referred to as **holographic vectors**; (c) a 0/1 value to characterize absence/presence of molecular features or signals. In the last case, the vectorial descriptor is usually referred to as a binary vector, bit-string or Boolean array.

A vector based on real numbers can always be transformed into a binary vector by setting 1 if the real number is greater than a threshold value and 0 otherwise. Similarly, a vector based on occurrences of structural features can always be transformed into a binary vector by unfolding the frequencies, which means assigning to each structural feature a number of bits, allowing the number of occurrences of that feature up to that number to be recorded. For example, a bin equal to 3 for nitrogen atom N can be replaced by 4 new binary bins, corresponding to 0N, 1N, 2N, and 3N, and only the last one will be set 1. Eventually, another bin can be added to store all the occurrences greater than 3. Another kind of unfolding may be using bins such as 0, *at least* 1N, *at least* 2N, *at least* 3N. In this case, all the bins apart from the first one will be set 1.

Several algorithms to obtain vectorial descriptors often produce vectors of variable length, depending on the number of atoms or bonds of the molecule or, more generally, on some features related to the algorithm. For example, the number of nonzero spectroscopic signals of a molecule closely depend on the molecule itself. In these cases, the algorithm must be implemented by a suitable transformation capable of obtaining **uniform-length descriptors**, which are vectors having the same cardinality independently of the molecule size and often comprising of homogeneous descriptors [Junghans and Pretsch, 1997; Baumann, 1999].

This implies at least one rule to define a fixed number L of elements in each vector and one rule to fill in the vector elements when the values are missing (for example, filling in with zero values).

Examples of uniform-length descriptors are $\rightarrow$ *EVA descriptors*, $\rightarrow$ *topological charge indices*, $\rightarrow$ *atomic walk count sequence*, $\rightarrow$ *SE-vectors*, $\rightarrow$ *molecular profiles*, $\rightarrow$ *spectra descriptors*, $\rightarrow$ *3D-MoRSE descriptors*, $\rightarrow$ *autocorrelation descriptors*, $\rightarrow$ *affinity fingerprints*, and most of the string representations of molecules based on $\rightarrow$ *substructure descriptors*.

Any set of molecular descriptors used to represent molecules in a fixed and ordered sequence was called **basis of descriptors** [Randić, 1992c; Randić and Trinajstić, 1993a; Baskin, Skvortsova *et al.*, 1995]. A basis of descriptors can be defined: (a) selecting a class of homogeneous naturally ordered descriptors ($\rightarrow$ *path counts*, $\rightarrow$ *connectivity indices*, $\rightarrow$ *3D-MoRSE descriptors*, $\rightarrow$ *EVA descriptors*, $\rightarrow$ *Burden eigenvalues*, etc.); (b) by using a few *ad hoc* selected descriptors such as one $\rightarrow$ *lipophilicity descriptor*, one $\rightarrow$ *steric descriptor* and one $\rightarrow$ *electronic descriptor*, as used in $\rightarrow$ *classical QSAR*.

A basis of descriptors can be viewed as a representation of the $\rightarrow$ *molecular structure* suitable for use in different applications. An interesting characteristic of a basis of descriptors is that they can be orthogonalized, thus providing a new basis of $\rightarrow$ *orthogonalized descriptors*.

- **Vector-matrix-vector multiplication** $\rightarrow$ graph invariants
- **vector semisum charge-transfer index** $\rightarrow$ topological charge indices
- **VED indices** $\rightarrow$ spectral indices ($\odot$ eigenvalues of the distance matrix)
- **VEM count** $\rightarrow$ ETA indices
- **VEM environment** $\equiv$ *Valence Electron Mobile environment* $\rightarrow$ ETA indices
- **VEM vertex count** $\rightarrow$ ETA indices
- **Vérhalom index** $\rightarrow$ path counts
- **Verloop parameters** $\equiv$ *Sterimol parameters*
- **Verloop Sterimol parameters** $\equiv$ *Sterimol parameters*
- **vertex adjacency matrix** $\equiv$ *adjacency matrix*
- **vertex centric indices** $\rightarrow$ centric indices
- **vertex chromatic decomposition** $\equiv$ *chromatic decomposition*
- **vertex chromatic information index** $\equiv$ *chromatic information index* $\rightarrow$ chromatic decomposition
- **vertex chromatic number** $\equiv$ *chromatic number* $\rightarrow$ chromatic decomposition
- **vertex complexity** $\rightarrow$ topological information indices
- **vertex-connectivity matrix** $\equiv$ χ *matrix* $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **vertex coordinate** $\rightarrow$ indices of neighborhood symmetry
- **vertex-cycle incidence matrix** $\rightarrow$ incidence matrices ($\odot$ cycle matrices)
- **vertex cyclic degree** $\rightarrow$ incidence matrices ($\odot$ cycle matrices)

■ vertex degree (δ)

The vertex degree δ of an atom is the count of its σ electrons in the $\rightarrow$ *H-depleted molecular graph*, that is, the number of adjacent nonhydrogen atoms. This quantity is a $\rightarrow$ *local vertex invariant* that is easily calculated from the $\rightarrow$ *adjacency matrix* **A** as the sum of the entries a_{ij} in the i th row or from the $\rightarrow$ *distance matrix* **D** as the sum of the entries equal to 1 in the i th row:

$$\delta_i \equiv VS_i(\mathbf{A}) = \sum_{j=1}^A a_{ij} = {}^1f_i$$

where A is the number of graph vertices and VS is the general symbol for matrix row sum ($\rightarrow$ *vertex sum operator*); ${}^1f_i$ is the $\rightarrow$ *vertex distance count* of first order (i.e., the number of distances equal to 1 from the i th vertex). It was demonstrated that the vertex degree δ_i coincides with the i th diagonal element of the second power of the adjacency matrix [Trinajstić, 1992], that is,

$$\delta_i = [\mathbf{A}^2]_{ii}$$

Two simple properties of the vertex degrees are [Favaron, Mahéo *et al.*, 2003]

$$\sum_{i=1}^A \delta_i = 2 \cdot B \quad \sum_{i=1}^A \delta_i^2 = \sum_{i=1}^A \sum_{j=1}^A a_{ij} \cdot (\delta_i + \delta_j)$$

where a_{ij} are the elements of the adjacency matrix that are equal to 1 only for adjacent vertices and zero otherwise.

The sum of the vertex degrees of the neighbors of the i th vertex is the first order $\rightarrow$ *extended connectivity*. The average value of the extended connectivity was called **dual degree** [Favaron, Mahéo *et al.*, 2003]:

$$\delta_i^* = \frac{\sum_{j=1}^A a_{ij} \cdot \delta_j}{\delta_i}$$

Moreover, the quantity

$$BI_i = \frac{\delta_i \cdot (\delta_i - 1)}{2}$$

is the number of paths of length 2 passing through the i th vertex, that is, the paths having vertex v_i as the central vertex. The sum of this local invariant over all the graph vertices is the $\rightarrow$ *Bertz branching index*.

The vertex degree describes the role of the atom in terms of its connectedness and the count of σ -electrons excluding hydrogen atoms; it is a **σ -electron descriptor**, that is,

$$\delta_i = \sigma_i - h_i$$

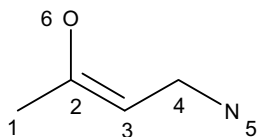
where σ_i and h_i are the number of σ electrons and hydrogens bonded to the i th atom, respectively [Kier and Hall, 1986]. The vertex degree does not account for bond multiplicity and chemical nature of atoms.

The degree of a vertex with a loop is taken to be the number of edges incident to this vertex plus two for the loop (for instance, a lone pair), because the loop contributes twice to the number of edges incident at that vertex [Lukovits, Miličević *et al.*, 2002].

Moreover, to account for atom chirality, chirality correction factors were proposed to be added to the vertex degrees [Schultz, Schultz *et al.*, 1995; Golbraikh, Bonchev *et al.*, 2001a], thus allowing calculation of $\rightarrow$ *topological chirality descriptors*.

Example V3

The vertex degrees for heavy atoms of 4-amino-2-hydroxy-but-2-ene.



$\delta_1 = \delta_5 = \delta_6 = 1$ because vertices 1, 5, and 6, being terminal atoms, have only one neighbor. $\delta_3 = \delta_4 = 2$ because vertices 3 and 4 are bonded to two vertices, respectively. $\delta_2 = 3$ because vertex 2 is bonded to vertices 1, 3, and 6. Note that δ values do not allow the distinguishing of atoms 1(C), 5(N), and 6(O) nor the distinguishing of atoms 3($=C<$) and 4($>C<$).

The **vertex degree matrix**, denoted by V , is a diagonal matrix of dimension $A \times A$ whose diagonal entries are the vertex degrees of molecule atoms:

$$[V]_{ij} = \begin{cases} \delta_i & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

δ_i being the vertex degree of the i th atom. This matrix is used in the calculation of the $\rightarrow$ *Laplacian matrix*. A **generalized vertex degree matrix** V^λ can be defined as

$$[V^\lambda]_{ij} = \begin{cases} \delta_i^\lambda & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Particular realizations of the generalized vertex degree matrix are the **vertex Zagreb matrix**, denoted by ${}^v\mathbf{ZM}$, and defined as [Janežič, Miličević *et al.*, 2007]

$$[{}^v\mathbf{ZM}]_{ij} = \begin{cases} \delta_i^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

and the **modified vertex Zagreb matrix**, denoted by ${}^{mv}\mathbf{ZM}$, and defined as [Janežič, Miličević *et al.*, 2007]

$$[{}^{mv}\mathbf{ZM}]_{ij} = \begin{cases} 1/\delta_i^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

The sum of the diagonal elements of the vertex Zagreb matrix results into the $\rightarrow$ *first Zagreb index*, while the sum of the diagonal elements of the modified vertex Zagreb matrix results into the $\rightarrow$ *modified first Zagreb index*.

For vertex- and edge-weighted graphs, the **valency of a vertex** (or **weighted vertex degree**) was defined as the sum of the weights of all the edges incident with vertex v_i [Ivanciuc and Ivanciuc, 1999; Ivanciuc, 2000i]:

$$val_i(w) = \sum_{j=1}^A a_{ij} \cdot w_{ij}$$

where w_{ij} are the edge weights for pairs of adjacent vertices, and zero otherwise; a_{ij} are the elements of the $\rightarrow$ *adjacency matrix*. For simple graphs, where all edge weights equal one, the valency coincides with the simple vertex degree. Moreover, vertex valency can be also computed as the sum of off-diagonal elements in the row of $\rightarrow$ *weighted adjacency matrices*.

The **bond vertex degree** δ_i^b is another local invariant that accounts for atom connectedness and also for $\rightarrow$ *bond multiplicity*. It is calculated from the $\rightarrow$ *atom connectivity matrix* C as the sum of row entries [Kier and Hall, 1986]:

$$\delta_i^b \equiv VS_i(C) = \sum_{j=1}^A a_{ij} \cdot \pi_{ij}^* \quad \pi_{ij}^* = 0 \text{ if } (i,j) \notin E(G)$$

where VS is the general symbol for local invariants defined as matrix row sums, π^* is the $\rightarrow$ *conventional bond order*, equal to 1 for single bonds, two for double bonds, three for triple bonds, 1.5 for conjugated bonds and zero otherwise; a_{ij} are the adjacency matrix elements equal to 1 only for vertices adjacent to the i th vertex. For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the bond vertex degree allows the distinguishing of atom 3 ($=C<$) and 4

(>C<) since δ^b of vertex 3 is equal to 3 (there are one single and one double bond), while δ^b of vertex 4 is equal to 2 (there are two incident single bonds). However, it still does not allow the distinguishing of heteroatoms: $\delta_1^b = \delta_5^b = \delta_6^b = 1$. The bond vertex degree can also be calculated using quantum-chemical derived bond orders instead of the conventional bond order [Estrada and Montero, 1993].

The bond vertex degree is closely related to the $\rightarrow$ atomic multigraph factor f_i as

$$\delta_i^b = \sum_{j=1}^A a_{ij} \cdot \pi_{ij}^* = \delta_i + f_i = \delta_i + \sum_{j=1}^A a_{ij} \cdot (\pi_{ij}^* - 1) \quad \pi_{ij}^* = 0 \text{ if } (i, j) \notin E(G)$$

To take into account all valence electrons of the i th atom, the vertex degree is replaced by the **valence vertex degree** δ_i^v (also called **vertex valence**) defined as

$$\delta_i^v = Z_i^v - h_i = \sigma_i + \pi_i + n_i - h_i$$

where Z_i^v is the number of valence electrons (σ electrons, π electrons and lone pair electrons n) of the i th atom and h_i is the number of hydrogen atoms bonded to it (Kier and Hall, 1986). This definition holds for atoms of the second principal quantum level (C, N, O, F). For atoms of higher principal quantum levels (P, S, Cl, Br, I), Kier and Hall proposed to account for both valence and nonvalence electrons, as the following:

$$\delta_i^v = \frac{(Z_i^v - h_i)}{(Z_i - Z_i^v - 1)}$$

where Z_i is the total number of electrons of the i th atom, that is, its atomic number. δ_i^v encodes the electronic identity of the atom in terms of both valence electron and core electron counts; it is a **valence electron descriptor**. It is useful to characterize heteroatoms and carbon atoms involved in multiple bonds. For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the valence vertex degree of vertex 2 is equal to 4, because the number of valence electrons Z^v is equal to 4 and there are no bonded hydrogens; all the terminal atoms are distinguished: δ^v of vertex 1 (C) is equal to 1 because Z^v is equal to 4 and there are three bonded hydrogens, δ^v of vertex 5 (N) is equal to 3 because the number of valence electrons Z^v is equal to 5 and there are two bonded hydrogens and δ^v of vertex 6 (O) is equal to 5 because the number of valence electrons Z^v is equal to 6 and there is one bonded hydrogen.

Vertex degrees, bond vertex degrees, and valence vertex degrees are used in the calculation of several molecular descriptors. The most popular ones are $\rightarrow$ Zagreb indices and $\rightarrow$ connectivity indices.

It was observed that the difference [Kier and Hall, 1981]

$$\delta_i^v - \delta_i = \pi_i + n_i$$

provides a quantitative measure of the potential of the atom for intermolecular interaction and reaction, the count being of π and n lone pair electrons of the atoms. A significant correlation with the $\rightarrow$ Mulliken electronegativity χ^{MU} was found as

$$\chi_i^{\text{MU}} = 2.05 \times (\delta_i^v - \delta_i) + 6.99$$

Therefore, the **Kier–Hall electronegativity** of the i th atom was proposed based on the difference between valence and simple vertex degree, as [Kier and Hall, 1981]

$$\chi_i^{KH} \equiv KHE_i = \frac{\delta_i^v - \delta_i}{L_i^2}$$

where the square of the principal quantum number L_i of the i th atom is used to account for the increase in the screening effect due to inner electrons of higher row elements. Since electronegativity for C_{sp^3} is equal to zero, this scale can also be thought of as relative electronegativity with respect to C_{sp^3} . Electronegativity for the hydrogen atom is assumed equal to -0.2 ; KHE is related to Mulliken electronegativity by the relationship:

$$\chi_i^{MU} = 7.99 \times \frac{(\delta_i^v - \delta_i)}{L_i^2} + 7.07$$

where δ^v in this equation is defined as

$$\delta_i^v = Z_i^v - h_i$$

for all the atoms, that is, it is the simple valence vertex degree defined above because the principal quantum number has already been considered. KHE values for some atom-types are collected in Table V1; an extended list of KHE values is given in Table E7 (see $\rightarrow$ *electrotopological state indices*).

Table V1 Valence vertex degrees δ^v , vertex degrees δ , their differences, and Kier–Hall electronegativity values for different atom-types.

Atom/hybrid	δ^v	δ	$\delta^v - \delta$	KHE
C_{sp^3}	4	4	0	0.00
C_{sp^2}	4	3	1	0.25
C_{sp}	4	2	2	0.50
N_{sp^3}	5	3	2	0.50
N_{sp^2}	5	2	3	0.75
N_{sp}	5	1	4	1.00
O_{sp^3}	6	2	4	1.00
O_{sp^2}	6	1	5	1.25
S_{sp^3}	6	2	4	0.44
S_{sp^2}	6	1	5	0.55
F	7	1	6	1.50
Cl	7	1	6	0.67
Br	7	1	6	0.38
I	7	1	6	0.24

Data from [Kier and Hall, 1999d].

Based on an analogous modification of the vertex degree, another electronegativity measure was defined in the framework of the $\rightarrow$ *ETA indices*.

Kier and Hall also proposed the **valence state indicator** [Hall and Kier, 1995] that is a combination of valence and simple vertex degree as

$$VSI_i = \delta_i^v + \delta_i$$

This local invariant was defined to calculate $\rightarrow$ *atom-type E-state indices*.

The **Kupchik vertex degree**, denoted as δ^{het} , was defined as [Kupchik, 1986, 1988]

$$\delta_i^{\text{het}} = \frac{R_C}{R_i} \cdot (Z_i^v - h_i)$$

where R_i and R_C are the covalent radius of the i th atom and the carbon atom, respectively; Z_i^v is the number of electrons, and h_i the number of hydrogen atoms bonded to the i th atom. As the valence vertex degree, the Kupchik vertex degree accounts both for heteroatoms and bond multiplicity.

For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the Kupchik vertex degree of vertex 5 (N) is equal to 3.080 because the atomic number Z^v is 5, the bonded hydrogens are 2 and the covalent radius is equal to 0.75 (the covalent radius of carbon atom is 0.77); the Kupchik vertex degree of vertex 6 (O) is equal to 5.274 because the atomic number Z^v is 6, there is only one bonded hydrogen and the covalent radius is equal to 0.73.

→ *Kupchik modified connectivity indices* are molecular descriptors derived from this vertex degree.

The **perturbation delta value** was defined in terms of the valence vertex degree δ^v modified by the atomic environment as [Gombar, Kumar *et al.*, 1987]

$$\delta_i^p = \delta_i^v + \sum_{j=1}^A a_{ij} \cdot \gamma_{ij} \cdot \delta_j^v$$

where the perturbation term of the i th atom is the sum of the valence vertex degrees of its first neighbors (a_{ij} being the elements of the adjacency matrix equal to 1 for vertices adjacent to the i th vertex), each weighted by parameter γ_{ij} accounting for the type of the bond i - j . γ values should be functions of the properties of the connected atoms i and j (e.g., between -0.30 and $+0.30$). For $\gamma = 0$, perturbation delta values coincide with the corresponding valence vertex degrees.

For example, assuming a γ value equal to 0.1 for all the bonds, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the vertex 2 has a perturbation delta value equal to $\delta_2^p = 4 + 0.1 \times (1 + 3 + 5) = 4.9$, which accounts for the presence of a heteroatom and a C_{sp2} in its environment.

→ *Perturbation connectivity indices* are molecular descriptors derived from this modified valence vertex degree.

Another modification of the valence vertex degree, still proposed by Kier and Hall [Kier and Hall, 1990a, 1999d] and used in → *electrotopological state indices*, is the **intrinsic state** I_i defined as

$$I_i = \frac{(2/L_i)^2 \cdot \delta_i^v + 1}{\delta_i}$$

where L_i is the principal quantum number, δ_i^v is the valence vertex degree and δ_i is the simple vertex degree of the i th atom in the H-depleted molecular graph; the term $(2/L_i)^2$ is equal to 1 for the elements of the second principal quantum level. For elements of higher levels, L is used to account for the increase in the screening effect of the inner electrons.

For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the intrinsic state of vertex 2 is equal to 1.667, that is, $I_2 = [(2/2)^2 \times 4 + 1]/3 = 1.667$.

The **Madan chemical degree**, denoted as δ^c , was defined as the row sum of the → *atomic weight-weighted adjacency matrix* $A(m)$ and calculated by summing up relative atomic weights of

all the vertices j adjacent to i [Goel and Madan, 1995; Gupta, Singh *et al.*, 2001a; Bajaj, Sambi *et al.*, 2005]:

$$\delta_i^c \equiv VS_i(\mathbf{A}, \mathbf{m}) = \sum_{j=1}^A a_{ij} \cdot \frac{m_j}{m_C}$$

where m_j and m_C are the atomic weights of the j th atom and the carbon atom, respectively; a_{ij} are the elements of the $\rightarrow$ *adjacency matrix*, which are equal to 1 only for vertices adjacent to vertex i . For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the Madan chemical degree for vertex 2 is equal to 3.332 because vertex 2 is bonded to vertex 1 (C) which gives a contribution of 1, to vertex 3 (C) that gives a contribution of 1, and to vertex 6 (O) that gives a contribution of $16/12.01 = 1.332$. It is noteworthy that this local invariant does not account for bond multiplicity and, unlike the other modified vertex degrees, information on the presence of heteroatoms is not coded in the heteroatom itself, but in the adjacent atoms.

Also, to account for bond multiplicity, the following modification of the Madan vertex degree, called **extended Madan degree**, is proposed as [Authors, This Book]

$$\delta_i^\pi = \sum_{j=1}^A a_{ij} \cdot \pi_{ij}^* \frac{m_j}{m_C} \quad \pi_{ij}^* = 0 \text{ if } (i,j) \notin E(G)$$

where π^* is the $\rightarrow$ *conventional bond order*, equal to 1 for single bonds, 2 for double bonds, 3 for triple bonds, 1.5 for aromatic bonds, and zero otherwise. For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the vertex degrees of vertices 2 and 3 become 4.332 and 3, respectively.

Summing up the Madan chemical degrees of all the vertices bonded to the i th vertex, the $\rightarrow$ *chemical extended connectivity* is derived. The Madan chemical degree has been used to calculate some of the $\rightarrow$ *eccentricity-based Madan indices*.

The **Hu–Xu vertex degree** was defined as [Hu and Xu, 1997]

$$\delta'_i = \delta_i \cdot \sqrt{Z_i}$$

where Z_i is the atomic number of the considered atom and δ_i the simple vertex degree. For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), δ' of vertex 1 (C) is $1 \times \sqrt{6} = 2.449$, δ' of vertex 5 (N) is $1 \times \sqrt{7} = 2.646$, and δ' of vertex 6 (O) is $1 \times \sqrt{8} = 2.828$. Note that δ' values of vertices 3 ($=C<$) and 4 ($>C<$) are both equal to $2 \times \sqrt{6} = 4.899$. This modified vertex degree was used to derive the $\rightarrow$ *Hu–Xu ID number*.

To prevent possible degenerative cases where vertex degrees are summed (e.g., $\sqrt{1} + \sqrt{4} = \sqrt{9}$), a modification of the Hu–Xu vertex degree, here called **Alikhanidi vertex degree** and denoted as δ_i^A , was proposed [Alikhanidi and Takahashi, 2006]:

$$\delta_i^A = \delta_i \cdot \sqrt{Z'_i}$$

where Z' is a function of the atomic numbers Z_j of the atoms adjacent to the i th atom; this function was called **consecutive AT number** and defined as

$$Z'_i = \left[\sum_{j=1}^A a_{ij} \cdot \sqrt{(\sqrt{2} + Z_j)} \right]^2$$

where a_{ij} denotes the elements of the adjacency matrix, which take value of one only for pairs of bonded atoms. Unlike the Hu–Xu vertex degree, in molecule 4-amino-2-hydroxy-but-2-ene

(Example V3), the Alikhanidi vertex degree allows the distinguishing of atoms 3 and 4:

$$\delta_3^A = 2 \times \left(\sqrt{\sqrt{2} + 6} + \sqrt{\sqrt{2} + 6} \right) = 10.892$$

$$\delta_4^A = 2 \times \left(\sqrt{\sqrt{2} + 6} + \sqrt{\sqrt{2} + 7} \right) = 11.247$$

since they are characterized by different environments.

The **augmented valence** of the i th vertex is defined as the sum of its vertex degree and vertex degrees of all the other vertices in the graph, each weighted by a quantity that decreases as the topological distance from the vertex v_i increases [Randić, 2001b; Randić and Plavšić, 2002, 2003]:

$$AV_i = \sum_{j=1}^A \frac{\delta_j}{2^{d_{ij}}} = \delta_i + \sum_{j \neq i}^A \frac{\delta_j}{2^{d_{ij}}}$$

where δ_j is the vertex degree of the j th atom and d_{ij} the $\rightarrow$ topological distance between the i th and the j th vertices [Randić and Plavšić, 2002, 2003]. Based on this local vertex invariant, the $\rightarrow$ Randić–Plavšić complexity index ξ was defined as the sum of augmented valences of all mutually nonequivalent vertices in the graph.

To account for heteroatoms and/or multiple bonds, the **Ren vertex degree**, denoted as δ^m , was defined as a modification of Kier–Hall intrinsic state as [Ren, 2002b]

$$\delta_i^m = \delta_i + \left[\left(\frac{2}{L_i} \right)^2 \cdot \delta_i^v + 1 \right]^{-1} = \delta_i + (I_i \cdot \delta_i)^{-1}$$

where δ_i is the simple vertex degree of the i th atom, L_i the principal quantum number, δ_i^v the valence vertex degree, and I_i the atom $\rightarrow$ intrinsic state. This formula is applied only to heteroatoms or carbon atoms with multiple bonds and/or bonded to heteroatoms; otherwise, the Ren vertex degree coincides with the simple vertex degree δ_i .

For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the intrinsic state of vertex 2 is equal to 1.667, and then the Ren vertex degree is $3 + (1.667 \times 3)^{-1} = 3.2$.

The Ren vertex degree was used to define $\rightarrow$ AI indices and $\rightarrow$ Xu index.

The **Li valence vertex degree**, denoted as δ^{Li} , was defined as [Li, Jalbout *et al.*, 2003]

$$\delta_i^{Li} = \frac{Z_i^v \cdot (Z_i^v - h_i)}{L_i^2}$$

where Z_i^v is the number of valence electrons and L_i the principal quantum number. For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the Li valence vertex degree of oxygen (6) is $(6 \times (6 - 1))/4 = 7.5$. The proposed Li vertex degree coincides with the valence vertex degree for all the carbon atoms, 4 being the number of valence electrons and 2 the principal quantum number (i.e., $Z_i^v/L_i^2 = 1$ and then $\delta_i^{Li} = Z_i^v - h_i = \delta_i^v$). Based on this vertex degree, a set of modified $\rightarrow$ valence connectivity indices was defined.

Moreover, the following relationship between Li valence vertex degree and Kier–Hall electronegativity *KHE* holds:

$$\delta_i^{Li} = Z_i^v \cdot \left(KHE_i + \frac{\delta_i}{L_i^2} \right)$$

A further modification of Kier–Hall vertex degree definition is the **Yang vertex degree**, denoted as δ^Y , conceived to better describe atoms also in complex organic compounds by using the **Yang's Electronegative Force Gauge** [Yang, 1992]. It is defined as [Jiang, Liu *et al.*, 2003]

$$\delta_i^Y = \frac{(Z_i^v - h_i) \cdot b_i}{L_i^2 \cdot Y_i}$$

where Z_i^v is the total number of electrons of the i th atom, b_i is the bonding electron number (i.e., the vertex degree in the H-filled molecular graph), L_i is the principal quantum number of the i th atom. Y_i is the Yang's Electronegative Force Gauge (Table V2), which reflects the atomic ability of attracting charge in the formed bond. For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the vertex degrees of vertices 2 and 6 become

$$\delta_2^Y = \frac{(4-0) \times 3}{2^2 \times 1.70} = 1.765 \quad \text{and} \quad \delta_6^Y = \frac{(6-1) \times 2}{2^2 \times 1.92} = 1.302, \quad \text{respectively.}$$

Table V2 Yang's Electronegative Force Gauge Y values of some atoms.

Atom	Y	Atom	Y	Atom	Y	Atom	Y	Atom	Y	Atom	Y
H	1.40	Al	1.26	Cr	1.56	Se	1.54	Pd	1.58	W	1.56
Li	0.97	Si	1.40	Mn	1.60	Br	1.62	Ag	1.46	Re	1.60
Be	1.42	P	1.54	Fe	1.44	Rb	0.63	Cd	1.40	Os	1.56
B	1.50	S	1.62	Co	1.48	Sr	0.85	In	1.12	Ir	1.57
C	1.70	Cl	1.72	Ni	1.52	Y	1.10	Sn	1.28	Pt	1.58
N	1.85	K	0.69	Cu	1.45	Zr	1.28	Sb	1.35	Au	1.57
O	1.92	Ca	0.92	Zn	1.35	Nb	1.54	Te	1.41	Hg	1.48
F	2.00	Sc	1.15	Ga	1.26	Mo	1.56	I	1.48	Tl	1.10
Na	0.85	Ti	1.34	Ge	1.38	Ru	1.53	Cs	0.58	Pb	1.21
Mg	1.12	V	1.47	As	1.48	Rh	1.57	Ba	0.82	Bi	1.29

Data from [Jiang, Liu *et al.*, 2003].

Replacing the simple vertex degree by the Yang vertex degree, the $\rightarrow$ *Yang connectivity index* was proposed [Jiang, Liu *et al.*, 2003].

Another modification of the vertex degree, denoted as δ^* , was proposed in terms of Mulliken $\rightarrow$ *population analysis*, as the sum of Mulliken overlap population of each atom [Li, Yu *et al.*, 2000]. A good relationship was found with the vertex degree.

A vertex degree, **CTvertex degree**, accounting for the number of adjacent atoms, heteroatoms, multiple bonds and conjugation is here defined [Authors, This Book] as

$$\delta_i^{CT} = \sqrt{\delta_i \cdot \left(\prod_{j=1}^A (\pi_{ij}^*)^{a_{ij}} \right)^{1/\delta_i} \cdot \sum_{j=1}^A a_{ij} \cdot \left(\frac{Z_i + Z_j}{2 \cdot Z_C} \right)} = \sqrt{\delta_i \cdot \Pi \cdot \mathbb{Z}}$$

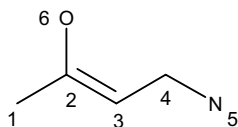
where a_{ij} are the elements of the adjacency matrix taking value equal to 1 only for pairs of bonded atoms, π^* the $\rightarrow$ *conventional bond order*, Z the atomic number. The term Π accounts for multiple bonds and conjugation, while the term $\mathbb{Z}$ for heteroatoms. For saturated hydrocarbons,

$\delta_i^{CT} = \delta_i$, Π being equal to 1 and $\mathbb{Z}$ equal to δ_i . Unlike the other vertex degrees, this vertex degree distinguishes conjugated atoms from nonconjugated atoms. For example, for carbon atom in benzene, carbon $C_2(sp^2)$ in butadiene, carbon $C_2(sp^2)$ in 2-butene, carbon $C_1(sp^2)$ in butadiene, carbon $C_1(sp^3)$ in 2-butene the CT vertex degree takes the following values: {2.449, 2.449, 2.378, 1.225, 1.000}. For the same atoms, valence vertex degrees δ^v , Ren vertex degrees δ^m , bond vertex degrees δ^b , and Yang vertex degrees δ^Y have the following values:

$$\begin{aligned}\delta^v &= \{3, 3, 3, 2, 1\} & \delta^m &= \{2.25, 2.25, 2.25, 1.33, 1.33\} \\ \delta^b &= \{3, 3, 3, 1.5, 1\} & \delta^Y &= \{1.765, 1.765, 1.765, 1.176, 0.588\}\end{aligned}$$

Example V4

Vertex degrees calculated according to different approaches for the atoms of 4-amino-2-hydroxy-but-2-ene.



		CH ₃	>C=	=CH–	>CH ₂	–NH ₂	–OH
Symbol	Name	1	2	3	4	5	6
δ	Vertex degree	1	3	2	2	1	1
δ^v	Valence vertex degree	1	4	3	2	3	5
δ^b	Bond vertex degree	1	4	3	2	1	1
VSI	Valence state indicator	2	7	5	4	4	6
δ^p	Perturbation delta value	1.4	4.9	3.6	2.6	3.2	5.4
δ^{het}	Kupchik vertex degree	1	3	2	2	3.080	5.274
δ'	Hu–Xu vertex degree	2.449	7.348	4.898	4.898	2.646	2.828
δ^A	Alikhanidi vertex degree	2.723	25.542	10.892	11.247	2.723	2.723
I	Intrinsic state	2	1.667	2	1.5	4	6
δ^m	Ren vertex degree	1	3.2	2.25	2	1.25	1.167
δ^{Li}	Li vertex degree	1	4	3	2	3.75	7.5
δ^Y	Yang vertex degree	0.588	1.765	1.324	1.176	1.216	1.302
δ^c	Madan chemical degree	1	3.332	2	2.167	1	1
δ^π	Extended Madan degree	1	4.332	3	2.167	1	1
δ^{CT}	CT vertex degree	1	3.460	2.378	2.041	1.041	1.080
δ^Z	Z-delta number	2	2	2	2	2.5	3

$\gamma = 0.1$ for all the bonds.

For hydrocarbons not containing multiple bonds, the values of the valence vertex degree, bond vertex degree, Kupchik vertex degree, Li vertex degree, Ren vertex degree, Madan vertex degree, and CT vertex degree coincide with the simple vertex degree, that is,

$$\delta_i = \delta_i^v = \delta_i^b = \delta_i^{Li} = \delta_i^m = \delta_i^c = \delta_i^{CT}$$

The **Z-delta number** is a local vertex invariant defined to distinguish heteroatoms in the molecular graph [Pogliani, 1996c]:

$$\delta_i^Z = \frac{Z_i^v}{L_i}$$

where Z_i^v is the number of valence electrons and L_i the principal quantum number. For example, in molecule 4-amino-2-hydroxy-but-2-ene (Example V3), the Z-delta number of all the carbon atoms (1, 2, 3, 4) is equal to $4/2 = 2$, that of nitrogen (5) is equal to $5/2 = 2.5$, and that of oxygen (6) is $6/2 = 3$.

Unlike the other vertex degrees, Z-delta-number does not account for the number of adjacent atoms; it was used to derive the $\rightarrow$ *Pogliani index*.

Besides the most popular $\rightarrow$ *connectivity indices*, $\rightarrow$ *Zagreb indices* and $\rightarrow$ *Schultz molecular topological index*, a lot of other topological indices were proposed as a function of vertex degrees of molecule atoms, such as, for example, $\rightarrow$ *Sh indices*, $\rightarrow$ *Xu index*, and $\rightarrow$ *eccentricity-based Madan indices*.

Some others are reported below.

The number of vertices with vertex degree equal to g is a $\rightarrow$ *vectorial descriptor* called **vertex degree count** ${}^gF$; therefore to each graph G , the vector

$$\{^1F, {}^2F, {}^3F, {}^4F\}$$

can be associated, provided the maximum vertex degree equals four.

Two molecular descriptors PR1 and PR2, called **ramification pair indices**, were proposed based on the vertex degree count of third order and defined as the number of pairs of vertices with a vertex degree equal to three at a topological distance equal to 1 and two, respectively [Rios-Santamarina, García-Domenech *et al.*, 1998].

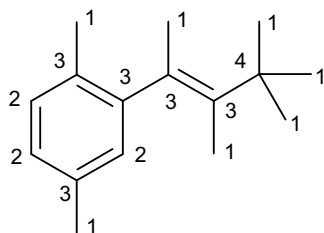
A simple **ramification index** was also proposed for acyclic graphs:

$$r = \sum_{\delta_i > 2} (\delta_i - 2) = {}^3F + 2 \cdot {}^4F$$

where the sum runs over all the vertex degrees greater than two [Araujo and De La Peña, 1998]. This index is quite similar to the $\rightarrow$ *quadratic index* and was previously used by Pitzer [Pitzer and Scott, 1941].

Example V5

Vertex degree count, ramification pair indices and ramification index for the molecule shown below. The numbers on the graph vertices represent the vertex degrees.



$$\begin{aligned} {}^1F &= 7, {}^2F = 3, {}^3F = 5, {}^4F = 1 \\ \text{PR1} &= 3, \text{PR2} = 3 \\ r &= 7 \end{aligned}$$

The **simple topological index S** (or better called **Narumi–Katayama index**) [Narumi and Katayama, 1984; Narumi, 1987; Gutman, 1990; Tomović and Gutman, 2001b] is a molecular descriptor related to $\rightarrow$ *molecular branching* proposed as the product of the vertex degrees δ

$$S = \prod_{i=1}^A \delta_i$$

where A is the number of atoms. The simple topological index S is quite similar to the $\rightarrow$ *total connectivity index* χ_T . For computational problems arising from molecules with a large number of atoms, the logarithm of this index should be adopted.

Other related molecular descriptors were proposed:

$$A = \frac{\sum_{i=1}^A \delta_i}{A} = \frac{2 \cdot B}{A} \quad H = \frac{A}{\sum_{i=1}^A 1/\delta_i} \quad G = \left(\prod_{i=1}^A \delta_i \right)^{1/A} = S^{1/A}$$

where A is the **arithmetic topological index**, H the **harmonic topological index**, and G the **geometric topological index** [Narumi, 1987]. B is the number of molecule bonds; obviously, the arithmetic topological index is the same for all isomers.

Among these molecular descriptors the following simple relationship holds:

$$A \geq G \geq H$$

- **vertex degree count** $\rightarrow$ vertex degree
- **vertex degree distance** $\rightarrow$ Schultz molecular topological index
- **vertex degree-distance index** $\rightarrow$ Cao–Yuan indices
- **vertex degree matrix** $\rightarrow$ vertex degree
- **vertex distance-delta matrix** $\equiv$ *delta matrix* $\rightarrow$ distance-path matrix
- **vertex distance code** $\rightarrow$ distance matrix
- **vertex distance complement matrix** $\equiv$ *distance complement matrix* $\rightarrow$ distance matrix
- **vertex distance complexity** $\rightarrow$ topological information indices
- **vertex distance counts** $\rightarrow$ distance matrix
- **vertex distance degree** $\equiv$ *distance degree* $\rightarrow$ distance matrix
- **vertex distance-delta matrix** $\equiv$ *delta matrix* $\rightarrow$ distance-path matrix
- **vertex distance matrix** $\equiv$ *distance matrix*
- **vertex distance-path matrix** $\equiv$ *distance-path matrix*
- **vertex distance sum** $\equiv$ *distance degree* $\rightarrow$ distance matrix
- **vertex–distance–vertex–degree matrix** $\rightarrow$ distance–degree matrices
- **vertex double sum** $\rightarrow$ local invariants ($\odot$ Local Vertex Invariants)
- **vertex eccentricity** $\equiv$ *atom eccentricity* $\rightarrow$ distance matrix
- **vertex-edge incidence matrix** $\rightarrow$ incidence matrices
- **vertex Harary matrix** $\equiv$ *Harary matrix* $\rightarrow$ distance matrix
- **vertex information distance index** $\equiv$ *vertex distance complexity* $\rightarrow$ topological information indices
- **vertex information layer index** $\rightarrow$ topological information indices ($\odot$ vertex distance complexity)
- **vertex matrices** $\rightarrow$ matrices of molecules

- **vertex orbital information content** → orbital information indices
- **vertex path code** → path counts
- **vertex path eccentricity** $\equiv$ *atom detour eccentricity* → detour matrix
- **vertex-path incidence matrix** → incidence matrices
- **vertex path sum** $\equiv$ *path degree* → path counts
- **vertex structural code** → self-returning walk counts
- **vertex sum** → local invariants ($\odot$ Local Vertex Invariants)
- **vertex sum operator** $\equiv$ *row sum operator* → algebraic operators
- **vertex topological state** $\equiv$ *topological state* → weighted matrices ($\odot$ weighted distance matrices)
- **vertex valence** $\equiv$ *valence vertex degree* → vertex degree
- **vertex-valence-weighted Schultz distance matrix** → weighted matrices ($\odot$ weighted distance matrices)
- **vertex-weighted Schultz distance matrix** → weighted matrices ($\odot$ weighted distance matrices)
- **vertex Zagreb matrix** → vertex degree
- **vertices** → graph
- **VHSE descriptor** → biodescriptors ($\odot$ amino acid descriptors)
- **V index** → topological information indices ($\odot$ relative vertex distance complexity)
- **V-like indices** → topological information indices ($\odot$ Balaban-like information indices)
- **VIN index** → environmental indices ($\odot$ leaching indices)
- **VIP score** → variable selection
- **VLOGP** → lipophilicity descriptors
- **VMV approach** $\equiv$ *vector-matrix-vector multiplication* → graph invariants
- **VolSurf descriptors** → grid-based QSAR techniques

■ volume descriptors (V)

These are → *steric descriptors* and/or → *size descriptors* representing the volume of a molecule. The volume of a molecule can be derived from experimental observation such as the volume of the unit cell in crystals or the molar volume of a solution or from theoretical calculations. Analytical and numerical approaches have been proposed for the calculation of molecular volume where the measure depends directly on the definition of → *molecular surface*, such as the → *solvent-excluded volume* that is a volume descriptor based on solvent-accessible surface.

A list of other common volume descriptors is given below.

• van der Waals volume (V^{vdw})

The van der Waals volume, also called **intrinsic molecular volume** V_I , is the volume of the space within the → *van der Waals molecular surface*. The **van der Waals radius** R^{vdw} is the distance at which the attractive and repulsive forces between two nonbonded atoms are balanced, thus the van der Waals volume may be regarded as an impenetrable volume for other molecules.

The accurate calculation of van der Waals volume, as well as of surface area, is quite complex and different approaches have been proposed, most being based on a numerical integration method [Leo, Hansch *et al.*, 1976; Testa and Purcell, 1978; Pearlman, 1980; Pavani and Ranghino, 1982; Gavezzotti, 1983; Meyer, 1985a, 1985b, 1986a, 1988a, 1988b, 1989; Motoc

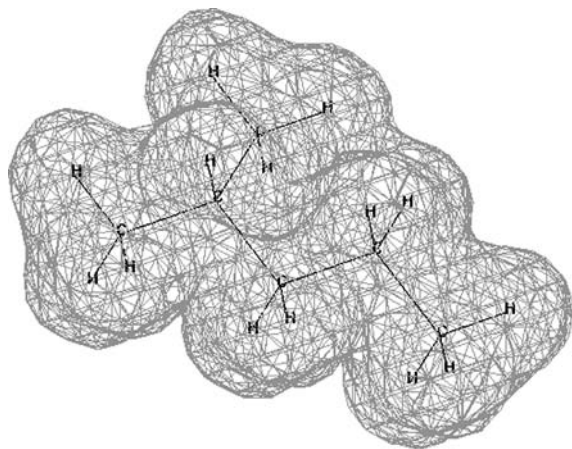


Figure V2 Van der Waals molecular volume of 2-methylpentane.

and Marshall, 1985; Stouch and Jurs, 1986; Bodor, Gabanyi *et al.*, 1989]. For example, a molecule or fragment can be described by a set of vectors $\mathbf{S}_i$, connecting the atomic nuclei to a chosen origin, and a set of radii, R_i (one for each atomic species, Table V3), each of which defines a sphere around each nucleus i [Gavezzotti, 1983].

In the hard-sphere model [Gavezzotti, 1983; Ciubotariu, Medeleanu *et al.*, 2004], the van der Waals volume of a molecule is defined as the total volume embedded by an envelope of hard-spheres, each representing an atomic van der Waals volume, and can be formally described as

$$V^{vdw} = \iiint dx \, dy \, dz$$

A simple estimate of this integral is performed by including the molecule into a bounding parallelepiped with volume V_p . Then, if n_t random points are generated within this volume and n_s are the random points that belong to the molecule, the van der Waals volume V^{vdw} is estimated as

$$V^{vdw} = \frac{n_s}{n_t} \cdot V_p$$

The points belonging to the molecule are the points that satisfy at least one of the following inequality:

$$(X_i - x)^2 + (Y_i - y)^2 + (Z_i - z)^2 \leq (R_i^{vdw})^2, \quad i = 1, \dots, A$$

where A is the number of atoms and R^{vdw} the van der Waals radius; X_i , Y_i , Z_i are the coordinates of the i th atom and x , y , z the coordinates of the generated points.

Bondi developed a method based on covalent bond distances and van der Waals radii (Table V3) to calculate van der Waals volume [Bondi, 1964]. The volume calculated in this way is sometimes called **Bondi volume**. It is obtained easily by summing up appropriate volume contributions of atoms and functional groups, as proposed by Bondi; note that the Bondi volume does not account for the overlaps which are possible whenever three or more atomic spheres intersect, it is roughly 60–70% of molecular volume.

Table V3 Van der Waals radii for some atoms.

Atom	Bondi ^a	Rohrbaugh–Jurs ^b	Gavezzotti ^c
H	1.20	1.20	1.17
C	1.70	1.70	1.75
N	1.55	1.55	1.55
O	1.52	1.52	1.40
S	1.80	1.80	—
F	1.47	1.50	1.30
Cl	1.75	1.75	1.77
Br	1.85	1.85	1.95
I	1.98	1.97	2.10

^a[Bondi, 1964];^b[Rohrbaugh and Jurs, 1987b];^c[Gavezzotti, 1983].

• molar volume ($\bar{V}$)

The molar volume of a substance is an important $\rightarrow$ *physico-chemical property* and is defined as the ratio of the volume of a sample of that substance (expressed in liters, for example) to the amount of substance (usually expressed in moles) in the sample.

It is experimentally measured as

$$\bar{V} = \frac{\text{MW}}{\rho}$$

where MW is the $\rightarrow$ *molecular weight* and ρ the density of the liquid. The SI unit of molar volume is cubic meters per mole (m^3/mol). It is the reciprocal of amount of substance concentration and it is related to $\rightarrow$ *molar refractivity* via $\rightarrow$ *refractive index* and to $\rightarrow$ *parachor* via Sugen equation.

For an ideal gas, the standard molar volume is the volume that is occupied by one mole of substance (in gaseous form) at standard temperature and pressure (STP) of 273.15 K (H_2O freezing temperature) and 101 325 Pa (1 atm). It is $0.022\,414\,\text{m}^3/\text{mol}$ or $22.414\,\text{l/mol}$ and is directly related to the universal gas constant R in the ideal gas law.

The molar volume is usually given for a solid substance at 298.15 K (temperature of standard state). Apart from temperature and density, it depends on phase and allotrope of the substance.

The **molecular volume** V is defined as the volume of the region within a molecule is constrained by its neighbors. It is calculated from the experimental density ρ of the liquid as

$$V = \frac{\text{MW}}{\rho \cdot N_A} = \frac{\bar{V}}{N_A}$$

where MW is the molecular weight and N_A is the Avogadro number [Meyer, 1985a]. The molecular volume as well as the molar volume characterize the bulk compound, comprising both the intrinsic molecular volume V^{vdw} and the volume of the empty “packing” space between molecules, sometimes called **free molecular volume** V^f , that is, $V = V^{vdw} + V^f$.

Molar volume can also be calculated by additive fragment methods [Elbro, Fredeslund *et al.*, 1991; Schotte, 1992] such as the fragment method of LeBas [Reid, Prausnitz *et al.*, 1988].

Moreover, it is frequently used as a measure of the $\rightarrow$ cavity term in $\rightarrow$ linear solvation energy relationships.

📖 [Immirzi and Perini, 1977; Horvath, 1992; van Haelst, Paulus *et al.*, 1997]

- **McGowan's characteristic volume (V_X)**

A steric descriptor defined as the sum of atomic volume parameters for all atoms in the molecule:

$$V_X = \sum_{i=1}^A w_i - 6.56 \cdot B$$

where w_i are the McGowan volume parameters (Table V4) and B the number of bonds [Abraham and McGowan, 1987; Zhao, Abraham *et al.*, 2003a].

Table V4 McGowan atomic parameters.

Atom	w_i	Atom	w_i
H	8.71	S	22.91
C	16.35	F	10.48
N	14.39	Cl	20.95
O	12.43	Br	26.21
P	24.87	I	34.53

The McGowan's characteristic volume is frequently used as a measure of the $\rightarrow$ cavity term in $\rightarrow$ linear solvation energy relationships.

- **Corey–Pauling–Koltun volume (V^{CPK}) ($\equiv$ CPK volume)**

It is an old measure of molecular volume obtained by immersing CPK models of molecules in a liquid.

- **geometric volume (V^{geom})**

Defined as the volume of the solid geometric shape of the molecule assuming the atoms as point masses. All the atoms in the molecule are interconnected in such a way that several regular and irregular tetrahedrons are formed, their volumes being respectively computable analytically and numerically. Therefore, the geometric volume is obtained by subtracting the common volume from the sum of the volumes of the constituent tetrahedrons [Bhattacharjee, Rao *et al.*, 1991; Bhattacharjee, Basak *et al.*, 1992; Bhattacharjee, 1994; Bhattacharjee and Dasgupta, 1994].

- **molecular volume index (MVI)**

This is a volume descriptor for steric effects defined as [Cheng and Yuan, 2006]

$$\text{MVI} = \sum_{i=1}^A \sum_{j=i+1}^A \frac{V_i^{\text{vdw}} \cdot V_j^{\text{vdw}}}{d_{ij}^2}$$

where the summation goes over all the pairs of nonhydrogen atoms, V^{vdw} is the $\rightarrow$ *van der Waals volume* of each group, the group being here intended as an atom plus the bonded hydrogen atoms, and d_{ij}^2 is the square $\rightarrow$ *topological distance* between atoms v_i and v_j . Van der Waals volumes for some atom-types are collected in Table V5. It should be noted that contribution of hydrogens to group volume is remarkable; for example, the increasing in percentage from C to CH_3 volume is about 82%.

Table V5 The van der Waals volumes (V^{vdw} , $10^{-2} \text{ \AA}^3$) of some atoms and functional groups.

Group	V^{vdw}	Group	V^{vdw}	Group	V^{vdw}	Group	V^{vdw}
H	0.056	S	0.244	I	0.388	NH	0.197
C	0.206	F	0.115	CH	0.262	NH ₂	0.253
N	0.141	Cl	0.244	CH ₂	0.318	NH ₃	0.309
O	0.115	Br	0.287	CH ₃	0.374	OH	0.171

📖 [Edward, 1982b; Meyer, 1985b, 1986c, 1988a, 1989; Dubois and Loukianoff, 1993; Jaworska and Schultz, 1993; Barratt, Basketter *et al.*, 1994; Connolly, 1994; van Haelst, Paulus *et al.*, 1997; Buchwald, 2000]

- **volume profiles** $\rightarrow$ molecular profiles
- **volume-to-surface profiles** $\rightarrow$ molecular profiles
- **Voronoi binding site models** $\rightarrow$ distance geometry
- **Voronoi Field Analysis** $\rightarrow$ grid-based QSAR techniques
- **Voronoi polyhedra** $\rightarrow$ grid-based QSAR techniques (⊙ Voronoi field analysis)
- **VRA indices** $\rightarrow$ spectral indices (⊙ eigenvalues of the adjacency matrix)
- **VRD indices** $\rightarrow$ spectral indices (⊙ eigenvalues of the distance matrix)
- **VSW descriptor** $\rightarrow$ biodescriptors (⊙ amino acid descriptors)
- **VTI indices** $\rightarrow$ local invariants (⊙ LOfcal Vertex Invariants)

W

- **walk** → graph
- **walk connectivity indices** → connectivity indices

■ walk counts

Walk counts are atomic and molecular descriptors obtained from an H -depleted molecular graph G , based on the graph → *walk* [Rücker and Rücker, 1993, 1994; Lu, Guo *et al.*, 2006b]. A random walk in a molecular graph is associated with a probability measure. There are two basic types of random walks depending on the probability measure [Klein, Palacios *et al.*, 2004]. **Simple random walks** are designated by a probability measure that entails starting walks from each vertex with equal probability and subsequent steps are such that each neighbor is stepped to with equal probability; this means that the probability of stepping from vertex i to vertex j is $1/\delta_i$, where δ_i is the → *vertex degree* of the i th vertex. **Equipose random walks** are walks associated with a probability measure that takes each possible walk of a given length as equally probable.

Let A be the → *adjacency matrix* of a graph G and A^k its k th → *power matrix* ($k = 1, 2, \dots, A - 1$) whose elements are denoted by $a_{ij}^{(k)}$ and A is the number of graph vertices. The $a_{ij}^{(k)}$ element of the A^k matrix corresponds to the number of equipose random walks of length k (*walk count*) from vertex v_i to vertex v_j . The diagonal entry $a_{ii}^{(k)}$ is the number of → *self-returning walks* for the i th vertex, that is, the random walks starting and ending at the i th vertex. Therefore, the **atomic walk count** of order k for the i th atom (also called k th-order **walk degree** ${}^k W_i$), denoted by $awc_i^{(k)}$, is calculated as the sum of the i th row entries of the k th power adjacency matrix A^k :

$$awc_i^{(k)} \equiv {}^k W_i \equiv VS_i(A^k) = \sum_{j=1}^A a_{ij}^{(k)}$$

where the symbol VS stands for the matrix → *row sum operator*. The atomic walk count $awc_i^{(k)}$ is the total number of equipose walks of length k starting from vertex v_i . → *Self-returning walk counts* are a particular case of walk counts with several important properties.

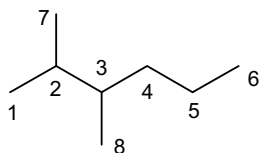
For $k = 1$, the matrix A^1 is simply the → *adjacency matrix* and therefore the atomic walk count coincides with the → *vertex degree* δ_i , that is,

$$awc_i^{(1)} = \sum_{j=1}^A a_{ij} \equiv {}^1 P_i \equiv \delta_i$$

where ${}^1 P_i$ is the → *atomic path count* of first order.

Example W1

Adjacency matrix **A** and its power matrices (order 2, 3, 4, 5) for 2,3-dimethylhexane. VS_i and CS_j are the matrix row and column sums, respectively; they give atomic walk counts.



A									
	1	2	3	4	5	6	7	8	VS_i
1	0	1	0	0	0	0	0	0	1
2	1	0	1	0	0	0	1	0	3
3	0	1	0	1	0	0	0	1	3
4	0	0	1	0	1	0	0	0	2
5	0	0	0	1	0	1	0	0	2
6	0	0	0	0	1	0	0	0	1
7	0	1	0	0	0	0	0	0	1
8	0	0	1	0	0	0	0	0	1
CS_j	1	3	3	2	2	1	1	1	14

A²									
	1	2	3	4	5	6	7	8	VS_i
1	1	0	1	0	0	0	1	0	3
2	0	3	0	1	0	0	0	1	5
3	1	0	3	0	1	0	1	0	6
4	0	1	0	2	0	1	0	1	5
5	0	0	1	0	2	0	0	0	3
6	0	0	0	1	0	1	0	0	2
7	1	0	1	0	0	0	1	0	3
8	0	1	0	1	0	0	0	1	3
CS_j	3	5	6	5	3	2	3	3	30

A³									
	1	2	3	4	5	6	7	8	VS_i
1	0	3	0	1	0	0	0	1	5
2	3	0	5	0	1	0	3	0	12
3	0	5	0	4	0	1	0	3	13
4	1	0	4	0	3	0	1	0	9
5	0	1	0	3	0	2	0	1	7
6	0	0	1	0	2	0	0	0	3
7	0	3	0	1	0	0	0	1	5
8	1	0	3	0	1	0	1	0	6
CS_j	5	12	13	9	7	3	5	6	60

A⁴									
	1	2	3	4	5	6	7	8	VS_i
1	3	0	5	0	1	0	3	0	12
2	0	11	0	6	0	1	0	5	23
3	5	0	12	0	5	0	5	0	27
4	0	6	0	7	0	3	0	4	20
5	1	0	5	0	5	0	1	0	12
6	0	1	0	3	0	2	0	1	7
7	3	0	5	0	1	0	3	0	12
8	0	5	0	4	0	1	0	3	13
CS_j	12	23	27	20	12	7	12	13	126

A⁵									
	1	2	3	4	5	6	7	8	VS_i
1	0	11	0	6	0	1	0	5	23
2	11	0	22	0	7	0	11	0	51
3	0	22	0	17	0	5	0	12	56
4	6	0	17	0	10	0	6	0	39
5	0	7	0	10	0	5	0	5	27
6	1	0	5	0	5	0	1	0	12
7	0	11	0	6	0	1	0	5	23
8	5	0	12	0	5	0	5	0	27
CS_j	23	51	56	39	27	12	23	27	258

The atomic walk count is a measure of something like “involvedness” or centrality of the atom in the graph, that is, a measure of the complexity of the vertex environment. Moreover, the atomic walk count is the $\rightarrow$ *extended connectivity* defined by Morgan [Razinger, 1982; Figueras, 1993; Rucker and Rucker, 1993; Lu, Guo *et al.*, 2006b].

The atomic walk count can also be evaluated by iterative summation of the walk degrees over all the first neighbors δ_i [Diudea, Topan *et al.*, 1994]:

$$awc_i^{(k+1)} \equiv {}^{k+1}W_i = [{}^{k+1}\mathbf{W}_C]_{ii} = \sum_{j=1}^A [C]_{ij} \cdot [{}^k\mathbf{W}_C]_{jj}; \quad [{}^0\mathbf{W}_C]_{jj} = 1$$

$$[{}^{k+1}\mathbf{W}_C]_{ij} = [{}^k\mathbf{W}_C]_{ij} = [C]_{ij}$$

where ${}^k\mathbf{W}_C$ is a matrix having k th-order walk degrees of atoms in the main diagonal and the $\rightarrow$ conventional bond order for each entry corresponding to pairs of bonded atoms and zero otherwise. This matrix is derived from the $\rightarrow$ atom connectivity matrix C ; therefore, the walk degrees obtained in such a way also account for bond multiplicity in the molecular graph.

To account also for heteroatoms, the $\rightarrow$ augmented adjacency matrix can be used instead of the simple adjacency matrix in the calculation of **weighted walk counts**, where atomic weights are placed on the main diagonal to distinguish among different atom types.

The distribution of simple random walks can be generated by powers of the **random walk Markov matrix**, which is a $\rightarrow$ weighted adjacency matrix belonging to the class of $\rightarrow$ stochastic matrices, defined as [Klein, Palacios *et al.*, 2004]

$$[\mathbf{MM}]_{ij} = \begin{cases} \frac{1}{\delta_j} & \text{if } i \neq j \wedge (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

where δ_j is the $\rightarrow$ vertex degree of the j th vertex adjacent to vertex i , $E(G)$ is the set of graph edges. It should be noted that this Markov matrix can be obtained as

$$\mathbf{MM} = \mathbf{A} \cdot \mathbf{V}^{-1}$$

where $\mathbf{A}$ and $\mathbf{V}$ are the $\rightarrow$ adjacency matrix and the diagonal $\rightarrow$ vertex degree matrix, respectively. The sum of the elements of the Markov matrix $\mathbf{MM}$ is equal to the number A of vertices of the graph. Moreover, another random walk-based matrix $\mathbf{H}$ was defined as [Klein, Palacios *et al.*, 2004]

$$\mathbf{H} = \mathbf{V}^{-1/2} \cdot \mathbf{MM} \cdot \mathbf{V}^{1/2}$$

Unlike matrix $\mathbf{MM}$, which usually is unsymmetrical, the matrix $\mathbf{H}$ is symmetric and related to $\mathbf{MM}$ by a similarity transformation, so that $\mathbf{H}$ and $\mathbf{MM}$ have the same eigenvalues and interrelated eigenvectors. The matrix $\mathbf{H}$ is also related to the $\rightarrow$ Laplacian matrix; moreover, the half sum of the elements of $\mathbf{H}$ coincides with the $\rightarrow$ Randić connectivity index:

$${}^1\chi = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{H}]_{ij}$$

Each off-diagonal element $[\mathbf{MM}^k]_{ij}$ of the k th power of the Markov matrix is interpreted as the probability for a simple random walk of length k beginning at vertex j to end at vertex i ; each diagonal element $[\mathbf{MM}^k]_{ii}$ is the probability for a simple random walk starting at vertex i to return to vertex i , after k steps.

The row sum of the k th power of the random walk Markov matrix $\mathbf{MM}^k$, called **random walk count**, is a local vertex invariant based on simple random walks defined by analogy with the atomic walk count:

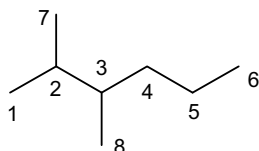
$$rwc_i^{(k)} = \sum_{j=1}^A [\mathbf{MM}^k]_{ij}$$

The random walk count is interpreted as A times the probability that a walk, after k steps, ends at vertex i after starting randomly at any vertex.

Note each column sum of $\mathbf{M}^k$ is always equal to 1, since it is the sum of the probabilities for walks of length k starting from vertex i to end at some other vertex in the graph.

Example W2

Random walk Markov matrix **MM** and its power matrices (order 2, 3, 4, 5) for 2,3-dimethylhexane. VS_i and CS_i are the matrix row and column sums, respectively; they give the random walk counts.



MM

	1	2	3	4	5	6	7	8	VS_i
1	0	1/3	0	0	0	0	0	0	1/3
2	1	0	1/3	0	0	0	1	0	7/3
3	0	1/3	0	1/2	0	0	0	1	11/6
4	0	0	1/3	0	1/2	0	0	0	5/6
5	0	0	0	1/2	0	1	0	0	3/2
6	0	0	0	0	1/2	0	0	0	1/2
7	0	1/3	0	0	0	0	0	0	1/3
8	0	0	1/3	0	0	0	0	0	1/3
CS_j	1	1	1	1	1	1	1	1	8

 MM^2 [illegible] MM^3 [illegible]
$$\mathbf{MM}^4$$
[illegible]MM⁵[illegible]

Atomic walk counts are used to build $\rightarrow$ *walk-sequence matrix* **SW**, from which several graph invariants can be derived (Table W1).

Table W1 Outline of a generic walk-sequence matrix **SW**.

Atom ID	Walk length k					Atomic walk count sums
	0	1	2	...	A-1	
1	1	$awc_1^{(1)}$	$awc_1^{(2)}$	...	$awc_1^{(A-1)}$	$awcs_1$
2	1	$awc_2^{(1)}$	$awc_2^{(2)}$	...	$awc_2^{(A-1)}$	$awcs_2$
...	...	...	...	...	...	...
...	...	...	...	...	...	...
A	1	$awc_A^{(1)}$	$awc_A^{(2)}$	...	$awc_A^{(A-1)}$	$awcs_A$
Molecular walk counts	A	$mwc^{(1)}$	$mwc^{(2)}$	...	$mwc^{(A-1)}$	TWC

The **walk count atomic code** of each i th atom is the ordered sequence of atomic walk counts of increasing length:

$$\{awc_i^{(1)}, awc_i^{(2)}, \dots, awc_i^{(A-1)}\}$$

Walks starting from two different vertices v_i and v_j are called **equipotent walks** if their walk count atomic codes are the same. Moreover, if the vertices v_i and v_j are nonequivalent, equipotent walks are called **unusual walks** and the corresponding vertices **unusual vertices** [Randić, Woodworth *et al.*, 1983].

Moreover, the **atomic walk count sum** of the i th atom is the sum of all walks of any length starting from the i th atom:

$$awcs_i \equiv VS_i(\mathbf{SW}) = \sum_{k=1}^{A-1} awc_i^{(k)}$$

where the symbol *VS* stands for the $\rightarrow$ *vertex sum operator*. Arranging in increasing order the A $awcs$ indices, a new molecular vector descriptor, called **ordered walk count molecular code**, is obtained:

$$\{awcs_{s(1)}, awcs_{s(2)}, \dots, awcs_{s(A)}\}$$

where $s(i)$ is an index for ordered sequence. The $awcs$ ranking is related to the increasing complexity of the atom molecular environment.

The **molecular walk count** $mwc^{(k)}$ of length k (also called **walk number** or **graph walk count**, $GWC^{(k)}$) is the total number of walks of length k in the molecular graph and, for any k different from zero, it is calculated as the half sum of all atomic walk counts of the same length k , that is, as the half sum of the entries in each column of the **SW** matrix:

$$mwc^{(k)} = \frac{1}{2} \cdot CS_k(\mathbf{SW}) = \frac{1}{2} \cdot \sum_{i=1}^A awc_i^{(k)} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A a_{ij}^{(k)}$$

where CS_k is the $\rightarrow$ *column sum operator* and $a_{ij}^{(k)}$ denotes the elements of the k th power of the adjacency matrix **A**. The molecular walk count of zero order $mwc^{(0)}$ is simply equal to the number A of graph vertices.

An alternative method for counting walks in a graph consists in calculating the eigenvectors and eigenvalues of the adjacency matrix for various values of k :

$$mwc^{(k)} = \sum_{i=1}^A \left(\sum_{j=1}^A \ell_{ij} \right)^2 \cdot \lambda_i^k$$

where ℓ_{ij} denotes the coefficients of the i th eigenvector and λ_i the corresponding eigenvalue [Lukovits, Miličević *et al.*, 2002].

The molecular walk count is related to $\rightarrow$ *molecular branching* and size and in general to the $\rightarrow$ *molecular complexity* of the graph. It was found that $mwc^{(k)}$ is directly related to the $\rightarrow$ *Lovasz–Pelikan index*, that is, the largest eigenvalue of the adjacency matrix [Cvetković and Gutman, 1977].

Table W2 Some molecular walk counts for C8 data set (Appendix C – Set 1); data are reported in log unit scale.

C8	$mwc^{(1)}$	$mwc^{(2)}$	$mwc^{(3)}$	$mwc^{(4)}$	$mwc^{(5)}$	$mwc^{(6)}$	$mwc^{(7)}$	$mwc^{(8)}$	$mwc^{(9)}$	$mwc^{(10)}$
n-octane	7	3.296	3.892	4.511	5.13	5.759	6.385	7.017	7.645	8.277
2M	7	3.367	3.97	4.654	5.283	5.974	6.611	7.303	7.943	8.636
3M	7	3.367	4.007	4.691	5.361	6.052	6.73	7.422	8.103	8.796
4M	7	3.367	4.007	4.71	5.371	6.084	6.751	7.468	8.136	8.854
3E	7	3.367	4.043	4.745	5.442	6.151	6.853	7.563	8.267	8.977
22MM	7	3.497	4.111	4.934	5.58	6.409	7.066	7.895	8.557	9.386
23MM	7	3.434	4.111	4.844	5.557	6.290	7.013	7.745	8.472	9.202
24MM	7	3.434	4.078	4.828	5.493	6.248	6.919	7.674	8.346	9.101
25MM	7	3.434	4.043	4.779	5.425	6.151	6.815	7.53	8.204	8.912
33MM	7	3.497	4.174	4.99	5.694	6.518	7.226	8.052	8.762	9.589
34MM	7	3.434	4.143	4.875	5.609	6.347	7.086	7.825	8.565	9.304
2M3E	7	3.434	4.143	4.890	5.617	6.368	7.099	7.851	8.584	9.336
3M3E	7	3.497	4.234	5.030	5.79	6.592	7.354	8.157	8.921	9.724
223MMM	7	3.555	4.263	5.094	5.838	6.666	7.419	8.245	9.001	9.827
224MMM	7	3.555	4.174	5.056	5.687	6.578	7.212	8.105	8.74	9.634
233MMM	7	3.555	4.29	5.106	5.883	6.693	7.479	8.288	9.076	9.884
234MMM	7	3.497	4.205	4.977	5.714	6.488	7.229	8.004	8.746	9.521
2233MMMM	7	3.664	4.394	5.273	6.075	6.925	7.749	8.588	9.419	10.255

A **normalized atomic walk count** of the i th atom can be defined from the previous indices by

$$\overline{awc}_i^{(k)} = \frac{awc_i^{(k)}}{mwc^{(k)}}$$

and can be considered as a $\rightarrow$ *weighting scheme* for graph vertices.

The **total walk count** TWC is the total number of walks of any length in the graph and is calculated as

$$TWC = \sum_{k=0}^{A-1} mwc^{(k)} = A + \frac{1}{2} \cdot \sum_{k=1}^{A-1} \sum_{i=1}^A awc_i^{(k)} = A + \frac{1}{2} \cdot \sum_{i=1}^A awc_i$$

The total walk count is a measure of $\rightarrow$ *molecular complexity*, increasing with size, branching, and cyclicity. Moreover, it was demonstrated that the total walk count increases with saturation,

that is, with increasing edge weights and with increasing vertex weights. A heteroatom is represented in the augmented adjacency matrix by a diagonal entry more than zero and thus will result in an increased sum of the matrix entries. In general, a heteroatom influences the total walk count more the more central its position within the graph; a second heteroatom increases *TWC* more if it is closer to the first heteroatom. *TWC* is also called **labyrinthicity** as it is a measure of how many possibilities are given to walk or oscillate through the graph along its edges and loops [Rücker and Rücker, 2000].

The $\rightarrow$ *walk connectivity indices* are molecular descriptors defined by analogy with the $\rightarrow$ *Randić connectivity index* by using the atomic walk counts in place of the vertex degrees [Razinger, 1986].

$\rightarrow$ *Local vertex invariants* proposed to account for both heteroatoms and multiple bonds are the **electronegativity-weighted walk degrees** ${}^k W_E$, defined as [Diudea, Topan *et al.*, 1994]

$${}^k W_{E,i} = {}^k W_i \cdot {}^k t_i$$

where ${}^k t_i$ is a weighting factor accounting for heteroatoms and multiple bonds derived by a recursive formula as

$${}^{k+1} t_i = \left[\prod_{j=1}^{\delta_i} \binom{{}^k t_j}{\pi_{ij}^*} \right]^{1/\sum_j \pi_{ij}^*}$$

where the product runs over the weights of the first neighbors of the *i*th atom each raised to the conventional bond order π_{ij}^* ; vertex weights at the first step ($k=1$) are valence group carbon-related electronegativities (Table W3) [Diudea, Kacso *et al.*, 1996].

Table W3 Valence group electronegativities used for the calculation of the electronegativity-weighted walk degrees [Diudea, Kacso *et al.*, 1996].

Atom/Hybrid	${}^1 t$	Atom/Hybrid	${}^1 t$	Atom/Hybrid	${}^1 t$
$>C<$	1.0000	$-CHBr_2$	1.0672	$-H$	0.9175
$>C=$	1.0747	$-CHI_2$	0.9914	$-N<$	1.2234
$-C\equiv$	1.1476	$-CF_3$	1.3260	$=N-$	1.3147
$=C=$	1.1581	$-CCl_3$	1.1932	$\equiv N$	1.5288
$>CH-$	0.9716	$-CBr_3$	1.1266	$-NH-$	1.1021
$=CH-$	1.0441	$-Cl_3$	1.0088	$=NH$	1.2474
$\equiv CH$	1.2142	$>C=O$	1.2397	$-NH_2$	1.0644
$-CH_2-$	0.9622	$-CH=O$	1.1596	$-NHCH_3$	1.0379
$=CH_2$	1.0891	$-COOH$	1.2220	$-N(CH_3)_2$	1.0292
$-CH_3$	0.9575	$-O-$	1.4634	$-C\equiv N$	1.2377
$-CH_2F$	1.0674	$=O$	1.6564	$-P<$	0.8988
$-CH_2Cl$	1.0305	$-OH$	1.2325	$=P-$	0.9658
$-CH_2Br$	1.0110	$-OCH_3$	1.1248	$-PH-$	0.9124
$-CH_2I$	0.9744	$-S-$	1.1064	$-PH_2$	0.9170
$-CH_2OH$	1.0228	$=S$	1.2523	$-F$	1.6514
$-CH_2SH$	0.9804	$-SCH_3$	1.0073	$-Cl$	1.3717
$-CHF_2$	1.1897	$-NO$	1.4063	$-Br$	1.2447
$-CHCl_2$	1.1089	$-NO_2$	1.4861	$-I$	1.0262

☞ [Randić, 1980c, 1995c; Gao and Hosoya, 1988; Kunz, 1989; Shalabi, 1991; Rücker and Rücker, 2000, 2001, 2003; Gutman, Rücker *et al.*, 2001; Palacios, 2001; Lukovits, Miličević *et al.*, 2002; Lukovits and Trinajstić, 2003; Braun, Kerber *et al.*, 2005; Vukicević, Miličević *et al.*, 2005]

- **walk count atomic code** → walk counts
- **walk degree** $\equiv$ *atomic walk count* → walk counts
- **walk degree layer matrix** → layer matrices
- **walk diagonal matrix** → walk matrices
- **W^* index** → weighted matrices ($\odot$ weighted distance matrices)
- **walk length** → graph

■ **walk matrices** ($\equiv$ *random walk matrices*)

The **walk diagonal matrix**, denoted by ${}^k W_M$, of k th order is a diagonal matrix whose diagonal elements are the k th-order **weighted walk degrees** ${}^k W_{M,i}$, that is, the sum of the weights (the property collected in matrix M) of all walks of length k starting from the i th vertex to any other vertex in the graph, directly calculated as

$${}^k W_{M,i} = [{}^k W_M]_{ii} = \sum_{j=1}^A [M^k]_{ij}$$

where A is the total number of the vertices in the graph and M^k is the k th power of the matrix M , which can be any square $A \times A$ topological matrix [Diudea, 1996a; Diudea and Randić, 1997]. An alternative algorithm to obtain the weighted walk degrees was proposed by Diudea and was called *${}^k W_M$ algorithm*. It is an algorithm based on the iterative summation of vertex contributions over all column entries of the i th vertex considered, defined as

$$\begin{aligned} M + I &= {}^0 W_M \\ {}^{k+1} W_{M,i} &= [{}^{k+1} W_M]_{ii} = \sum_{j=1}^A a_{ij} \cdot [M]_{ij} \cdot [{}^k W_M]_{ii} \\ [{}^{k+1} W_M]_{ij} &= [{}^k W_M]_{ij} = [M]_{ij} \end{aligned}$$

where a_{ij} denotes the elements of the adjacency matrix and the sum in the second expression effectively accounts only for contributions from vertices adjacent to the i th vertex (i.e., $a_{ij} = 1$).

If $M = A$, A being the → *adjacency matrix*, then the weighted walk degree of the i th atom ${}^k W_{M,i}$ is just the → *atomic walk count* of length k of the i th atom, that is, the number of all walks of length k starting from the i th vertex to any other vertex in the graph.

The **weighted walk numbers** are molecular descriptors obtained as the half sum of the weighted walk degrees of all vertices in the graph:

$${}^k W_M = \frac{1}{2} \cdot \sum_{i=1}^A {}^k W_{M,i} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [M^k]_{ij}$$

In the case of $M = A$, the global walk number ${}^k W_A$ is just the → *molecular walk count* of k th order $mw_c^{(k)}$, that is, the total number of k th length walks in the graph.

Based on weighted walk degrees of different orders calculated by the algorithm defined above, the **walk matrix** $W_{(M_1, M_2, M_3)}$ is an unsymmetrical square $A \times A$ matrix whose i - j entry is

defined as

$$[\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}]_{ij} = [\mathbf{M}_2]_{ij} W_{\mathbf{M}_1, i} [\mathbf{M}_3]_{ij}$$

where $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ are any square $A \times A$ matrices, $W_{\mathbf{M}_1, i}$ is the weighted walk degree of the i th vertex, based on the property collected in matrix $\mathbf{M}_1$, $[\mathbf{M}_2]_{ij}$ is the i - j th entry of the $\mathbf{M}_2$ matrix giving the length of the walk, and the i - j th entry $[\mathbf{M}_3]_{ij}$ of the third matrix is used as a weighting factor. The diagonal entries are usually equal to zero.

Summing the entries in each row, $\rightarrow$ *local vertex invariants* representing a sum of weighted walk degrees of different orders are obtained:

$$W_{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, i} = \sum_{j=1}^A \left([\mathbf{M}_2]_{ij} W_{\mathbf{M}_1, i} \cdot [\mathbf{M}_3]_{ij} \right)$$

Global weighted walk numbers $W_{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3}$ are obtained applying the $\rightarrow$ *Wiener operator* Wi to the walk matrix as

$$W_{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3} \equiv Wi(\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \left([\mathbf{M}_2]_{ij} W_{\mathbf{M}_1, i} [\mathbf{M}_3]_{ij} \right)$$

Appropriate combinations of $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ give various unsymmetrical walk matrices and hence various local and molecular descriptors.

For example, the combination $(\mathbf{A}, \mathbf{1}, \mathbf{1})$, where $\mathbf{A}$ is the adjacency matrix and $\mathbf{1}$ represents a matrix whose off-diagonal elements are all equal to 1, gives a walk matrix where in each i th row the vertex degree of the i th vertex considered is repeated in all the columns, the vertex degree being the row sum of the adjacency matrix and coincident with the first-order atomic walk count $awc_i^{(1)}$. If the matrix $\mathbf{A}$ is substituted by a general matrix $\mathbf{M}_1$, the row sums of $\mathbf{M}_1$ are collected in the rows of the walk matrix.

Next, considering the combination $(\mathbf{M}_1, \mathbf{1}, \mathbf{M}_3)$, the corresponding walk matrix is obtained simply by the following relationship:

$$\mathbf{W}_{(\mathbf{M}_1, \mathbf{1}, \mathbf{M}_3)} = \mathbf{W}_{(\mathbf{M}_1, \mathbf{1}, \mathbf{1})} \otimes \mathbf{M}_3$$

where the symbol $\otimes$ represents the $\rightarrow$ *Hadamard matrix product*. Moreover, the global walk number $W_{\mathbf{M}_1, \mathbf{1}, \mathbf{M}_3}$ is equal to the half sum of the entries of the matrix product $\mathbf{M}_1 \mathbf{M}_3$:

$$W_{\mathbf{M}_1, \mathbf{1}, \mathbf{M}_3} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{W}_{(\mathbf{M}_1, \mathbf{1}, \mathbf{M}_3)}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{M}_1 \cdot \mathbf{M}_3]_{ij}$$

This relationship was demonstrated recalling that the sum of the entries of the matrix product $\mathbf{M}_1 \mathbf{M}_3$ is equal to the product of the $\mathbf{M}_1$ column sum vector for the $\mathbf{M}_3$ row sum vector:

$$S(\mathbf{M}_1 \mathbf{M}_3) = \mathbf{u}^T \mathbf{M}_1 \mathbf{M}_3 \mathbf{u} = \mathbf{cs}(\mathbf{M}_1) \cdot \mathbf{rs}(\mathbf{M}_3)$$

where S is the $\rightarrow$ *total sum operator*, $\mathbf{u}$ is a column unit vector, $\mathbf{cs}$ and $\mathbf{rs}$ the $\rightarrow$ *column sum vector* and the $\rightarrow$ *row sum vector*, respectively.

The global walk number obtained by the particular walk matrix defined by $(\mathbf{A}, \mathbf{1}, \mathbf{D})$, where $\mathbf{D}$ is the $\rightarrow$ distance matrix, coincides with the half of the Schultz $\rightarrow S$ index.

Moreover, if also $\mathbf{M}_1 = \mathbf{M}_3$, the global walk number $W_{\mathbf{M}_1, \mathbf{1}, \mathbf{M}_1}$ represents the half sum of the square $\mathbf{M}_1$ matrix. If $\mathbf{M}_2$ is a matrix having all nondiagonal entries equal to an integer n , the global walk number is $^{n+1}W_{\mathbf{M}_1, n, \mathbf{M}_1}$ of order $n + 1$ coinciding with the walk number $^{n+1}W_{\mathbf{M}_1}$.

An interesting walk matrix is $\mathbf{W}_{(\mathbf{A}, \mathbf{D}, \mathbf{1})}$, where $\mathbf{M}_1$ is the adjacency matrix $\mathbf{A}$, $\mathbf{M}_2$ is the distance matrix $\mathbf{D}$, and $\mathbf{M}_3$ is a matrix whose elements are equal to 1. In this matrix, the atomic walk counts of increasing orders are collected where the $\rightarrow$ topological distance between each pair of vertices gives the order.

The reciprocal walk matrix $\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}^{-1}$ is a square $A \times A$ matrix whose entries are the reciprocal of the corresponding entries of the walk matrix $\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}$:

$$\left[\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)}^{-1} \right]_{ij} = \left[\mathbf{W}_{(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3)} \right]_{ij}^{-1}$$

The diagonal entries are always zero.

$\rightarrow$ Harary indices are derived from these matrices by applying the $\rightarrow$ Wiener operator.

A restricted random walk matrix **RRW** was also proposed [Randić, 1995c] as a $A \times A$ dimensional square unsymmetrical matrix that enumerates restricted (i.e., selected) random walks over a molecular graph G . The i - j entry of the matrix is the probability of a random walk starting at vertex v_i and ending at vertex v_j of the length equal to the topological distance d_{ij} between the vertices considered:

$$[\mathbf{RRW}]_{ij} = \begin{cases} \frac{[\mathbf{A}^{d_{ij}}]_{ij}}{d_{ij} W_i} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases}$$

where $d_{ij} W_i$ is the d_{ij} -order atomic walk count of the i th vertex, that is, the number of walks of length d_{ij} starting at vertex v_i (the elements of the walk matrix $\mathbf{W}_{(\mathbf{A}, \mathbf{D}, \mathbf{1})}$); $[\mathbf{A}^{d_{ij}}]_{ij}$ is the number of all walks of length d_{ij} starting from vertex v_i and ending at vertex v_j ; it is equal to 1 for any pair of vertices in acyclic graphs while it can be greater than 1 for a pair of vertices in cyclic graphs; $\mathbf{A}^{d_{ij}}$ is the power equal to the distance d_{ij} of the adjacency matrix $\mathbf{A}$. Therefore, to build the **RRW** matrix the powers of the adjacency matrix $\mathbf{A}$ are used to calculate walk counts of increasing length and the distance matrix $\mathbf{D}$ to impose the restrictions on the length of walks.

A matrix $\mathbf{Q}$ was defined by analogy to **RRW**, but based on $\rightarrow$ simple random walks instead of equipoise random walks. The off-diagonal elements of the matrix $\mathbf{Q}$ are probabilities for a simple random walk of length d_{ij} starting from vertex j to end at vertex i and are calculated as [Klein, Palacios *et al.*, 2004]:

$$[\mathbf{Q}]_{ij} = \begin{cases} \frac{[\mathbf{MM}^{d_{ij}}]_{ij}}{d_{ij} W_j} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases}$$

where $\mathbf{MM}$ is the $\rightarrow$ random walk Markov matrix.

For acyclic graphs, the restricted random walk matrix is simply the reciprocal walk matrix where $\mathbf{M}_1 = \mathbf{A}$, $\mathbf{M}_2 = \mathbf{D}$, and $\mathbf{M}_3 = \mathbf{1}$:

$$\mathbf{RRW} = \mathbf{W}_{(\mathbf{A}, \mathbf{D}, \mathbf{1})}^{-1}$$

Several matrix invariants can be calculated from the restricted random walk matrix such as the $\rightarrow$ *eigenvalues*, the $\rightarrow$ *determinant*, and the coefficients of the $\rightarrow$ *characteristic polynomial*. Moreover, $\rightarrow$ *weighted path counts* of m th order are obtained by summing all matrix entries corresponding to vertices separated by a distance of length m . The first-order weighted path count ^1R is always equal to the number A of vertices; the second-order weighted path count ^2R is a connection additive index that correlates well with some $\rightarrow$ *physico-chemical properties*. This sequence is a weighted $\rightarrow$ *molecular path code*

$$\{^1R, ^2R, \dots, ^LR\}$$

where L is equal to $A - 1$, A being the number of vertices. Summing all path numbers from the first order to $A - 1$ order, the corresponding $\rightarrow$ *molecular ID number* is obtained. It is called **restricted walk ID number (RWID)** and is defined as

$$RWID = \sum_{m=1}^{A-1} {}^mR = A + \sum_{m=2}^{A-1} {}^mR$$

📖 [Diudea, Minailiuc *et al.*, 1996; Diudea, 1996b, 1999; Diudea and Randić, 1997; Diudea and Gutman, 1998]

- **walk matrix** $\rightarrow$ walk matrices
- **walk number** $\equiv$ *molecular walk count* $\rightarrow$ walk counts
- **walk-sequence matrix** $\rightarrow$ sequence matrices
- **Watson nonmetric distance** $\rightarrow$ similarity/diversity (☉ Table S10)
- **Wave-Hedges distance** $\rightarrow$ similarity/diversity (☉ Table S7)
- **wavelet coefficient descriptors** $\rightarrow$ TAE descriptor methodology
- **wavelet analysis** $\rightarrow$ spectra descriptors
- **wavelet transforms** $\rightarrow$ spectra descriptors
- **WCD** $\equiv$ *wavelet coefficient descriptors* $\rightarrow$ TAE descriptor methodology
- **WDEN index** $\equiv$ *WHIM density* $\rightarrow$ WHIM descriptors (☉ global WHIM descriptors)
- **Weckwerth solute descriptors** $\rightarrow$ Linear Solvation Energy Relationships
- **weighted adjacency matrices** $\rightarrow$ weighted matrices
- **weighted atomic self-returning walk counts** $\rightarrow$ self-returning walk counts
- **weighted classification accuracy** $\rightarrow$ classification parameters
- **weighted combination of COSV and ISDFP** $\rightarrow$ molecular shape analysis
- **weighted detour matrix** $\rightarrow$ detour matrix
- **weighted distances** $\rightarrow$ similarity/diversity
- **weighted distance matrices** $\rightarrow$ weighted matrices
- **weighted edge adjacency matrix** $\rightarrow$ edge adjacency matrix
- **weighted edge degree** $\rightarrow$ edge adjacency matrix
- **weighted electronic connectivity matrix** $\rightarrow$ weighted matrices (☉ weighted adjacency matrices)
- **weighted electronic distance** $\rightarrow$ weighted matrices (☉ weighted adjacency matrices)
- **weighted graph** $\rightarrow$ graph
- **Weighted Holistic Invariant Molecular descriptors** $\equiv$ *WHIM descriptors*
- **weighted ID number** $\rightarrow$ ID numbers
- **weighted information indices by volume** $\rightarrow$ indices of neighborhood symmetry

■ weighted matrices

These matrices are derived from vertex- and/or edge-weighted molecular graphs representing molecules that contain heteroatoms and/or multiple bonds. A vertex- and edge-weighted molecular graph is obtained by defining a vertex weight set and an edge weight set according to a specific $\rightarrow$ *weighting scheme* w . Usually, in a vertex- and edge-weighted graph, hydrogen atoms are not considered, thus resulting in a $\rightarrow$ *H-depleted molecular graph*, and the weight of a vertex corresponding to a carbon atom is zero, whereas the weight of an edge corresponding to a carbon-carbon single bond is 1 [Ivanciuc, 2000i]. Typical vertex weights are physico-chemical $\rightarrow$ *atomic properties*, such as atomic numbers, atomic mass, and atomic electronegativity, and $\rightarrow$ *local vertex invariants* such as the $\rightarrow$ *vertex degree*, which is the number of adjacent vertices. Edge weights usually are bond parameters such as $\rightarrow$ *conventional bond order*, force constant, ionic character, $\rightarrow$ *dipole moment*, and $\rightarrow$ *bond distance*. Edge weights can also be derived from some combination of the weights of the two vertices incident on the edge.

Row sums of weighted matrices are $\rightarrow$ *local vertex invariants*, which can be used to calculate molecular descriptors encoding chemical information. In principle, all the molecular descriptors defined from matrices representing simple molecular graphs can be computed from weighted molecular graphs by using the same formulas applied to selected weighted matrices. In effect, all the $\rightarrow$ *graph-theoretical matrices* defined for a simple molecular graph can also be computed for a corresponding vertex- and edge-weighted graph so that new matrices and, accordingly, new descriptors are obtained, which encode different chemical information depending on the weighting scheme. The most common weighted graph-theoretical matrices are the weighted adjacency matrices and the weighted distance matrices; these are extensively presented below. Moreover, $\rightarrow$ *distance-degree matrices*, $\rightarrow$ *distance-path matrix*, $\rightarrow$ *distance complement matrix*, $\rightarrow$ *complementary distance matrix*, and their corresponding $\rightarrow$ *reciprocal matrices* were derived from weighted molecular graphs to generate $\rightarrow$ *Wiener-type indices* [Ivanciuc, 2000i]. Together with the Wiener-type indices, generally denoted by $W_i(\mathbf{M}, w)$, where $\mathbf{M}$ is the graph-theoretical matrix and w the weighting scheme, other classes of molecular descriptors were computed from weighted graphs, such as the $\rightarrow$ *Balaban-like indices* $J(\mathbf{M}, w)$, the $\rightarrow$ *spectral indices* $MinSp(\mathbf{M}, w)$ and $MaxSp(\mathbf{M}, w)$, and the $\rightarrow$ *Balaban-like information indices* $U(\mathbf{M}, w)$, $V(\mathbf{M}, w)$, $X(\mathbf{M}, w)$, and $Y(\mathbf{M}, w)$.

• weighted adjacency matrices

These are square $\rightarrow$ *sparse matrices* encoding information about adjacencies between graph vertices. The diagonal elements of a weighted adjacency matrix are not necessarily equal to zero and the off-diagonal elements corresponding to pairs of bonded atoms can be any real positive numbers.

While the vertex $\rightarrow$ *adjacency matrix* $\mathbf{A}$ and $\rightarrow$ *edge adjacency matrix* $\mathbf{E}$ contain information only about vertex and edge connectivities in the graph, weighted adjacency matrices allow to distinguish different bonds and/or atoms in a molecule.

When the adjacency between a pair of atoms is represented by a bond weight and/or each atom is represented by some atomic property, several weighted-vertex adjacency matrices called **atom connectivity matrices (ACM)** can be defined as [Spialter, 1963, 1964a, 1964b]

$$[\mathbf{A}(w)]_{ij} = \begin{cases} w_{ij} & \text{if } (i, j) \in E(G) \\ w_i & \text{if } i = j \\ 0 & \text{if } (i, j) \notin E(G) \end{cases}$$

where w_i depends on the chemical nature of the i th atom and is usually equal to zero for carbon atom; w_{ij} depends on the chemical nature of i th and j th bonded atoms as well as on the bond multiplicity [Ivanciuc, Ivanciuc *et al.*, 1997]. $E(G)$ is the set of the graph edges.

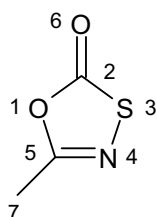
To account only for heteroatoms in the molecule, the **augmented adjacency matrix** ${}^aA$ was proposed by Randić [Randić, 1991b, 1991e; Randić and Dobrowolski, 1998] replacing the zero diagonal entries of the adjacency matrix of the simple graph with values characterizing different atoms in the molecule:

$$[{}^aA(w)]_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E(G) \\ w_i & \text{if } i = j \\ 0 & \text{if } (i,j) \notin E(G) \end{cases}$$

The diagonal elements w_i usually are some atomic physico-chemical properties or local vertex invariants; however, diagonal elements can also be atom-type variable parameters ($x, y, z, \dots$) that are optimized to enhance the estimate of the studied property by regression analysis [Randić and Pompe, 2001b; Lučić, Miličević *et al.*, 2003].

Example W3

Augmented adjacency matrix of 5-methyl-1,3,4-oxathiazol-2-one calculated on the H-depleted molecular graph whose vertices are weighted by relative atomic numbers Z ; VS_i indicates the matrix row sums.



$${}^aA(Z) =$$

Atom	1	2	3	4	5	6	7	VS_i
1	1.33	1	0	0	1	0	0	3.33
2	1	1	1	0	0	1	0	4
3	0	1	2.67	1	0	0	0	4.67
4	0	0	1	1.17	1	0	0	3.17
5	1	0	0	1	1	0	1	4
6	0	1	0	0	0	1.33	0	2.33
7	0	0	0	0	1	0	1	2

The calculation was based on the following atomic numbers: $Z_1 = 8$, $Z_2 = 6$, $Z_3 = 16$, $Z_4 = 7$, $Z_5 = 6$, $Z_6 = 8$, and $Z_7 = 6$.

From the variable augmented adjacency matrix, a number of topological indices, called $\rightarrow$ *variable descriptors*, are derived. Moreover, the row sums of an augmented adjacency matrix are $\rightarrow$ *local vertex invariants* encoding information about the connectivity of each atom and its atom type; therefore, they can be viewed as **augmented vertex degrees**. The inverse of the square root of the product of the augmented degrees of the vertices incident with a bond is used as bond weight in calculating the $\rightarrow$ *weighted path counts*.

The most known weighted adjacency matrix is the **atom connectivity matrix C** (or simply known as **connectivity matrix**), obtained from multigraphs using the $\rightarrow$ *conventional bond order* π^* to represent the adjacency of a pair of vertices [Spialter, 1963, 1964a, 1964b]:

$$[C]_{ij} = \begin{cases} \pi_{ij}^* & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

where π^* takes value 1 for single bonds, 2 for double bonds, 3 for triple bonds, and 1.5 for aromatic bonds. The atom connectivity matrix accounts for bond multiplicity but does not allow to distinguish different atoms in molecules.

Note that the connectivity matrix was also defined by Ivanciuc as [Ivanciuc, Ivanciuc *et al.*, 1999a]

$$[C]_{ij} = \begin{cases} 1/\pi_{ij}^* & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

where each edge in the graph is weighted by the reciprocal of the conventional bond order, a quantity related to the bond length.

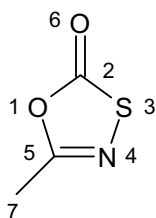
The **adjacency matrix of a multigraph** is a variant of the atom connectivity matrix defined as [Janežič, Miličević *et al.*, 2007]

$$[{}^m\mathbf{A}]_{ij} = \begin{cases} m_{ij} & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

where the element m_{ij} is the multiplicity of the edge connecting vertices v_i and v_j , which can take value 1 for single bonds, 2 for double bonds, and 3 for triple bonds; bonds belonging to aromatic systems are not uniquely identified since they can be represented as single or double bonds, depending on the Kekulé structure considered.

Example W4

Atom connectivity matrix **C** of 5-methyl-1,3,4-oxathiazol-2-one; VS_i indicates the matrix row sums.



$$\mathbf{C} =$$

Atom	1	2	3	4	5	6	7	VS_i
1	0	1	0	0	1	0	0	2
2	1	0	1	0	0	2	0	4
3	0	1	0	1	0	0	0	2
4	0	0	1	0	2	0	0	3
5	1	0	0	2	0	0	1	4
6	0	2	0	0	0	0	0	2
7	0	0	0	0	1	0	0	1

Note that for this molecule the adjacency matrix of the multigraph ${}^m\mathbf{A}$ coincides with the atom connectivity matrix **C** since no aromatic bonds are present in 5-methyl-1,3,4-oxathiazol-2-one.

Another variant of the connectivity matrix is the **adjacency matrix of a general graph** defined in such a way that all the connectivity matrix entries of $\pi^* = 1.5$ corresponding to aromatic bonds are differently defined [Marrero-Ponce, 2004a; Marrero-Ponce, Castillo-Garit *et al.*, 2004a]. The adjacency matrix of a general graph is defined as

$$[{}^g\mathbf{A}]_{ij} = \begin{cases} m_{ij} & \text{if } (i,j) \in E(G) \\ 1 & \text{if } i = j \wedge i \in \text{loop} \\ 0 & \text{otherwise} \end{cases}$$

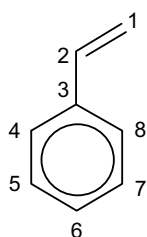
where m_{ij} is the bond multiplicity, which can take value 1 for single bonds, 2 for double bonds, and 3 for triple bonds. Aromatic rings with more than one canonical structure are represented

like a $\rightarrow$ *pseudograph*. This occurs for aromatic compounds such as benzene, pyridine, naphthalene, quinoline, and so on, for which the presence of π electrons is accounted for by means of loops in each atom of the aromatic ring. Bonds in these aromatic rings are represented by 1 instead of 1.5 in the adjacency matrix, but the incident atoms are represented by 1 instead of 0 in the matrix diagonal. Conversely, aromatic rings with only one canonical structure, such as furan, thiophene, and pyrrole, are represented like a $\rightarrow$ *multigraph*. Therefore, the bonds are represented by 1 or 2, depending on the specific bond, and the diagonal entries corresponding to the incident atoms are equal to zero.

The **molecular pseudograph's adjacency matrix** is the original name used to denote the adjacency matrix of a general graph; in this book, the latter name has been preferred because molecular graphs containing both aromatic rings represented by loops and multiple bonds are correctly referred to as general graphs instead of pseudographs. Derived from this matrix are the $\rightarrow$ *TOMOCOMD descriptors*.

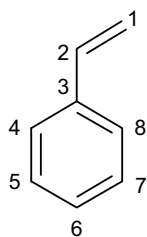
Example W5

Adjacency matrix of the multigraph ${}^m\mathbf{A}$ for two different canonical structures of styrene and the corresponding adjacency matrix of the general graph ${}^g\mathbf{A}$ and the atom connectivity matrix $\mathbf{C}$; VS_i indicates the matrix row sums.



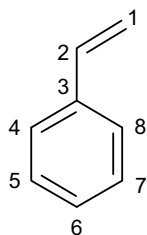
$$\mathbf{C} =$$

Atom	1	2	3	4	5	6	7	8	VS_i
1	0	2	0	0	0	0	0	0	2
2	2	0	1	0	0	0	0	0	3
3	0	1	0	1.5	0	0	0	1.5	4
4	0	0	1.5	0	1.5	0	0	0	3
5	0	0	0	1.5	0	1.5	0	0	3
6	0	0	0	0	1.5	0	1.5	0	3
7	0	0	0	0	0	1.5	0	1.5	3
8	0	0	1.5	0	0	0	1.5	0	3



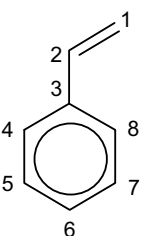
$${}^m\mathbf{A} =$$

Atom	1	2	3	4	5	6	7	8	VS_i
1	0	2	0	0	0	0	0	0	2
2	2	0	1	0	0	0	0	0	3
3	0	1	0	1	0	0	0	2	4
4	0	0	1	0	2	0	0	0	3
5	0	0	0	2	0	1	0	0	3
6	0	0	0	0	1	0	2	0	3
7	0	0	0	0	0	2	0	1	3
8	0	0	2	0	0	0	1	0	3



$${}^m\mathbf{A} =$$

Atom	1	2	3	4	5	6	7	8	VS_i
1	0	2	0	0	0	0	0	0	2
2	2	0	1	0	0	0	0	0	3
3	0	1	0	2	0	0	0	1	4
4	0	0	2	0	1	0	0	0	3
5	0	0	0	1	0	2	0	0	3
6	0	0	0	0	2	0	1	0	3
7	0	0	0	0	0	1	0	2	3
8	0	0	1	0	0	0	2	0	3

	Atom	1	2	3	4	5	6	7	8	VS _i
	1	0	2	0	0	0	0	0	0	2
	2	2	0	1	0	0	0	0	0	3
	3	0	1	1	1	0	0	0	1	4
	4	0	0	1	1	1	0	0	0	3
	5	0	0	0	1	1	1	0	0	3
	6	0	0	0	0	1	1	1	0	3
	7	0	0	0	0	0	1	1	1	3
	8	0	0	1	0	0	0	1	1	3

The $\rightarrow$ *bond length-weighted adjacency matrix* is another important connectivity matrix whose elements corresponding to pairs of adjacent vertices are bond lengths from computational chemistry and zero otherwise.

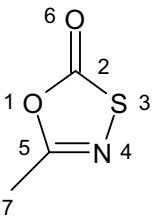
Moreover, another connectivity matrix is obtained by weighting each bond between vertices v_i and v_j by the $\rightarrow$ *edge connectivity* $(\delta_i \cdot \delta_j)^{-1/2}$, δ being the $\rightarrow$ *vertex degree* of the atoms. This matrix is known as **χ matrix** [Randić, 1992c] or more formally **edge- χ matrix** or **vertex-connectivity matrix** [Janežič, Miličević *et al.*, 2007] or **degree-adjacency matrix** [Rodriguez and Sigarreta, 2005] and is defined as

$$[\chi_e]_{ij} = \begin{cases} 1/\sqrt{\delta_i \cdot \delta_j} & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

Analogous connectivity matrices are obtained using the $\rightarrow$ *bond vertex degree* or $\rightarrow$ *valence vertex degree* in place of the classic vertex degree. The half sum of the off-diagonal elements of χ matrix is the $\rightarrow$ *Randić connectivity index*. Moreover, based on the different powers of the χ matrix, **higher order χ matrices**, denoted as χ^k , were defined from which $\rightarrow$ *$k\alpha$ descriptors* were calculated by applying the $\rightarrow$ *Wiener operator*.

Example W6

Edge- χ matrix (χ_e) of 5-methyl-1,3,4-oxathiazol-2-one; VS_i is the matrix row sum.

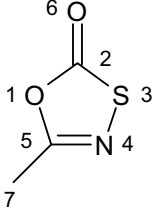
	Atom	1	2	3	4	5	6	7	VS _i
	1	0	0.408	0	0	0.408	0	0	0.816
	2	0.408	0	0.408	0	0	0.577	0	1.393
	3	0	0.408	0	0.500	0	0	0	0.908
	4	0	0	0.500	0	0.408	0	0	0.908
	5	0.408	0	0	0.408	0	0	0.577	1.393
	6	0	0.577	0	0	0	0	0	0.577
	7	0	0	0	0	0.577	0	0	0.577

The calculation was based on the following vertex degrees: $\delta_1=2$, $\delta_2=3$, $\delta_3=2$, $\delta_4=2$, $\delta_5=3$, $\delta_6=1$, and $\delta_7=1$. The Randić connectivity index is ${}^1\chi = 0.408 \times 4 + 0.577 \times 2 + 0.500 = 3.286$.

The **distance-sum-connectivity matrix** ${}^{\sigma}\chi_e$ is analogous to the χ matrix but based on the $\rightarrow$ vertex distance sum σ instead of the vertex degree δ [Szymanski, Müller *et al.*, 1986b]:

$$[{}^{\sigma}\chi_e]_{ij} = \begin{cases} 1/\sqrt{\sigma_i \cdot \sigma_j} & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

The distance-sum-connectivity matrix is used to calculate the $\rightarrow$ *Weighted ID number (WID)*.

Example W7									
Distance-sum-connectivity matrix of 5-methyl-1,3,4-oxathiazol-2-one; VS_i is the matrix row sum.									
 ${}^{\sigma}\chi_e =$	Atom	1	2	3	4	5	6	7	VS_i
	1	0	0.100	0	0	0.100	0	0	0.200
	2	0.100	0	0.095	0	0	0.082	0	0.277
	3	0	0.095	0	0.091	0	0	0	0.186
	4	0	0	0.091	0	0.095	0	0	0.186
	5	0.100	0	0	0.095	0	0	0.082	0.277
	6	0	0.082	0	0	0	0	0	0.082
	7	0	0	0	0	0.082	0	0	0.082
The calculation was based on the following vertex distance degrees: $\sigma_1 = \sigma_2 = \sigma_5 = 10$; $\sigma_3 = \sigma_4 = 11$; and $\sigma_6 = \sigma_7 = 15$.									

Zagreb matrices are a generalization of the χ matrix in terms of a variable exponent λ as

$$[ZM_e(\lambda)]_{ij} = \begin{cases} (\delta_i \cdot \delta_j)^\lambda & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

For $\lambda = -1/2$, the Zagreb matrix obviously reduces to the edge- χ matrix.

The **edge-Zagreb matrix**, denoted by ZM_e , was defined for $\lambda = 1$ as [Janežič, Miličević *et al.*, 2007]

$$[ZM_e]_{ij} = \begin{cases} \delta_i \cdot \delta_j & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

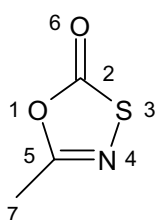
and the **modified edge-Zagreb matrix**, denoted by ${}^mZM_e$, was defined for $\lambda = -1$ as [Janežič, Miličević *et al.*, 2007]

$$[{}^mZM_e]_{ij} = \begin{cases} 1/(\delta_i \cdot \delta_j) & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

The half sum of the off-diagonal elements of the edge-Zagreb matrix is the $\rightarrow$ *second Zagreb index*, whereas the half sum of the off-diagonal elements of the modified edge-Zagreb matrix is the $\rightarrow$ *modified second Zagreb index*.

Example W8

Edge-Zagreb matrix ($\mathbf{ZM}_e$) and modified edge-Zagreb matrix (${}^m\mathbf{ZM}_e$) of 5-methyl-1,3,4-oxathiazol-2-one; VS_i is the matrix row sum.



$\mathbf{ZM}_e =$

Atom	1	2	3	4	5	6	7	VS_i
1	0	6	0	0	6	0	0	12
2	6	0	6	0	0	3	0	15
3	0	6	0	4	0	0	0	10
4	0	0	4	0	6	0	0	10
5	6	0	0	6	0	0	3	15
6	0	3	0	0	0	0	0	3
7	0	0	0	0	3	0	0	3

${}^m\mathbf{ZM}_e =$

Atom	1	2	3	4	5	6	7	VS_i
1	0	0.167	0	0	0.167	0	0	0.334
2	0.167	0	0.167	0	0	0.333	0	0.667
3	0	0.167	0	0.250	0	0	0	0.417
4	0	0	0.250	0	0.167	0	0	0.417
5	0.167	0	0	0.167	0	0	0.333	0.667
6	0	0.333	0	0	0	0	0	0.333
7	0	0	0	0	0.333	0	0	0.333

The calculation was based on the following vertex degrees: $\delta_1=2$, $\delta_2=3$, $\delta_3=2$, $\delta_4=2$, $\delta_5=3$, $\delta_6=1$, and $\delta_7=1$. The second Zagreb index is $M_2=6 \times 4 + 3 \times 2 + 4=34$; the modified second Zagreb index is ${}^mM_2=0.167 \times 4 + 0.333 \times 2 + 0.25=1.584$.

Related to the χ matrix are also the $\rightarrow$ *extended adjacency matrices* based on a sort of average vertex degree and the **edge-XI matrix**, which is derived from $\rightarrow$ *distance-valency matrices Dval* where the simple vertex degree of the χ matrix is replaced with the $\rightarrow$ *valency of a vertex* [Ivanciuc, 1999c]:

$$[\mathbf{XI}_e(w)]_{ij} \equiv [\mathbf{Dval}_e(0, -0.5, -0.5; w)]_{ij} = \begin{cases} 1/\sqrt{\text{val}_i(w) \cdot \text{val}_j(w)} & \text{if } (i, j) \in E(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases}$$

where $\text{val}(w)$ is the valency of a vertex calculated from a weighting scheme w as the sum of the weights of the edges to the first neighbors. Obviously, when no weighting scheme is applied to graph vertices or, which is the same, the unitary scheme is chosen (i.e., $w_i=1$), then the edge-XI matrix coincides with the χ matrix since the valency of a vertex reduces to the vertex degree.

The **edge-weighted Harary matrix**, denoted as ${}^w\mathbf{H}_e$, is a symmetric weighted adjacency matrix defined by weighting each edge between adjacent vertices v_i and v_j as the following [Lučić, Miličević *et al.*, 2002; Janežič, Miličević *et al.*, 2007]:

$$w_{ij} = \sum_m c_{ij}^m \cdot \frac{1}{m^2}$$

where m is the path length and c_{ij}^m is the number of paths of length m , which include the edge between vertices v_i and v_j ; the summation goes up to the maximum path length in the graph. Note that $c_{ij}^1=1$ and the diagonal elements of this matrix are zero by definition. It should also be noted that edge weights are calculated from paths weighted by the reciprocal path length, hence the name *Harary matrix*.

The **edge-Harary index** is defined as the half sum of the elements of the edge-weighted Harary matrix:

$$Wi({}^wH_e) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [{}^wH_e]_{ij}$$

where Wi is the $\rightarrow$ Wiener operator and A the number of graph vertices. It is worth to note that, for acyclic graphs, the edge-Harary index coincides with the Harary index from the reciprocal distance matrix; for cycle-containing graphs, the two indices are usually different.

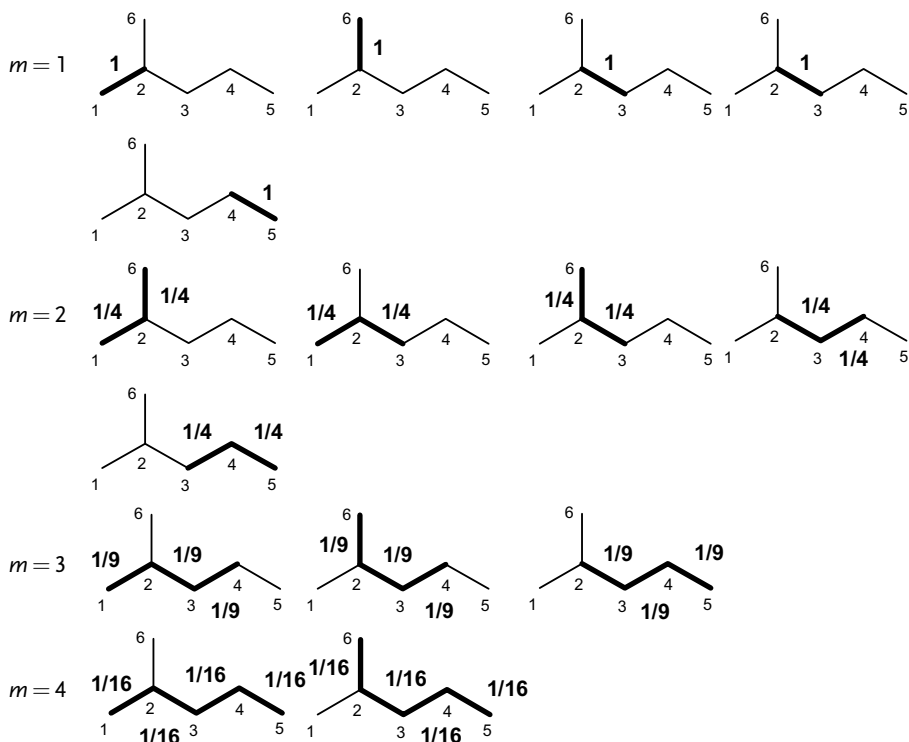
The **modified edge-weighted Harary matrix**, denoted as ${}^{mw}H_e$, is defined as the reciprocal matrix of the edge-weighted Harary matrix as [Lučić, Miličević *et al.*, 2002; Janežič, Miličević *et al.*, 2007]

$$[{}^{mw}H_e]_{ij} = \begin{cases} 1/[{}^wH_e]_{ij} & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

The **modified edge-weighted Harary index** was defined as the half sum of the elements of the modified edge-weighted Harary matrix [Lučić, Miličević *et al.*, 2002].

Example W9

Edge-weighted Harary matrix (${}^wH_e$) for 2-methylpentane; edge weights from paths of different length are shown.



$w_{12} = w_{26} = 1 + 2 \times \frac{1}{4} + \frac{1}{9} + \frac{1}{16} = 1.674$	${}^w\mathbf{H}_e =$	Atom	1	2	3	4	5	6
$w_{23} = 1 + 3 \times \frac{1}{4} + 3 \times \frac{1}{9} + 2 \times \frac{1}{16} = 2.208$		1	0	1.674	0	0	0	0
$w_{34} = 1 + 2 \times \frac{1}{4} + 3 \times \frac{1}{9} + 2 \times \frac{1}{16} = 1.958$		2	1.674	0	2.208	0	0	1.674
$w_{45} = 1 + \frac{1}{4} + \frac{1}{9} + 2 \times \frac{1}{16} = 1.486$		3	0	2.208	0	1.958	0	0
		4	0	0	1.958	0	1.486	0
		5	0	0	0	1.486	0	0
		6	0	1.674	0	0	0	0

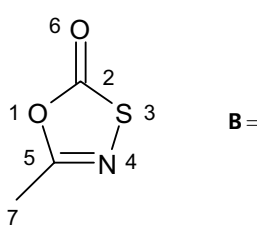
The **Burden matrix** is another interesting weighted adjacency matrix from which $\rightarrow$ *Burden eigenvalues* are computed and used in QSAR/QSPR modeling. This is defined as [Burden, 1989]

$$[\mathbf{B}]_{ij} = \begin{cases} \pi_{ij}^* \cdot 10^{-1} & \text{if } (i,j) \in E(G) \\ Z_i & \text{if } i = j \\ 0.001 & \text{if } (i,j) \notin E(G) \end{cases}$$

where the diagonal elements are the atomic numbers Z_i of atoms; the off-diagonal elements representing two bonded atoms v_i and v_j are function of the $\rightarrow$ *conventional bond order* π^* , that is, 0.1, 0.2, 0.3, and 0.15 for a single, double, triple, and aromatic bond, respectively; all other matrix elements are set at 0.001. Moreover, matrix elements corresponding to terminal bonds are augmented by 0.01.

Example W10

Burden matrix of 5-methyl-1,3,4-oxathiazol-2-one; VS_i is the matrix row sum.

	$\mathbf{B} =$	Atom	1	2	3	4	5	6	7	VS_i
		1	8	0.1	0.001	0.001	0.1	0.001	0.001	8.204
		2	0.1	6	0.1	0.001	0.001	0.21	0.001	6.413
		3	0.001	0.1	16	0.1	0.001	0.001	0.001	16.204
		4	0.001	0.001	0.1	7	0.2	0.001	0.001	7.304
		5	0.1	0.001	0.001	0.2	6	0.001	0.11	6.413
		6	0.001	0.21	0.001	0.001	0.001	8	0.001	8.215
		7	0.001	0.001	0.001	0.001	0.11	0.001	6	6.115

A generalization of the Burden matrix was proposed in $\rightarrow$ *DRAGON descriptors* where instead of the atomic numbers Z , atomic masses (m), van der Waals volumes (v), Sanderson electronegativities (e), and polarizabilities (p) are used as the weighting schemes for graph vertices. Thus, a general definition of the Burden matrix in terms of a vertex weighting scheme w_i is

$$[\mathbf{B}]_{ij} = \begin{cases} \pi_{ij}^* \cdot 10^{-1} & \text{if } (i,j) \in E(G) \\ w_i & \text{if } i = j \\ 0.001 & \text{if } (i,j) \notin E(E) \end{cases}$$

This kind of generalization was also proposed by Ivanciuc [Ivanciuc, 2001f], who suggested to use in the main diagonal $\rightarrow$ *local vertex invariants* instead of the atomic numbers Z . Moreover, an extension of the Burden matrix was proposed to derive $\rightarrow$ *BCUT descriptors*.

The **weighted electronic connectivity matrix** (or **CEP matrix**) of an edge-weighted molecular graph is a symmetric weighted adjacency matrix, of size $A \times A$, whose elements are defined as [Berinde and Berinde, 2004]

$$[\text{CEP}]_{ij} = \begin{cases} \text{wed}_{ij} & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases} \quad \text{wed}_{ij} = \frac{1}{\pi_{ij}^*} \cdot \frac{z'_i + z'_j}{\delta_i \cdot \delta_j}$$

where wed_{ij} is the weighting scheme for graph edges, called **weighted electronic distance** and defined in terms of the $\rightarrow$ conventional bond order π^* and the $\rightarrow$ vertex degrees δ_i and δ_j of two adjacent vertices. The term z' is a local vertex invariant called **formal degree** of a vertex and defined as

$$z'_i = Z_i \cdot \delta_i$$

where Z_i is the atomic number associated with the i th vertex and δ_i is its vertex degree. The row sum of the weighted electronic connectivity matrix is a local vertex invariant, denoted as SEP_i , defined as

$$\text{SEP}_i \equiv \text{VS}_i[\text{CEP}]_{ij} = \sum_{j=1}^A [\text{CEP}]_{ij}$$

where VS_i is the $\rightarrow$ row sum operator.

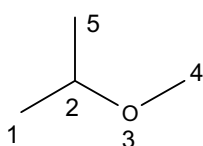
CEP matrix allows to differentiate multiple bonds, heteroatoms, and connectivities. Three molecular descriptors were derived from the CEP matrix:

$$\text{ZEP} = \sum_{i=1}^A \text{SEP}_i^{1/2} \quad \text{RZ} = \sum_{u,v} (\text{wed}_u \cdot \text{wed}_v)^{-1/2} \quad V_\gamma = \sum_{i=1}^A \frac{\text{SEP}_i}{\delta_i}$$

where u and v are two adjacent edges; the summation in the RZ index goes over all the pairs of adjacent edges.

Example W11

CEP matrix and related molecular descriptors for methyl-isopropyl-ether.

	CEP =						SEP _i

$$\text{wed}_{12} = \frac{1}{1} \times \frac{6 \times 1 + 6 \times 3}{1 \times 3} = 8 \quad \text{wed}_{23} = \frac{1}{1} \times \frac{6 \times 3 + 8 \times 2}{2 \times 3} = 5.67$$

$$\text{wed}_{25} = \frac{1}{1} \times \frac{6 \times 1 + 6 \times 3}{1 \times 3} = 8 \quad \text{wed}_{34} = \frac{1}{1} \times \frac{8 \times 2 + 6 \times 1}{2 \times 1} = 11$$

$$\text{ZEP} = \sqrt{8} + \sqrt{21.67} + \sqrt{16.67} + \sqrt{11} + \sqrt{8} = 17.71$$

$$\text{RZ} = \frac{1}{\sqrt{8 \times 5.67}} + \frac{1}{\sqrt{8 \times 8}} + \frac{1}{\sqrt{5.67 \times 8}} + \frac{1}{\sqrt{5.67 \times 11}} = 0.549$$

$$V_\gamma = \frac{8}{1} + \frac{21.67}{3} + \frac{16.67}{2} + \frac{11}{1} + \frac{8}{1} = 42.56$$

Another class of weighted vertex adjacency matrices consists of unsymmetric $A \times A$ matrices representing heteroatoms by some values different from 1 outside the matrix diagonal. The diagonal elements of these matrices are all equal to zero as in the adjacency matrix or to an

atomic property w_i , and each element i - j corresponding to a pair of bonded atoms is a vertex weight w_j associated with the bonded atom v_j :

$$[{}^w\mathbf{A}]_{ij} = \begin{cases} w_j & \text{if } (i,j) \in E(G) \\ w_i \text{ or } 0 & \text{if } i = j \\ 0 & \text{if } (i,j) \notin E(G) \end{cases}$$

The vertex weight w_j can be any physico-chemical atomic property or a local vertex invariant of the j th atom bonded to the i th atom. It is worth to note that, unlike the other weighted adjacency matrices, whose off-diagonal entries corresponding to edges are some edge weights, these adjacency matrices have vertex weights instead of edge weights.

To highlight heteroatom characteristics, relative physico-chemical properties with respect to carbon atom are usually used. As a consequence, if no heteroatom is present, the adjacency matrix is obtained independently of the weighting scheme. These matrices are unsymmetrical because for each i th atom the physico-chemical property of the j th bonded atom is the i - j matrix element.

The **electronegativity-weighted adjacency matrix** ${}^{\chi}\mathbf{A}$ is an example of these unsymmetrical weighted adjacency matrices and was proposed to calculate $\rightarrow$ *MARCH-INSIDE descriptors*. It is defined as [González Díaz, Olazabal *et al.*, 2002]:

$$[{}^{\chi}\mathbf{A}]_{ij} = \begin{cases} \chi_j & \text{if } (i,j) \in E(G) \\ \chi_i & \text{if } i = j \\ 0 & \text{if } (i,j) \notin E(G) \end{cases}$$

where χ denotes atomic electronegativities.

The **atomic weight-weighted adjacency matrix** (or **chemical adjacency matrix**) is another example of unsymmetrical weighted adjacency matrices defined as [Bajaj, Sami *et al.*, 2005]:

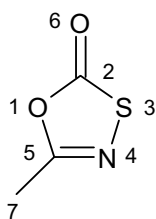
$$[{}^m\mathbf{A}]_{ij} = \begin{cases} \frac{m_j}{m_C} & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

where m_j is the atomic mass associated with the j th vertex bonded to v_i and m_C the atomic mass of the carbon atom.

Chemical adjacency matrices based on relative atomic masses were used to calculate the $\rightarrow$ *atomic molecular connectivity index*, $\rightarrow$ *Zagreb topochemical indices*, and the $\rightarrow$ *superadjacency topochemical index*, all defined in terms of the $\rightarrow$ *Madan chemical degree* δ^c , which is the row sum of the atomic weight-weighted adjacency matrix.

Example W12

Atomic weight-weighted adjacency matrix of 5-methyl-1,3,4-oxathiazol-2-one. δ^c is the Madan chemical degree and CS_j refers to the matrix column sums.



Atom	1	2	3	4	5	6	7	δ^c
1	0	1	0	0	1	0	0	2
2	1.332	0	2.670	0	0	1.332	0	5.334
3	0	1	0	1.166	0	0	0	2.166
4	0	0	2.670	0	1	0	0	3.670
5	1.332	0	0	1.166	0	0	1	3.498
6	0	1	0	0	0	0	0	1
7	0	0	0	0	1	0	0	1
CS_j	2.664	3	5.340	2.332	3	1.332	1	18.668

1.332, 1.166, and 2.670 are the relative atomic masses of oxygen, nitrogen, and sulfur, respectively. Note that column sums CS_j of the atomic weight-weighted adjacency matrix are

$$CS_j = \frac{m_j}{m_C} \cdot \delta_j$$

By extension of the $\rightarrow$ *additive adjacency matrix* to account for heteroatoms, the **additive chemical adjacency matrix** was derived from the vertex adjacency matrix substituting row elements equal to 1, corresponding to bonded atoms, with the $\rightarrow$ *Madan chemical degree* of the bonded atom as [Bajaj, Sambi *et al.*, 2004b]:

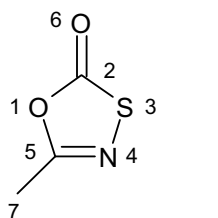
$$[{}^cA]_{ij} = \begin{cases} \delta_j^c & \text{if } (i,j) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

This matrix is defined according to the scheme of the chemical adjacency matrix by using the Madan chemical degree as the weighting scheme for graph vertices.

The row sum of this matrix is the **chemical extended connectivity** of first-order EC^{1c} , which is a modification of the $\rightarrow$ *extended connectivity* defined by Morgan to account for heteroatoms. This local invariant was used to calculate the $\rightarrow$ *superadjacency topochemical index*.

Example W13

The additive chemical adjacency matrix of 5-methyl-1,3,4-oxathiazol-2-one; EC^{1c} is the chemical extended connectivity and CS_j refers to the matrix column sums.

	Atom	1	2	3	4	5	6	7	EC^{1c}
	1	0	5.334	0	0	3.498	0	0	8.832
	2	2	0	2.166	0	0	1	0	5.166
	3	0	5.334	0	3.670	0	0	0	9.004
	4	0	0	2.166	0	3.498	0	0	5.664
	5	2	0	0	3.670	0	0	1	6.670
	6	0	5.334	0	0	0	0	0	5.334
	7	0	0	0	0	3.498	0	0	3.498
${}^cA =$	CS_j	4	16.002	4.332	7.340	10.494	1	1	44.168

The matrix elements different from zero are the Madan vertex degrees taken from Example W12. In the column vector, row sums of the matrix are collected, which correspond to the chemical extended connectivities of the molecule atoms.

Another unsymmetrical weighted adjacency matrix was defined so as to account for the number of lone-pairs and π bonds on the main diagonal and for the number of hydrogens bonded to a neighboring atom out of the main diagonal [Yang and Zhong, 2003]:

$$[{}^+A]_{ij} = \begin{cases} 1 + h_j/6 & \text{if } (i,j) \in E(G) \\ n_i^{LP} + n_i^\pi & \text{if } i = j \\ 0 & \text{if } (i,j) \notin E(G) \end{cases}$$

where n_i^{LP} and n_i^π are the number of lone-pair electrons and the number of π bonds of the i th atom, respectively, and h_j is the number of hydrogen atoms bonded to the j th atom, which in turn is bonded to the i th atom. From this matrix, modified $\rightarrow$ *vertex degrees* and $\rightarrow$ *connectivity indices* were derived.

Other important weighted vertex adjacency matrices are the $\rightarrow$ *extended adjacency matrices*, the $\rightarrow$ *edge-Wiener matrix*, $\rightarrow$ *edge-Cluj matrices*, $\rightarrow$ *edge-Szeged matrices*, and the $\rightarrow$ *random walk Markov matrix*.

Analogous to weighted vertex adjacency matrices, weighted edge adjacency matrices are derived from weighted graphs assigning each edge a weight that can be a bond order or a bond distance. Examples are the $\rightarrow$ *bond order-weighted edge adjacency matrix* and the $\rightarrow$ *bond distance-weighted edge adjacency matrix*. Moreover, the χ^E **matrix** [Ivanciuc, Ivanciuc *et al.*, 1997] or **edge-connectivity matrix** [Janežič, Miličević *et al.*, 2007] of the graph G is the $\rightarrow$ χ *matrix* of the $\rightarrow$ *line graph* of G , that is,

$$[\chi^E]_{ij} = \begin{cases} 1/\sqrt{\epsilon_i \cdot \epsilon_j} & \text{if } (i,j) \in C(G) \\ 0 & \text{otherwise} \end{cases}$$

where ϵ is the $\rightarrow$ *edge degree* and C is the set of connections in the graph, that is, the set of pairs of adjacent edges. Also $\rightarrow$ *Zagreb matrices* were defined in terms of edge degrees used in place of the vertex degrees [Janežič, Miličević *et al.*, 2007].

• weighted distance matrices

These are a generalization of the distance matrix for heteroatom molecular systems with possible multiple bonds. Weighted distance matrices are calculated from vertex- and/or edge-weighted molecular graphs. Depending on the weighting scheme w , several weighted distance matrices were proposed for QSAR/QSPR modeling of organic compounds.

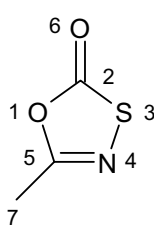
To only account for heteroatoms in the molecule, the **augmented distance matrix** $^aD(w)$ was derived from vertex-weighted graphs replacing the zero diagonal entries of the usual distance matrix with values characterizing different atoms in the molecule [Randić and Pompe, 2001b; Lučić, Miličević *et al.*, 2003]:

$$[^aD(w)]_{ij} = \begin{cases} d_{ij} & \text{if } i \neq j \\ w_i & \text{if } i = j \end{cases}$$

where d_{ij} is the $\rightarrow$ *topological distance* between vertices v_i and v_j and w_i is a weight for the i th atom. The diagonal elements usually are some atomic physico-chemical properties or local vertex invariants; however, diagonal elements can also be atom-type variable parameters ($x, y, z, \dots$) that are optimized to enhance the estimate of the studied property by regression analysis.

Example W14

Augmented distance matrix of 5-methyl-1,3,4-oxathiazol-2-one derived by using relative atomic numbers (i.e., Z_i/Z_C) as the vertex weights; VS_i is the matrix row sum.

	Atom	1	2	3	4	5	6	7	VS_i
	1	1.33	1	2	2	1	2	2	11.33
	2	1	1	1	2	2	1	3	11.00
	3	2	1	2.67	1	2	2	3	13.67
	4	2	2	1	1.17	1	3	2	12.17
	5	1	2	2	1	1	3	1	11.00
	6	2	1	2	3	3	1.33	4	16.33
	7	2	3	3	2	1	4	1	16.00

The calculation was based on the following relative atomic numbers: $Z_1 = 1.33$, $Z_2 = 1$, $Z_3 = 2.67$, $Z_4 = 1.17$, $Z_5 = 1$, $Z_6 = 1.33$, and $Z_7 = 1$.

From the augmented distance matrix, a number of topological indices, called $\rightarrow$ *variable descriptors*, are derived.

For vertex- and edge-weighted molecular graphs, the weighted distance matrix is generally defined as a square symmetric $A \times A$ matrix, A being the number of graph vertices, as

$$[\mathbf{D}(w)] = \begin{cases} d_{ij}(w) & \text{if } i \neq j \\ w_i & \text{if } i = j \end{cases}$$

where $d_{ij}(w)$ is the distance between vertices v_i and v_j computed as a function of the weights assigned to the edges involved in the path connecting v_i and v_j ; w_i is the weight of the vertex v_i , that is, some atomic physico-chemical property or local vertex invariant. The edge weight can be a physico-chemical bond property such as bond order, bond distance, resonance integral, and so on, or some combination of the properties of the two vertices incident to the bond. The weighted distance $d_{ij}(w)$ between vertices v_i and v_j can be calculated in different ways; the most common one is to sum up the weights of all the edges along the shortest path between v_i and v_j :

$$d_{ij}(w) = \sum_{b=1}^{d_{ij}} w_b$$

where the summation goes over all the bonds involved in the path, d_{ij} is the number of bonds along the shortest path, and w_b is the bond weight.

This approach to the calculation of weighted distances naturally derives from the definition of topological distance between two vertices, which is the sum of the bonds along the shortest path connecting the vertices considered, each bond contributing one to the sum. When more than one shortest path exists between a pair of vertices, some rules must be adopted to calculate the corresponding matrix element. For example, average values from all the shortest paths can be calculated or the minimum value can be chosen. Another way to obtain weighted distance matrices is to take, as matrix entries, the minimum sum of the edge weights w_b along the path p_{ij} between the vertices considered, which is not necessarily the shortest possible path between them:

$$d_{ij}(w) = \min_{p_{ij}} (w_{ij})$$

where w_{ij} is the path weight calculated as

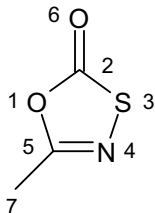
$$w_{ij} = \sum_{b=1}^{|p_{ij}|} w_b$$

where $|p_{ij}|$ is the number of bonds along the path p_{ij} .

The diagonal entries of these weighted matrices can be chosen different from zero to encode information about the chemical nature of the molecule atoms.

Example W15

Calculation of weighted distances in the H-depleted molecular graph of 5-methyl-1,3,4-oxathiazol-2-one.



The edge weights computed according to the Barysz $\rightarrow$ *Z weighting scheme* ($w(Z)$, Table W8) are the following: $w_{12} = 0.750$, $w_{23} = 0.375$, $w_{34} = 0.321$, $w_{45} = 0.429$, $w_{15} = 0.750$, $w_{26} = 0.375$, and $w_{57} = 1.000$. There exist two paths between vertices v_2 and v_5 , namely, $p_{25} = \{v_2, v_1, v_5\}$, which is the shortest path, and $p_{25} = \{v_2, v_3, v_4, v_5\}$. The weight associated with the path $p_{25} = \{v_2, v_1, v_5\}$ is $w_{25} = w_{12} + w_{15} = 0.750 + 0.750 = 1.500$, while the weight associated with the path $p_{25} = \{v_2, v_3, v_4, v_5\}$ is $w_{25} = w_{23} + w_{34} + w_{45} = 0.375 + 0.321 + 0.429 = 1.125$. Therefore, the weighted distance $d_{25}(Z)$ can be 1.500 if only the shortest path is considered in the distance computation; otherwise, the weighted distance is 1.125, because this is the minimum weight between the two paths joining the vertices.

The **multigraph distance matrix** ${}^*\mathbf{D}$ is a weighted distance matrix derived from an edge-weighted molecular graph by using the $\rightarrow$ *conventional bond order* as the weighting scheme for edges. The distance from vertex v_i to vertex v_j is obtained by counting the edges in the shortest path between them, where each edge counts as the reciprocal of the conventional bond order π^* , that is, the $\rightarrow$ *relative topological distance*, and therefore contributes $1/\pi^*$ to the overall path length [Balaban, 1982; Balaban, 1983a]. The multigraph distance matrix is defined as

$$[{}^*\mathbf{D}]_{ij} = \begin{cases} d_{ij}(\pi^*) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where

$$d_{ij}(\pi^*) = \min_{p_{ij}}(w_{ij}) \quad w_{ij} = \sum_{b=1}^{d_{ij}} (\pi_b^*)^{-1}$$

where b runs over all the bonds involved in the shortest path p_{ij} and d_{ij} is the number of bonds along the path p_{ij} .

The sum of the matrix elements in each row is called **multigraph distance degree** ${}^*\sigma$ and is defined as

$${}^*\sigma_i \equiv VS_i({}^*\mathbf{D}) = \sum_{j=1}^A [{}^*\mathbf{D}]_{ij}$$

where the symbol VS stands for the $\rightarrow$ vertex sum operator and A for the number of graph vertices. This is a local vertex invariant taking edge multiplicity into account. $\rightarrow$ Balaban modified distance connectivity indices are calculated from the multigraph distance degrees used in place of the common distance degrees σ derived from the simple graph.

Moreover, the **bond order-weighted Wiener index** was proposed by applying the $\rightarrow$ Wiener operator Wi to a variant of the multigraph distance matrix $\mathbf{D}(\pi)$, where the conventional bond orders are replaced by the Mulliken bond orders [Jalbout and Li, 2003c]. Then, the bond order-weighted Wiener index was defined as

$$Wi(\mathbf{D}(\pi)) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij}(\pi)$$

where $d_{ij}(\pi)$ is the bond order-weighted distance between vertices v_i and v_j , and A is the total number of graph vertices.

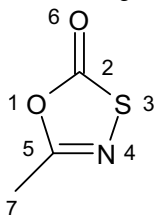
The **chemical distance matrix** is a variant of the multigraph distance matrix defined by using the $\rightarrow$ chemical distance as the weighting scheme for edges; therefore, the path weight w_{ij} in terms of bond chemical distances is calculated as [Balaban, Bonchev *et al.*, 1993]:

$$w_{ij} = \sum_{b=1}^{d_{ij}} (\pi_b^*)^{-1/4}$$

where the summation goes over all the bonds along the shortest path p_{ij} ; d_{ij} is the length of the path p_{ij} between vertices v_i and v_j , and π^* the $\rightarrow$ conventional bond order.

Example W16

Multigraph distance matrix $^*\mathbf{D}$ of 5-methyl-1,3,4-oxathiazol-2-one; $^*\sigma_i$ is the multigraph distance degree.



$^*\mathbf{D} =$

Atom	1	2	3	4	5	6	7	$^*\sigma_i$
1	0	1	2	1.5	1	1.5	2	10
2	1	0	1	2	2	0.5	3	9.5
3	2	1	0	1	1.5	1.5	2.5	9.5
4	1.5	2	1	0	0.5	2.5	1.5	9
5	1	2	1.5	0.5	0	2.5	1	8.5
6	1.5	0.5	1.5	2.5	2.5	0	3.5	12
7	2	3	2.5	1.5	1	3.5	0	13.5

The **bond length-weighted distance matrix** is defined in a similar way as the multigraph distance matrix but using as the weighting scheme for graph edges the bond lengths r or the relative bond lengths r^* in place of the conventional bond orders [Castro, Gutman *et al.*, 2002; Lu, Guo *et al.*, 2006b, 2006c]:

$$[\mathbf{D}(r)]_{ij} = \begin{cases} d_{ij}(r) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where

$$d_{ij}(r) = \sum_{b=1}^{d_{ij}} r_b$$

where b runs over all the bonds involved in the shortest path p_{ij} , d_{ij} is the number of bonds along the shortest path p_{ij} , and r is the bond length. This distance matrix encodes information on the molecular geometry, but not on the molecular conformations, as the $\rightarrow$ *geometry matrix* does.

The **bond length-weighted Wiener index** is calculated from this matrix by applying the $\rightarrow$ *Wiener operator* Wi as [Castro, Gutman *et al.*, 2002]:

$$Wb \equiv Wi(\mathbf{D}(r)) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}(r)]_{ij}$$

This index can be calculated both from H-depleted and H-filled molecular graphs.

From the weighted distance matrix based on relative bond lengths r^* , the $\rightarrow$ *Lu index*, the $\rightarrow$ *DAI indices*, and the $\rightarrow$ *Nt index* are calculated. The relative bond length r^* is calculated as the ratio between each bond length r and the bond length of C–C bond (1.54 Å).

For vertex-weighted graphs, where no specific weight is assigned to edges, one can derive an edge weight w_{ij} by combining the weights w_i and w_j of the two incident vertices and, then, compute the aforementioned weighted distance from edge weights. Alternatively, a weight w_{ij} can be calculated for the path p_{ij} connecting vertices v_i and v_j by some function of the weights w of all the vertices along the considered path p_{ij} as

$$w_{ij} = f(w_i, w_k, \dots, w_j)$$

and then the weighted distance between vertices v_i and v_j is usually taken as the minimum path weight over all the paths connecting vertices v_i and v_j :

$$d_{ij}(w) = \min_{p_{ij}}(w_{ij})$$

The weighted distance $d_{ij}(w)$ usually corresponds to the weight of the shortest path between vertices v_i and v_j .

The **path- χ matrix** [Hall, 1990] is an example of weighted distance matrix derived from a vertex-weighted graph. It is based on the path contributions arising in constructing $\rightarrow$ *connectivity indices*. Using the $\rightarrow$ *vertex degree* δ of the atoms as the weighting scheme, each path p_{ij} between the vertices v_i and v_j is weighted by the **path connectivity** defined as

$$w_{ij} = (\delta_i \cdot \delta_k \cdot \dots \cdot \delta_j)^{-1/2}$$

where $\delta_i, \delta_k, \dots, \delta_j$ are the vertex degrees of the atoms along the path p_{ij} .

Using the $\rightarrow$ *valence vertex degree* δ^v or the $\rightarrow$ *bond vertex degree* δ^b instead of the simple vertex degree, the corresponding path weights are defined as

$$w_{ij} = (\delta_i^v \cdot \delta_k^v \cdot \dots \cdot \delta_j^v)^{-1/2} \quad w_{ij} = (\delta_i^b \cdot \delta_k^b \cdot \dots \cdot \delta_j^b)^{-1/2}$$

Therefore, based on path connectivity, the path- χ matrix is defined as

$$[\chi_p]_{ij} = \begin{cases} d_{ij}(\delta) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $d_{ij}(\delta)$ is the path connectivity of the shortest path between vertices v_i and v_j . Diagonal entries are zero assuming that there is no path bonding to the atom itself. This matrix is an extension of the $\rightarrow$ *χ matrix*, which only accounts for pairs of adjacent vertices.

For acyclic graphs, $\rightarrow$ connectivity indices for path subgraphs can be calculated by the Wiener operator applied to the product of the path- χ matrix and the $\rightarrow$ geodesic matrix ${}^k\mathbf{B}$ whose elements are all equal to zero except for those corresponding to the shortest paths $i-j$ of length k that are equal to 1:

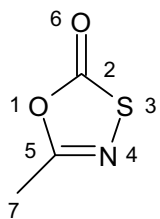
$${}^k\chi_p \equiv Wi(\chi_p \otimes {}^k\mathbf{B}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\chi_p \otimes {}^k\mathbf{B}]_{ij}$$

where Wi is the $\rightarrow$ Wiener operator and the symbol $\otimes$ indicates the $\rightarrow$ Hadamard matrix product. Other graph invariants can be derived from the path- χ matrix such as the largest positive eigenvalue, the spectrum, and row sums for any graphs. For example, the $\rightarrow$ characteristic root index is the sum of the positive eigenvalues of the path- χ matrix.

A variant of the path- χ matrix is obtained by using the vertex distance sum σ in place of the vertex degree as the vertex-weighting scheme. The matrix obtained in this way can be considered an extension of the $\rightarrow$ distance-sum-connectivity matrix and thus called **path-distance-sum-connectivity matrix**, denoted by ${}^\sigma\chi_p$.

Example W17

Path- χ matrix χ_p and path-distance sum-connectivity matrix ${}^\sigma\chi_p$ of 5-methyl-1,3,4-oxathiazol-2-one; VS_i is the matrix row sum.

	$\chi_p =$	Atom	1	2	3	4	5	6	7	VS_i
		1	0	0.408	0.289	0.289	0.408	0.408	0.408	2.210
		2	0.408	0	0.408	0.289	0.236	0.577	0.236	2.154
		3	0.289	0.408	0	0.500	0.289	0.408	0.289	2.183
		4	0.289	0.289	0.500	0	0.408	0.289	0.408	2.183
		5	0.408	0.236	0.289	0.408	0	0.236	0.577	2.154
		6	0.408	0.577	0.408	0.289	0.236	0	0.236	2.154
		7	0.408	0.236	0.289	0.408	0.577	0.236	0	2.154
	${}^\sigma\chi_p =$	Atom	1	2	3	4	5	6	7	VS_i
		1	0	0.100	0.030	0.030	0.100	0.026	0.026	0.312
		2	0.100	0	0.095	0.029	0.032	0.082	0.008	0.346
		3	0.030	0.095	0	0.091	0.029	0.078	0.007	0.277
		4	0.030	0.029	0.091	0	0.095	0.007	0.025	0.277
		5	0.100	0.032	0.029	0.095	0	0.008	0.082	0.346
		6	0.026	0.082	0.078	0.007	0.008	0	0.007	0.155
		7	0.026	0.008	0.007	0.025	0.082	0.007	0	0.155

The calculation of the two matrices was based on the following vertex degrees: $\delta_1 = 2$, $\delta_2 = 3$, $\delta_3 = 2$, $\delta_4 = 2$, $\delta_5 = 3$, $\delta_6 = 1$, and $\delta_7 = 1$ and vertex distance degrees $\sigma_1 = \sigma_2 = \sigma_5 = 10$, $\sigma_3 = \sigma_4 = 11$, and $\sigma_6 = \sigma_7 = 15$.

Moreover, still based on the product of vertex degrees of all the vertices along the path, the **topological state matrix**, denoted as **T**, is a square symmetric matrix of dimension $A \times A$, A being the number of graph vertices, defined as [Hall and Kier, 1990; Hu and Xu, 1994]

$$[\mathbf{T}]_{ij} = \begin{cases} \sum_{p_{ij}} GM_{ij}^b \cdot f(n_{ij})^c & \text{if } i \neq j \\ \delta_i & \text{if } i = j \end{cases}$$

where the summation goes over all the paths p_{ij} between vertices v_i and v_j , and the exponents can be $b = \pm 1$ and $c = 0, \pm 1, \pm 2, \pm 3$; GM_{ij} is the geometric mean of the $\rightarrow$ vertex degree δ of the vertices involved in the path i - j and defined as

$$GM_{ij} = \left(\prod_{a=1}^{n_{ij}} \delta_a \right)^{1/n_{ij}} \quad a \in p_{ij}$$

where n_{ij} is the number of vertices in the path p_{ij} , $f(n_{ij})$ is a function of the separation between the vertices v_i and v_j . The simplest function (as proposed by Hall-Kier) is the reciprocal of the number of vertices in the path:

$$f(n_{ij}) = \frac{1}{n_{ij}} = \frac{1}{m+1}$$

where m is the path length; note that, for acyclic graphs, the function reduces to $1/(d_{ij} + 1)$, d_{ij} being the $\rightarrow$ topological distance between the two vertices. The simplest topological state matrix is defined by $b = 1$ and $c = 1$; moreover, the reciprocal of GM terms ($b = -1$) was considered in defining the topological state matrix **T**.

The diagonal entries of **T** correspond to paths of length zero, thus they are simply equal to the vertex degrees. Moreover, for cyclic graphs, there can be more than one path between the vertices v_i and v_j , hence each topological state matrix element is a sum of the contributions of all paths between the vertices considered.

From the topological state matrix, a local vertex invariant S_i , called **topological state** (or **vertex topological state**), can be calculated as

$$S_i \equiv VS_i(\mathbf{T}) = \sum_{j=1}^A [\mathbf{T}]_{ij}$$

where VS stands for the $\rightarrow$ row sum operator.

A **total topological state index** τ having a highly discriminating power was also derived as the following:

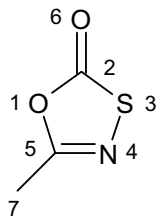
$$\tau = \sum_{i=1}^A [\mathbf{T}]_{ii} + \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{T}]_{ij}$$

As the diagonal entries are just the vertex degrees δ , the first term on the right is the $\rightarrow$ total adjacency index A_v . The topological state index is one among the $\rightarrow$ molecular ID numbers.

Valence topological state matrix $\mathbf{T}^v$, valence topological state S^v and total valence topological state index τ^v can be obtained using $\rightarrow$ valence vertex degree δ^v instead of the simple vertex degrees to encode the atom identity. Valence topological states were used to search for topological equivalence among molecule atoms to define the $\rightarrow$ Kier symmetry index.

Example W18

Topological state matrix **T** of 5-methyl-1,3,4-oxathiazol-2-one. VS_i is the matrix row sum. The parameters for matrix calculation are $b = -1$ and $c = 1$.

	Atom	1	2	3	4	5	6	7	VS_i
	1	2	0.204	0.146	0.146	0.204	0.183	0.183	3.066
	2	0.204	3	0.204	0.146	0.127	0.289	0.121	4.091
	3	0.146	0.204	2	0.250	0.146	0.183	0.134	3.063
	4	0.146	0.146	0.250	2	0.204	0.134	0.183	3.063
	5	0.204	0.127	0.146	0.204	3	0.121	0.289	4.091
	6	0.183	0.289	0.183	0.134	0.121	1	0.160	2.070
	7	0.183	0.121	0.134	0.183	0.289	0.160	1	2.070

The calculation of the **T** matrix was based on the following vertex degrees: $\delta_1 = 2$, $\delta_2 = 3$, $\delta_3 = 2$, $\delta_4 = 2$, $\delta_5 = 3$, $\delta_6 = 1$, and $\delta_7 = 1$.

The total topological state index is $\tau = 14 + 3.758 = 17.758$.

The **atomic weight-weighted distance matrix** ${}^c\mathbf{D}$ is an unsymmetrical weighted distance matrix derived from a vertex-weighted graph to account for heteroatoms [Bajaj, Sambi *et al.*, 2004a, 2004b]. This was proposed according to a scheme similar to that used for the path- χ matrix defined above, where the product of the weights of path vertices is replaced by their sum and relative atomic weights are used as the weighting scheme instead of vertex degrees. Thus, the path weight w_{ij} is calculated as

$$w_{ij} = \left(\frac{m_k}{m_C} + \dots + \frac{m_j}{m_C} \right) = \left(\frac{m_k + \dots + m_j}{m_C} \right)$$

where $m_k, \dots, m_j$ are the atomic masses of the vertices along the path p_{ij} , m_C refers to the atomic mass of the carbon atom, and the subscript k is the index of the first atom bonded to the i th atom along the path considered. It is worth to note that in the calculation of this path weight, the vertex v_i from which the path starts is not accounted for. The atomic weight-weighted distance matrix is then defined as

$$[{}^c\mathbf{D}]_{ij} = \begin{cases} d_{ij}(m) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $d_{ij}(m)$ is the minimum path weight over all the shortest paths between vertices v_i and v_j :

$$d_{ij}(m) = \min_{p_{ij}}(w_{ij})$$

The row sum of this matrix is a local vertex invariant called **chemical distance degree**; the maximum entry in each row is called **chemical atom eccentricity**, which is the maximum distance weighted by relative atomic weights from each atom. The chemical atom eccentricity is used in the calculation of the $\rightarrow$ *superadjacency topochemical index*, while the chemical distance degree in the calculation of the *Wiener topochemical index*.

The **Wiener topochemical index** was proposed as a modification of the $\rightarrow$ *Wiener index* to account for heteroatoms. It is calculated as the half sum of the elements of the atomic weight-

weighted distance matrix [Bajaj, Sambi *et al.*, 2004a, 2004b]:

$$W_C \equiv W_i({}^c\mathbf{D}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [{}^c\mathbf{D}]_{ij}$$

where $[{}^c\mathbf{D}]_{ij}$ denotes the elements of the atomic weight-weighted distance matrix and W_i indicates the $\rightarrow$ *Wiener operator*.

Note. The atomic weight-weighted distance matrix was originally called *chemical distance matrix*; however, this name cannot be used here because it was previously attributed by Balaban [Balaban, Bonchev *et al.*, 1993] to another weighted distance matrix (see below).

Example W19

Atomic weight-weighted distance matrix ${}^c\mathbf{D}$ of 5-methyl-1,3,4-oxathiazol-2-one; VS_i indicates the matrix row sum that is the chemical distance degree; CS_j is the matrix column sum; and η_i^c is the chemical atom eccentricity.

Atom	1	2	3	4	5	6	7	VS_i	η_i^c
1	0	1	3.670	2.166	1	2.332	2	12.168	3.670
2	1.332	0	2.670	3.836	2.332	1.332	3.332	14.834	3.836
3	2.332	1	0	1.166	3.670	2.332	4.670	15.170	4.670
4	2.332	3.670	2.670	0	1	5.002	2	16.674	5.002
5	1.332	2.332	3.836	1.166	0	3.664	1	13.330	3.836
6	2.332	1	3.670	4.836	3.332	0	4.332	19.502	4.836
7	2.332	3.332	4.836	2.166	1	4.664	0	18.330	4.836
CS_j	11.992	12.334	21.352	15.336	12.334	19.326	17.334	110.008	

The calculation was based on the following relative atomic weights for carbon, oxygen, nitrogen, and sulfur, respectively: 1, 1.332, 1.166, and 2.670. The Wiener topochemical index is $W_C = 1/2 \times (12.168 + 14.834 + 15.170 + 16.674 + 13.330 + 19.502 + 18.330) = 55.004$.

Derived from vertex-weighted molecular graphs, the **augmented vertex degree matrix** ${}^a\mathbf{D}(\delta)$ is an unsymmetrical weighted distance $A \times A$ matrix based on the concept of $\rightarrow$ *augmented valence* and defined as [Randić, 2001b; Janežič, Miličević *et al.*, 2007]:

$$[{}^a\mathbf{D}(\delta)]_{ij} = \begin{cases} \frac{\delta_j}{2^{d_{ij}}} & \text{if } i \neq j \\ \delta_i & \text{if } i = j \end{cases}$$

where δ is the $\rightarrow$ *vertex degree* and d_{ij} the $\rightarrow$ *topological distance* between vertices v_i and v_j . The row sum of the augmented vertex degree matrix coincides with the $\rightarrow$ *augmented valence* of a vertex AV_i :

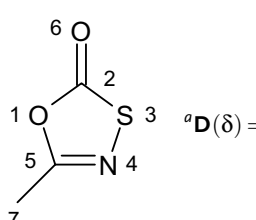
$$VS_i({}^a\mathbf{D}(\delta)) \equiv AV_i = \sum_{j=1}^A \frac{\delta_j}{2^{d_{ij}}}$$

where VS_i is the $\rightarrow$ *row sum operator*.

Moreover, from the row sums of the augmented vertex degree matrix, the $\rightarrow$ *Randić–Plavsic complexity index* is derived.

Example W20

Augmented vertex degree matrix ${}^a\mathbf{D}(\delta)$ of 5-methyl-1,3,4-oxathiazol-2-one; AV_i is the augmented valence of the vertex.

	Atom	1	2	3	4	5	6	7	AV_i
	1	2	1.500	0.500	0.500	1.500	0.250	0.250	6.500
	2	1	3	1	0.500	0.750	0.500	0.125	6.875
	3	0.500	1.500	2	1	0.750	0.250	0.125	6.125
	4	0.500	0.750	1	2	1.500	0.125	0.250	6.125
	5	1	0.750	0.500	1	3	0.125	0.500	6.875
	6	0.500	1.500	0.500	0.250	0.375	1	0.063	4.188
	7	0.500	0.375	0.250	0.500	1.500	0.063	1	4.188
CS_j		6.000	9.375	5.750	5.750	9.375	2.313	2.313	40.876

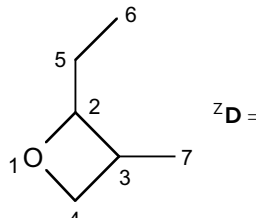
Derived from vertex- and edge-weighted molecular graphs, the **Barysz distance matrix** ${}^Z\mathbf{D}$ is a symmetric weighted distance matrix accounting contemporarily for the presence of heteroatoms and multiple bonds in the molecule; it is defined as [Barysz, Jashari *et al.*, 1983]:

$$[{}^Z\mathbf{D}]_{ij} = \begin{cases} d_{ij}(Z, \pi^*) & \text{if } i \neq j \\ 1 - \frac{Z_C}{Z_i} & \text{if } i = j \end{cases} \quad d_{ij}(Z, \pi^*) = \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \cdot \frac{Z_C^2}{Z_{b(1)} \cdot Z_{b(2)}} \right)$$

where Z_C is the atomic number of the carbon atom, Z_i is the atomic number of the i th atom, and π^* is the $\rightarrow$ *conventional bond order*; $d_{ij}(Z, \pi^*)$ is a weighted topological distance calculated by summing the edge weights over all d_{ij} bonds involved in the shortest path between vertices v_i and v_j , d_{ij} being the topological distance, and the subscripts $b(1)$ and $b(2)$ represent the two vertices incident to the b bond considered. Note that diagonal elements are atomic weights based on relative atomic numbers and the quantity in parenthesis is a bond weight based on conventional bond order and atomic numbers of the vertices incident to the b bond considered; some values of these atomic and bond parameters are listed in Tables W7 and W8.

Example W21

Barysz distance matrix for the H-depleted molecular graph of 2-ethyl-3-methyl-oxetane.

	Atom	1	2	3	4	5	6	7	VS_i
	1	0.143	0.857	1.857	0.857	1.857	2.857	2.857	11.285
	2	0.857	0	1	2	1	2	2	8.857
	3	1.857	1	0	1	2	3	1	9.857
	4	0.857	2	1	0	3	4	2	12.857
	5	1.857	1	2	3	0	1	3	11.857
	6	2.857	2	3	4	1	0	4	16.857
	7	2.857	2	1	2	3	4	0	14.857

The **Barysz index** J_{het} [Barysz, Jashari *et al.*, 1983] is a $\rightarrow$ *Balaban-like index* calculated by applying the $\rightarrow$ *Ivanciuc–Balaban operator IB* to the Barysz distance matrix:

$$J_{\text{het}} \equiv IB({}^Z\mathbf{D}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (VS_i({}^Z\mathbf{D}) \cdot VS_j({}^Z\mathbf{D}))^{-1/2}$$

where a_{ij} denotes the elements of the adjacency matrix $\mathbf{A}$, A is the number of atoms, B and C are the number of edges and rings, respectively, and VS_i and VS_j are the matrix row sums corresponding to two adjacent vertices. Moreover, a $\rightarrow$ *Wiener-type index* and a $\rightarrow$ *Schultz-type index* were calculated from the Barysz distance matrix and the *Barysz distance-plus-adjacency matrix*, respectively [Nikolić, Trinajstić *et al.*, 1993].

The **reciprocal Barysz distance matrix** is a square symmetric $A \times A$ matrix obtained by inverting the off-diagonal elements of the Barysz distance matrix as [Ivanciuc, Ivanciuc *et al.*, 1999a; Ivanciuc, 2000i]

$$[{}^Z\mathbf{D}^{-1}]_{ij} = \begin{cases} \frac{1}{d_{ij}(Z, \pi^*)} & \text{if } i \neq j \\ 1 - \frac{Z_C}{Z_i} & \text{if } i = j \end{cases}$$

The **complement Barysz distance matrix** is a square symmetric $A \times A$ matrix defined as [Ivanciuc, 2000i; Janežič, Miličević *et al.*, 2007]

$$[{}^C\mathbf{Z}\mathbf{D}]_{ij} = \begin{cases} A - d_{ij}(Z, \pi^*) & \text{if } i \neq j \\ 1 - \frac{Z_C}{Z_i} & \text{if } i = j \end{cases}$$

where A is the number of vertices. The **reciprocal complement Barysz distance matrix** was also defined in terms of the reciprocal of the off-diagonal elements of the complement Barysz distance matrix.

Example W22

Barysz distance matrix ${}^Z\mathbf{D}$, reciprocal Barysz distance matrix ${}^Z\mathbf{D}^{-1}$, and complement Barysz distance matrix ${}^C\mathbf{Z}\mathbf{D}$ of 5-methyl-1,3,4-oxathiazol-2-one; VS_i is the matrix row sum.

Atom	1	2	3	4	5	6	7	VS_i
1	0.250	0.750	1.125	1.179	0.750	1.125	1.750	6.929
2	0.750	0	0.375	0.696	1.500	0.375	2.500	6.196
3	1.125	0.375	0.625	0.321	0.750	0.750	1.750	5.696
4	1.179	0.696	0.321	0.143	0.429	1.071	1.429	5.268
5	0.750	1.500	0.750	0.429	0	1.875	1	6.304
6	1.125	0.375	0.750	1.071	1.875	0.250	2.875	8.321
7	1.750	2.500	1.750	1.429	1	2.875	0	11.304

${}^Z\mathbf{D}^{-1} =$	Atom	1	2	3	4	5	6	7	VS_i
	1	0.250	1.333	0.889	0.848	1.333	0.889	0.571	6.114
	2	1.333	0	2.667	1.436	0.667	2.667	0.400	9.169
	3	0.889	2.667	0.625	3.111	1.333	1.333	0.571	10.530
	4	0.848	1.436	3.111	0.143	2.333	0.933	0.700	9.505
	5	1.333	0.667	1.333	2.333	0	0.533	1	7.200
	6	0.889	2.667	1.333	0.933	0.533	0.250	0.348	6.953
	7	0.571	0.400	0.571	0.700	1	0.348	0	3.591
${}^CZ\mathbf{D} =$	Atom	1	2	3	4	5	6	7	VS_i
	1	6.750	6.250	5.875	5.821	6.250	5.875	5.250	42.071
	2	6.250	7	6.625	6.304	5.500	6.625	4.500	42.804
	3	5.875	6.625	6.375	6.679	6.250	6.250	5.250	43.304
	4	5.821	6.304	6.679	6.857	6.571	5.929	5.571	43.732
	5	6.250	5.500	6.250	6.571	7	5.125	6	42.696
	6	5.875	6.625	6.625	5.929	5.125	6.750	4.125	40.679
	7	5.250	4.500	5.250	5.571	6	4.125	7	37.696

The Barysz weighting scheme was generalized by Ivanciuc [Ivanciuc, Ivanciuc *et al.*, 1999a; Ivanciuc, 2000a, 2000i] in terms of the conventional bond order π^* and any atomic property. Based on the $\rightarrow$ *Ivanciuc weighting schemes*, several **Ivanciuc weighted distance matrices** were proposed to represent vertex- and edge-weighted molecular graphs [Ivanciuc, 2000i]. The general definition of these weighted matrices in terms of a generic atomic property w_i is the following:

$$[{}^w\mathbf{D}]_{ij} \equiv [\mathbf{D}(w)]_{ij} = \begin{cases} \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \cdot \frac{w_C^2}{(w_{b(1)} \cdot w_{b(2)})} \right) & \text{if } i \neq j \\ 1 - \frac{w_C}{w_i} & \text{if } i = j \end{cases}$$

where w can be equal to Z (atomic number), A (atomic mass), P (atomic polarizability), E (atomic electronegativity), R (covalent radius), X (relative atomic electronegativity), Y (relative covalent radius), and AH (atomic mass corrected for hydrogens). π^* is the $\rightarrow$ *conventional bond order*, the summation goes over all d_{ij} bonds involved in the shortest path between vertices v_i and v_j , d_{ij} being the topological distance, and the subscripts $b(1)$ and $b(2)$ represent the two vertices incident to the b th bond.

From Ivanciuc-weighted distance matrices, the corresponding **Ivanciuc weighted adjacency matrices** are derived by setting equal to zero all the matrix entries corresponding to pairs of nonadjacent vertices:

$${}^w\mathbf{D}_e = {}^w\mathbf{D} \otimes \mathbf{A}$$

where $\otimes$ is the $\rightarrow$ *Hadamard matrix product* and $\mathbf{A}$ the adjacency matrix.

Moreover, other graph-theoretical matrices were defined for molecular graphs weighted according to the schemes proposed by Ivanciuc [Ivanciuc, 2000i]. These are the weighted $\rightarrow$ *reciprocal distance matrix*, the weighted $\rightarrow$ *distance-valency matrices*, the weighted $\rightarrow$ *distance-path matrix*, the weighted $\rightarrow$ *reciprocal distance-path matrix*, the weighted $\rightarrow$ *distance complement matrix*, the weighted $\rightarrow$ *reciprocal distance complement matrix*, the weighted $\rightarrow$

complementary distance matrix, and the weighted $\rightarrow$ reciprocal complementary distance matrix. Several graph invariants can be derived from these weighted matrices applying the common matrix operators.

Still for vertex- and edge-weighted molecular graphs, a variant of the $\rightarrow$ bond length-weighted distance matrix was defined by including the atomic Pauling's electronegativity used as the weighting scheme for vertices. This is defined for $\rightarrow$ H-filled molecular graphs as [Yang, Wang *et al.*, 2003b]

$$[\mathbf{S}^*]_{ij} = \begin{cases} \frac{\alpha_{ij} \cdot \sqrt[n_{ij}]{\chi_i^{\text{PA}} \cdot \chi_k^{\text{PA}} \cdot \dots \cdot \chi_j^{\text{PA}}}}{d_{ij}(r)} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \quad d_{ij}(r) = \sum_{b=1}^{d_{ij}} r_b \quad \alpha_{ij} = \frac{\sum_{b=1}^{d_{ij}} \left(\frac{Z_{b(1)}}{Z_{b(2)}} \right)^{1/2}}{d_{ij}}$$

where $d_{ij}(r)$ is the sum of the bond lengths r of the edges along the shortest path connecting vertices v_i and v_j ; $\chi_i^{\text{PA}} \cdot \chi_k^{\text{PA}} \cdot \dots \cdot \chi_j^{\text{PA}}$ denote the Pauling's electronegativity values of the vertices along the path from v_i to v_j , n_{ij} is the total number of vertices along the path (i.e., $d_{ij} + 1$, d_{ij} being the topological distance between vertices v_i and v_j); and α_{ij} is the arithmetic mean of all α values for all the bonds involved in the path from v_i to v_j , each bond α value being the square root of the ratio of the atomic number $Z_{b(1)}$ of the positive atom over the atomic number $Z_{b(2)}$ of the negative atom forming the considered bond b . Note that in α calculation, hydrogen atoms are always taken as positive atoms.

The $\mathbf{W}^*$ index is a $\rightarrow$ Wiener-type index derived from the matrix $\mathbf{S}^*$ as

$$W^* \equiv Wi(\mathbf{S}^*) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{S}^*]_{ij}$$

where Wi is the $\rightarrow$ Wiener operator.

Moreover, the same index was also calculated from a variant of the matrix $\mathbf{S}^*$, defined in terms of the total energies of valence electrons of vertices as [Yang, Wang *et al.*, 2003a]

$$[{}^E\mathbf{S}^*]_{ij} = \begin{cases} \frac{\sqrt[n_{ij}]{E_i \cdot E_k \cdot \dots \cdot E_j}}{d_{ij}(Z)} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \quad d_{ij}(Z) = \sum_{b=1}^{d_{ij}} \left[(Z_{b(1)} - Z_{b(1)}^v - 1) + (Z_{b(2)} - Z_{b(2)}^v - 1) \right]_b$$

where $d_{ij}(Z)$ is a weighted topological distance calculated by summing over all edges b in the shortest path between vertices v_i and v_j , a sort of bond length approximated by the total number of inner-shell electrons of the two atoms $b(1)$ and $b(2)$ forming the bond b ; Z is the atomic number and Z^v the number of valence electrons of an atom. E_i is the total energy of the valence electrons of the i th atom calculated as:

$$E_i = Z_i^v \cdot \bar{e}_i \quad \bar{e}_i = \frac{\sum_k e_{ik} \cdot n_{ik}}{\sum_k n_{ik}} \quad e_{ik} = L_i + \sqrt{2 \cdot l_k} \quad l_k = 0, 1, 2, 3$$

where $\bar{e}_i$ is the average energy of valence electrons of the i th atom, e_{ik} the energy of an electron in the subshell k , n_{ik} the number of electrons in the subshell, L_i the principal quantum number, and l_k the azimuthal quantum number, which takes values 0, 1, 2, and 3 for s, p, d, and f atomic orbital shells.

Also **Schultz weighted distance matrices** were proposed, weighting both vertices and edges in the H-depleted molecular graph to account contemporarily for heteroatoms and bond multiplicity [Schultz, Schultz *et al.*, 1994]. Three different graph weighting schemes were defined, which allow the calculation of different weighted matrices and the corresponding graph invariants.

The vertex weighting w_i is based on atomic numbers:

$$w_i(Z) = 1 + Z_i - Z_C = 1 + Z_i - 6$$

where Z_i is the atomic number of the i th vertex and Z_C is the atomic number of carbon atom. Note that the vertex weight is equal to unity for all carbon atoms and is defined for characterizing heteroatoms, the heavier the heteroatoms the greater the weight value. Values of the weights for the most common atoms are reported in Table W4; as it can be noted some vertex weights have negative values.

Table W4 Schultz's vertex weights for the most common atoms.

Atom	w	Atom	w
Li	-2	Al	8
Be	-1	Si	9
B	0	P	10
C	1	S	11
N	2	Cl	12
O	3	Br	30
F	4	I	48

The vertex weights are collected into a diagonal matrix $\mathbf{W}^V$, which is used as a premultiplier of the $\rightarrow$ distance matrix $\mathbf{D}$ to give the **vertex-weighted Schultz distance matrix** $\mathbf{D}^V$ as

$$\mathbf{D}^V = \mathbf{W}^V \cdot \mathbf{D}$$

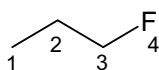
Thus, elements of the vertex-weighted distance matrix are formally defined as

$$[\mathbf{D}^V]_{ij} = \begin{cases} w_i \cdot d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where w_i is the weight of the i th vertex and d_{ij} the $\rightarrow$ topological distance between vertices v_i and v_j .

Example W23

The H-depleted molecular graph of 1-fluoropropane is



The vertex weights are $w_1(Z) = 1$, $w_2(Z) = 1$, $w_3(Z) = 1$, and $w_4(Z) = 4$. The distance matrix of 1-fluoropropane is

$$\mathbf{D} = \begin{vmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{vmatrix}$$

Then, the vertex-weighted distance matrix $\mathbf{D}^V$ of 1-fluoropropane is

$$\mathbf{D}^V = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix} \times \begin{vmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 12 & 8 & 4 & 0 \end{vmatrix}$$

It is worth noting that the vertex-weighted Schultz distance matrix is unsymmetrical.

The weight w_{ij} associated with the edge between vertices v_i and v_j is defined as

$$w_{ij} = m_{ij} + (Z_i - Z_C) + (Z_j - Z_C)$$

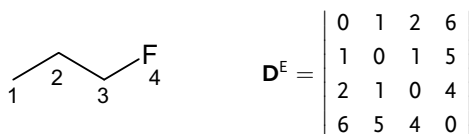
where m_{ij} is the multiplicity of the edge between vertices v_i and v_j , having value of 1 for a single, 2 for a double, and 3 for a triple bond; Z_i and Z_j are the atomic numbers of the bonded atoms; and Z_C is the atomic number of the carbon atom. Note that for carbon–carbon bonds, the edge weights simplify to the bond multiplicity; moreover, for single carbon–heteroatom bonds and in the case of elements lying to the left of carbon in the second quantum level of the periodic table, the edge weights can be zero or negative. The **edge-weighted Schultz distance matrix** $\mathbf{D}^E$ is obtained by summing up the weights of all edges involved in the shortest path between two considered vertices.

$$[\mathbf{D}^E]_{ij} = \begin{cases} d_{ij}(w) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \quad d_{ij}(w) = \min_{p_{ij}} \left(\sum_{b=1}^{d_{ij}} w_b \right)$$

where w_b is the edge weight, the sum runs over all the edges along the path p_{ij} , and the i – j matrix entry is the minimum value over all the shortest paths between vertices v_i and v_j ; d_{ij} is the topological distance, that is, the number of edges along the shortest path.

Example W24

The edge-weighted Schultz distance matrix of 1-fluoropropane, derived from the edge weights $w_{12}(Z) = 1$, $w_{23}(Z) = 1$, and $w_{34}(Z) = 4$, is



The **edge-vertex-weighted Schultz distance matrix** $\mathbf{D}^{EV}$ is derived by multiplying the diagonal matrix $\mathbf{W}^V$ collecting the vertex weights by the edge-weighted distance matrix collecting weighted distances as

$$\mathbf{D}^{\text{EV}} = \mathbf{W}^{\text{V}} \cdot \mathbf{D}^{\text{E}}$$

Thus, elements of the edge–vertex-weighted distance matrix are formally defined as

$$[\mathbf{D}^{\text{EV}}]_{ij} = \begin{cases} w_i \cdot d_{ij}(w) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where w_i is the weight of the i th vertex, and $d_{ij}(w)$ the weighted topological distance between vertices v_i and v_j , calculated as the sum of the weights of the edges along the shortest path.

Example W25

The edge–vertex-weighted Schultz distance matrix of 1-fluoropropane is

$$\mathbf{D}^{\text{EV}} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix} \times \begin{vmatrix} 0 & 1 & 2 & 6 \\ 1 & 0 & 1 & 5 \\ 2 & 1 & 0 & 4 \\ 6 & 5 & 4 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 1 & 2 & 6 \\ 1 & 0 & 1 & 5 \\ 2 & 1 & 0 & 4 \\ 24 & 20 & 16 & 0 \end{vmatrix}$$

Note that, as for the vertex-weighted distance matrix, the matrix inner product leads to an unsymmetrical matrix.

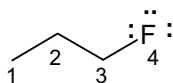
The third weighting scheme is related to vertex valence and takes into account the number of bonds incident to a vertex and its pairs of unshared electrons. Each electron pair present counts one, each missing electron pair contributes -1 to the total valence of atoms, and in the case of free radicals the unique electron present in the outer valence shell contributes half a bond to the total valence. The vertex weight w_i is then defined as

$$w_i(\delta) = \delta_i + n_i^{(\bullet\bullet)} - n_i^{(\ominus\ominus)} + \frac{1}{2} I_i^{(\bullet)}$$

where δ is the $\rightarrow$ vertex degree, $n^{(\bullet\bullet)}$ is the number of unshared lone-pairs, $n^{(\ominus\ominus)}$ is the number of missing lone-pairs to complete the electron octet in the outer valence shell, and $I^{(\bullet)}$ denotes an indicator variable for radicals. This weighting scheme allows to distinguish the charged/uncharged atoms and thus the derived molecular descriptors are also suitable for describing properties for charged species.

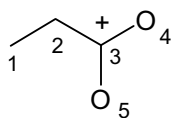
Example W26

The molecular graph of 1-fluoropropane where electron pairs of F are indicated is:



While valence weights of carbon atoms are all equal to their vertex degrees, that is, $w_1(\delta) = 1$, $w_2(\delta) = 2$, and $w_3(\delta) = 2$, the weight for fluorine is equal to the vertex degree plus 3, which is the number of unshared electron pairs, that is, $w_4(\delta) = 1 + 3 = 4$.

The molecular graph of carbonyl oxygen protonated propanoic acid is



The carbon atom 3 has valence weight $w_3(\delta) = 3 - 1 = 2$, because it is positively charged.

The valence weights are collected into a diagonal matrix $\mathbf{W}^\delta$ used to give the **valence-weighted Schultz distance matrix** $\mathbf{D}^\delta$:

$$\mathbf{D}^\delta = \mathbf{W}^\delta \cdot \mathbf{D}$$

and the **edge-vertex-valence-weighted Schultz distance matrix** $\mathbf{D}^{\text{EV}\delta}$:

$$\mathbf{D}^{\text{EV}\delta} = \mathbf{W}^\delta \cdot \mathbf{D}^{\text{EV}}$$

Analogously, the **edge-valence-weighted Schultz distance matrix** $\mathbf{D}^{\text{E}\delta}$ and the **vertex-valence-weighted Schultz distance matrix** $\mathbf{D}^{\text{V}\delta}$ are defined as

$$\mathbf{D}^{\text{E}\delta} = \mathbf{W}^\delta \cdot \mathbf{D}^{\text{E}} \quad \text{and} \quad \mathbf{D}^{\text{V}\delta} = \mathbf{W}^\delta \cdot \mathbf{D}^{\text{V}}$$

Example W27

The valence-weighted distance matrix of 1-fluoropropane is

$$\mathbf{D}^\delta = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix} \times \begin{vmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 1 & 2 & 3 \\ 2 & 0 & 2 & 4 \\ 4 & 2 & 0 & 2 \\ 12 & 8 & 4 & 0 \end{vmatrix}$$

The edge-vertex-valence-weighted distance matrix of 1-fluoropropane is

$$\mathbf{D}^{\text{EV}\delta} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix} \times \begin{vmatrix} 0 & 1 & 2 & 6 \\ 1 & 0 & 1 & 5 \\ 2 & 1 & 0 & 4 \\ 24 & 20 & 16 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 1 & 2 & 6 \\ 2 & 0 & 2 & 10 \\ 4 & 2 & 0 & 8 \\ 96 & 80 & 64 & 0 \end{vmatrix}$$

The edge-valence-weighted distance matrix of 1-fluoropropane is

$$\mathbf{D}^{\text{E}\delta} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix} \times \begin{vmatrix} 0 & 1 & 2 & 6 \\ 1 & 0 & 1 & 5 \\ 2 & 1 & 0 & 4 \\ 6 & 5 & 4 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 1 & 2 & 6 \\ 2 & 0 & 2 & 10 \\ 4 & 2 & 0 & 8 \\ 24 & 20 & 16 & 0 \end{vmatrix}$$

The vertex-valence-weighted distance matrix of 1-fluoropropane is

$$\mathbf{D}^{\text{V}\delta} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix} \times \begin{vmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 12 & 8 & 4 & 0 \end{vmatrix} = \begin{vmatrix} 0 & 1 & 2 & 3 \\ 2 & 0 & 2 & 4 \\ 4 & 2 & 0 & 2 \\ 48 & 32 & 16 & 0 \end{vmatrix}$$

The most common topological indices can be derived from the above defined weighted distance matrices, for example, the $\rightarrow$ *Schultz-type indices* were calculated as well as the $\rightarrow$ *MTI' index*, $\rightarrow$ *determinant*, $\rightarrow$ *permanent*, $\rightarrow$ *product of row sums*, and the $\rightarrow$ *hafnian*.

From the vertex–valence-weighted distance matrix $\mathbf{D}^{\text{V}\delta}$, a **geometric modification number** GM was also derived to account for the geometric isomerism of compounds [Schultz, Schultz *et al.*, 1995]. **Geometric factors** GF equal to +1 are assigned to vertices with *Z* (*cis*) geometry and –1 to vertices with *E* (*trans*) geometry, all other vertices in the graph are assigned geometric factors equal to zero; for large molecules, the GF values of the various paired centers of geometric isomerism are algebraically summed to arrive at correct GF values for each vertex. The geometric factors are collected into a diagonal matrix $\mathbf{W}^{\text{GF}}$. Then, this matrix is used as a premultiplier of the sum of the vertex–valence-weighted distance matrix $\mathbf{D}^{\text{V}\delta}$ with its transposed matrix $(\mathbf{D}^{\text{V}\delta})^{\text{T}}$:

$$\mathbf{W}^{\text{GF}} \cdot [\mathbf{D}^{\text{V}\delta} + (\mathbf{D}^{\text{V}\delta})^{\text{T}}]$$

where the matrix sum $\mathbf{D}^{\text{V}\delta} + (\mathbf{D}^{\text{V}\delta})^{\text{T}}$ is performed to obtain a symmetrical matrix and to incorporate as much information as possible into rows representing the geometric centers of the molecule. The row sums of the final matrix relative to geometric centers give the geometric modification number GM of the molecule. The GM numbers for two isomers (*Z* and *E*) will have opposite signs. The derived GM values can be incorporated (i.e., added) into any given topological index to achieve discrimination among geometric isomers.

The same technique was proposed to discriminate among optical isomers, adding to topological indices a **chiral modification number** CM that is calculated in the same way as the geometric modification number GM, the only difference being to assign **chiral factors** CF of +1 to chiral stereocenters with *R* configuration and chiral factors of –1 to stereocenters with *S* configuration; all other vertices in the graph are assigned factor values of zero.

Moreover, **rotamer factors** RF of +1 were proposed to be assigned to single bonded vertices with *C* geometry and –1 to vertices with *T* geometry. The **rotamer modification number** RMN based on rotamer factors and vertex–valence-weighted distance matrix allows to discriminate among conformational isomers when it is incorporated into the topological indices. It is calculated in the same way as the geometric modification number GM and chiral modification number CM [Schultz, Schultz *et al.*, 1996].

Two general classes of symmetric $A \times A$ weighted distance matrices are the **interaction graph matrices**, denoted by **IM**, and **perturbation graph matrices**, denoted by **PM**, which are defined as [Authors, This Book]

$$[\mathbf{IM}(w; \alpha, \beta, \gamma, \lambda)]_{ij} = \begin{cases} \frac{(w_i \cdot w_j)^\lambda}{f(d_{ij}, \gamma)} & \text{if } i \neq j \\ \alpha \cdot w_i^\beta & \text{if } i = j \end{cases}$$

$$[\mathbf{PM}(w; \alpha, \beta, \gamma, \lambda)]_{ij} = \begin{cases} \frac{|w_j - w_i|^\lambda}{f(d_{ij}, \gamma)} & \text{if } i \neq j \\ \alpha \cdot w_i^\beta & \text{if } i = j \end{cases}$$

where w is a weighting scheme for graph vertices, d_{ij} the topological distance between vertices v_i and v_j , f is a smoothing function used to modulate the role of distances in defining contributions

from vertices far apart. α , β , γ , and λ are user-defined real parameters. α and β are tuning parameters for the diagonal matrix elements, associated with the vertex weights; α can be set at zero, when one does not want explicit information on atom properties, or at 1 to generate augmented weighted matrices or, for instance, chosen equal to the reciprocal of the property of carbon atom ($\alpha = 1/w_C$) to have relative atomic properties on the main diagonal. γ is a distance smoothing parameter while the parameter λ is a real exponent associated with the interaction between the properties of two vertices.

The most common distance smoothing functions are

$$f_1(d_{ij}, \gamma) = d_{ij}^\gamma \quad f_2(d_{ij}, \gamma) = (d_{ij} + 1)^\gamma \quad f_3(d_{ij}, \gamma) = 2^{\gamma \cdot d_{ij}} \quad f_4(d_{ij}, \gamma, x) = (d_{ij} \cdot x^{(d_{ij}-1)})^\gamma$$

It can be noted that function f_1 was used by Ivanciuc to define the $\rightarrow$ *distance-valency matrices*, function f_2 was used by Kier-Hall to define the $\rightarrow$ *electrotopological state indices*, function f_3 was proposed by Randić to calculate the $\rightarrow$ *augmented valence* of vertices, and function f_4 was proposed by Estrada in the $\rightarrow$ *generalized molecular-graph matrix*, where x is a user-defined parameter or a parameter to be optimized during the searching for the best QSAR correlations.

It can be easily seen that interaction matrices reduce to some well-known graph-theoretical matrices by using specific combinations of α , β , γ , and λ parameters and distance smoothing functions (Tables W5 and W6). For instance, the interaction matrix calculated from a simple graph (i.e., $w_i = 1$) by using the distance function $f_1 = d_{ij}^{-1}$ with $\gamma = -1$ and setting $\alpha = 0$ to have diagonal elements equal to zero coincides with the $\rightarrow$ *distance matrix* $\mathbf{D}$: $\mathbf{IM}(1; 0, 0, -1, 0) \equiv \mathbf{D}$. The $\rightarrow$ *Harary matrix* $\mathbf{H}$ is obtained by the same parameter combination, except for γ parameter that need to be set at 1: $\mathbf{IM}(1; 0, 0, 1, 0) \equiv \mathbf{H}$. $\rightarrow$ *Augmented distance matrices* $^a\mathbf{D}$ are obtained in the same way as the distance matrix but by weighting vertices by some property w_i different from 1 and setting $\alpha = \beta = 1$ to retain atomic properties on the main diagonal of the matrix: $\mathbf{IM}(w; 1, 1, -1, 0) \equiv ^a\mathbf{D}$.

It is also worth noting to point out that, unlike the other weighted distance matrices, whose elements are calculated by combining weights of all the vertices along the shortest path, in interaction and perturbation graph matrices, only the properties of the two terminal vertices of the shortest path are considered. Some function of the length of this shortest path is then used to smooth interactions between the terminal vertices. Moreover, instead of the simple topological distance d_{ij} , which is the path length obtained when each edge contributes 1, a weighted distance $d_{ij}(w)$ calculated as the sum of the bond lengths or the reciprocal of the bond orders of the edges along the shortest path can be used. In this case, the $\rightarrow$ *multigraph distance matrix*, the $\rightarrow$ *atomic weight-weighted distance matrix*, and the $\rightarrow$ *bond length-weighted distance matrix* can be obtained from interaction graph matrices by using the distance smoothing function $f_1 = d_{ij}^\gamma(w)$ and the following set of parameters: $\alpha = 0$, $\beta = 0$, $\gamma = -1$, and $\lambda = 0$.

An interesting perturbation matrix is that derived from molecular graphs where vertices are weighted by the $\rightarrow$ *intrinsic states* I_i and choosing the second distance smoothing function f_2 ; then, for $\alpha = 1$, $\beta = 1$, and $\lambda = 1$, this perturbation matrix is defined as

$$[\mathbf{PM}(I; 1, 1, \gamma, 1)]_{ij} = \begin{cases} \frac{|I_j - I_i|}{(d_{ij} + 1)^\gamma} & \text{if } i \neq j \\ I_i & \text{if } i = j \end{cases}$$

The row sum of this perturbation matrix is a local vertex invariant defined as

$$VS_i(\mathbf{PM}(I; 1, 1, \gamma, 1)) = I_i + \sum_{j=1}^A \frac{|I_j - I_i|}{(d_{ij} + 1)^\gamma} = I_i + \Delta I_i$$

where VS_i is the matrix row sum and ΔI_i a global perturbation term accounting for the differences between the intrinsic state of the i th atom and the intrinsic states of the other atoms in the environment of i .

It can be noted that the definition of this local vertex invariant is very similar to that of the electrotopological state index S_i , but this is based on absolute values of the perturbation terms $|I_j - I_i|$, with the consequence that the sum of ΔI_i over all the vertices is different from zero. Absolute values were introduced to avoid the sum of the matrix off-diagonal elements results in zero for any weighting scheme w considered.

Unsymmetrical interaction graph matrices, denoted by **UIM**, are defined by considering only the property w_j of the vertex v_j that is the terminal vertex of the shortest path starting from v_i as [Authors, This Book]

$$[\mathbf{UIM}(w; \alpha, \beta, \gamma, \lambda)]_{ij} = \begin{cases} \frac{w_j^\lambda}{f(d_{ij}, \gamma)} & \text{if } i \neq j \\ \alpha \cdot w_i^\beta & \text{if } i = j \end{cases}$$

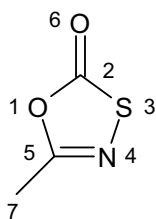
where w is any weighting scheme for graph vertices, d_{ij} denotes the $\rightarrow$ *topological distance* between vertices v_i and v_j ; α , β , γ , and λ are the user-defined real parameters defined above. By choosing the distance function f_3 with $\alpha = 1$, $\beta = 1$, $\gamma = 1$, and $\lambda = 1$, this matrix can be regarded as generalization of the $\rightarrow$ *augmented vertex degree matrix* by using an atomic property w in place of the vertex degree δ . Each row sum of the unsymmetrical interaction matrix is a local vertex invariant that encodes information about the sum of the properties of all vertices in the graph from the point of view of a single vertex. Property contributions usually decrease as the distance from the focused vertex increases.

By symmetrization of the unsymmetrical interaction matrices by the Hadamard matrix product, interaction graph matrices **IM** are obtained for a certain combination of the α , β , γ , and λ parameters:

$$\mathbf{IM}(w; \alpha', \beta', \gamma', \lambda') = \mathbf{UIM}(w; \alpha, \beta, \gamma, \lambda)^T \otimes \mathbf{UIM}(w; \alpha, \beta, \gamma, \lambda)$$

Example W28

Perturbation graph matrix $\mathbf{PM}(m; 1, 1, 1, 1)$, unsymmetrical $\mathbf{UIM}(m; 1, 1, 1, 1)$, and symmetric $\mathbf{IM}(m; 1, 2, 2, 1)$ interaction matrices of 5-methyl-1,3,4-oxathiazol-2-one; VS_i is the matrix row sum and CS_j the matrix column sum. Calculation was performed on the following relative atomic masses: 1, 1.332, 1.166, and 2.670 for carbon, oxygen, nitrogen, and sulfur, respectively; the chosen distance function was $f_3(d_{ij}, \gamma) = 2^{\gamma \cdot d_{ij}}$. $\mathbf{IM}(m; 1, 2, 2, 1)$ was generated by symmetrization of $\mathbf{UIM}(m; 1, 1, 1, 1)$. $Wi(\mathbf{PM}(m; 1, 1, 1, 1))$ and $Wi(\mathbf{IM}(m; 1, 2, 2, 1))$ are the $\rightarrow$ *Wiener-type indices*.

**PM(m;1,1,1,1)**

	1	2	3	4	5	6	7	VS_i
1	1.332	0.166	0.334	0.041	0.166	0	0.083	2.122
2	0.166	1	0.835	0.042	0	0.166	0	2.209
3	0.334	0.835	2.667	0.752	0.417	0.334	0.209	5.548
4	0.041	0.042	0.752	1.166	0.083	0.021	0.042	2.147
5	0.166	0	0.417	0.083	1	0.042	0	1.708
6	0	0.166	0.334	0.021	0.042	1.332	0.020	1.915
7	0.083	0	0.209	0.042	0	0.020	1	1.354
CS_j	2.122	2.209	5.548	2.147	1.708	1.915	1.354	17.003

$$Wi(\mathbf{PM}(m;1,1,1,1)) = 8.502$$

UIM(m;1,1,1,1)

	1	2	3	4	5	6	7	VS_i
1	1.332	0.500	0.667	0.292	0.500	0.333	0.250	3.874
2	0.666	1	1.335	0.292	0.250	0.666	0.125	4.334
3	0.333	0.500	2.67	0.583	0.250	0.333	0.125	4.794
4	0.333	0.250	1.335	1.166	0.500	0.167	0.250	4.001
5	0.666	0.250	0.667	0.583	1	0.167	0.500	3.833
6	0.333	0.500	0.667	0.146	0.125	1.332	0.063	3.166
7	0.333	0.125	0.334	0.292	0.500	0.083	0	1.667
CS_j	3.996	3.125	7.676	3.352	3.125	3.080	1.313	25.67

IM(m;1,2,2,1)

	1	2	3	4	5	6	7	VS_i
1	1.774	0.333	0.222	0.097	0.333	0.111	0.083	2.954
2	0.333	1	0.668	0.073	0.063	0.333	0.016	2.485
3	0.222	0.668	7.129	0.778	0.167	0.222	0.042	9.228
4	0.097	0.073	0.778	1.360	0.292	0.024	0.073	2.696
5	0.333	0.063	0.167	0.292	1	0.021	0.250	2.125
6	0.111	0.333	0.222	0.024	0.021	1.774	0.005	2.491
7	0.083	0.016	0.042	0.073	0.250	0.005	0	0.469
CS_j	2.954	2.485	9.228	2.696	2.125	2.491	0.469	22.447

$$Wi(\mathbf{IM}(m;1,2,2,1)) = 11.223$$

When $\gamma = 0$, a special class of interaction and perturbation matrices is obtained, accounting only for relationships between properties of pairs of vertices, independent of their mutual location in the graph. These matrices are then defined as

$$[\mathbf{IM}(w; \alpha, \beta, 0, \lambda)]_{ij} = \begin{cases} (w_i \cdot w_j)^\lambda & \text{if } i \neq j \\ \alpha \cdot w_i^\beta & \text{if } i = j \end{cases}$$

$$[\mathbf{PM}(w; \alpha, \beta, 0, \lambda)]_{ij} = \begin{cases} |w_j - w_i|^\lambda & \text{if } i \neq j \\ \alpha \cdot w_i^\beta & \text{if } i = j \end{cases}$$

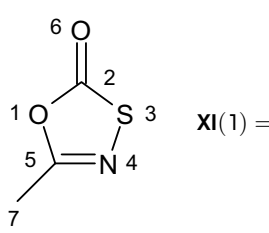
The **XI matrix**, proposed by Ivanciuc [Ivanciuc, 1999c], is an example of these interaction graph matrices with $\alpha = 0$, $\gamma = 0$, and $\lambda = -1/2$. This matrix is derived from the $\rightarrow$ distance–valency matrices **Dval** as

$$[\mathbf{XI}(w)]_{ij} \equiv [\mathbf{Dval}(0, -0.5, -0.5; w)]_{ij} = \begin{cases} 1/\sqrt{\text{val}_i(w) \cdot \text{val}_j(w)} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $\text{val}(w)$ is the $\rightarrow$ valency of a vertex calculated from a weighting scheme w , that is, the sum of the weights of the edges to the first neighbors.

Example W29

XI matrix ($w = 1$) of 5-methyl-1,3,4-oxathiazol-2-one.

	Atom	1	2	3	4	5	6	7
	1	0	0.408	0.500	0.500	0.408	0.707	0.707
	2	0.408	0	0.408	0.408	0.333	0.577	0.577
	3	0.500	0.408	0	0.500	0.408	0.707	0.707
	4	0.500	0.408	0.500	0	0.408	0.707	0.707
	5	0.408	0.333	0.408	0.408	0	0.577	0.577
	6	0.707	0.577	0.707	0.707	0.577	0	1.000
	7	0.707	0.577	0.707	0.707	0.577	1.000	0

The calculation of the matrix was based on the following vertex degrees, which coincide with the vertex valencies $\text{val}_i(1)$ for a unitary weighting scheme: $\delta_1 = \text{val}_1(1) = 2$, $\delta_2 = \text{val}_2(1) = 3$, $\delta_3 = \text{val}_3(1) = 2$, $\delta_4 = \text{val}_4(1) = 2$, $\delta_5 = \text{val}_5(1) = 3$, $\delta_6 = \text{val}_6(1) = 1$, and $\delta_7 = \text{val}_7(1) = 1$.

From interaction and perturbation graph matrices with $\gamma = 0$, other matrices can be derived by introducing again information on distances between pairs of vertices with the aid of the $\rightarrow$ geodesic matrices ${}^k\mathbf{B}$. The resulting matrices, called **interaction geodesic matrices** ${}^k\mathbf{IM}$ and **perturbation geodesic matrices** ${}^k\mathbf{PM}$, are then calculated as [Authors, This Book]

$${}^k\mathbf{IM}(w; 0, 0, 0, \lambda) = \mathbf{IM}(w; \alpha, \beta, 0, \lambda) \otimes {}^k\mathbf{B} \quad {}^k\mathbf{PM}(w; 0, 0, 0, \lambda) = \mathbf{PM}(w; \alpha, \beta, 0, \lambda) \otimes {}^k\mathbf{B}$$

where $\otimes$ indicates the Hadamard matrix product and ${}^k\mathbf{B}$ is the k th order geodesic matrix, whose elements are equal to 1 only for vertices v_i and v_j at topological distance k and zero otherwise. The elements of the interaction and perturbation geodesic matrices are formally defined as

$$[{}^k\mathbf{IM}(w; 0, 0, 0, \lambda)]_{ij} = \begin{cases} (w_i \cdot w_j)^\lambda & \text{if } i \neq j \wedge d_{ij} = k \\ 0 & \text{otherwise} \end{cases}$$

$$[{}^k\mathbf{PM}(w; 0, 0, 0, \lambda)]_{ij} = \begin{cases} |w_j - w_i|^\lambda & \text{if } i \neq j \wedge d_{ij} = k \\ 0 & \text{otherwise} \end{cases}$$

where w is any atomic property, k is the matrix order, and d_{ij} is the topological distance between vertices v_i and v_j . These matrices consist of a few nonvanishing elements that encode information on properties of vertices located at a given distance k in the graph.

It is straightforward to observe that, for $\lambda = 1$, the half sum of the elements of the k th order interaction geodesic matrix reduces to the $\rightarrow$ *Moreau–Broto autocorrelation* descriptor of order k (ATS_k):

$$ATS_k = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [{}^k\mathbf{IM}(w; 0, 0, 0, 1)]_{ij}$$

Therefore, Moreau–Broto autocorrelation descriptors can be thought as the $\rightarrow$ *Wiener-type indices* derived from interaction geodesic matrices.

It should be noted that from interaction and perturbation matrices geodesic augmented matrices cannot be obtained, being always zeroed all the diagonal elements.

Interaction and perturbation geodesic matrices of order 1 are also referred to as *edge-interaction* and *edge-perturbation graph matrices*, denoted by $\mathbf{IM}_e$ and $\mathbf{PM}_e$, since they encode information only about pairs of adjacent vertices.

Moreover, from the interaction geodesic matrices, with $\alpha = \beta = \gamma = 0$, and $\lambda = 1$, **atom-type interaction matrices** can be obtained by using in place of the vertex weights atom-type indicator variables: [Authors, This Book]

$$[{}^k\mathbf{IM}(u, v; 0, 0, 0, 1)]_{ij} = \begin{cases} \delta(i; u) \times \delta(j; v) & \text{if } i \neq j \wedge d_{ij} = k \\ 0 & \text{otherwise} \end{cases}$$

where u and v are two atom-types; $\delta(i; u)$ is a Kronecker delta function equal to 1 if the vertex i is of type u , and zero otherwise; analogously, $\delta(j; v)$ is a Kronecker delta function equal to 1 if the vertex j is of type v , and zero otherwise. If atom-types u and v are different from each other, then these matrices are unsymmetrical.

The sum of the elements of this matrix is the $\rightarrow$ *atom-type autocorrelation*, that is,

$$ATAC_k(u, v) = \sum_{i=1}^A \sum_{j=1}^A [{}^k\mathbf{IM}(u, v; 0, 0, 0, 1)]_{ij} \quad v \neq u$$

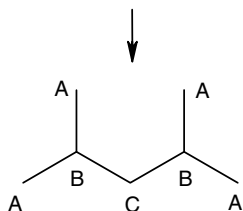
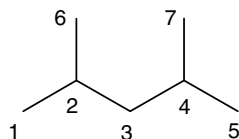
where $ATAC_k(u, v)$ is the number of occurrences of the pairs of atoms of types u and v at topological distance of k . If atom-types u and v coincide (e.g., lipophilic–lipophilic), then the atom-type autocorrelation is calculated as the half sum of the matrix elements:

$$ATAC_k(u, u) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [{}^k\mathbf{IM}(u, u; 0, 0, 0, 1)]_{ij}$$

Example W30

Atom-type interaction matrices ${}^1\mathbf{IM}(A, B; 0, 0, 0, 1)$, ${}^3\mathbf{IM}(A, B; 0, 0, 0, 1)$, and ${}^2\mathbf{IM}(A, A; 0, 0, 0, 1)$ for two pairs of atom types, AB and AA, and different distance values.

Atoms are assigned the following atom types: A, B, and C. VS_i and CS_j are the matrix row and column sums, respectively; $ATAC_k$ indicates the atom-type autocorrelation.

 ${}^1\mathbf{IM}(A, B; 0, 0, 0, 1)$

	1	2	3	4	5	6	7	VS_i
1	0	1	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	0	1
6	0	1	0	0	0	0	0	1
7	0	0	0	1	0	0	0	1
CS_j	0	2	0	2	0	0	0	4

$$ATAC_1(A, B) = 4$$

 ${}^3\mathbf{IM}(A, B; 0, 0, 0, 1)$

	1	2	3	4	5	6	7	VS_i
1	0	0	0	1	0	0	0	1
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	1
6	0	0	0	1	0	0	0	1
7	0	1	0	0	0	0	0	1
CS_j	0	2	0	2	0	0	0	4

$$ATAC_3(A, B) = 4$$

 ${}^2\mathbf{IM}(A, A; 0, 0, 0, 1)$

	1	2	3	4	5	6	7	VS_i
1	0	0	0	0	0	1	0	1
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	1
6	1	0	0	0	0	0	0	1
7	0	0	0	0	1	0	0	1
CS_j	1	0	0	0	1	1	1	4

$$ATAC_2(A, A) = 2$$

Then, for each topological distance k , a square symmetric general atom-type matrix, called **atom-type autocorrelation matrix**, denoted as ${}^k\mathbf{ATAC}$, of size $N_{at} \times N_{at}$, where N_{at} is the total number of defined atom types, can be defined as [Authors, This Book]

$$[{}^k\mathbf{ATAC}]_{uv} = ATAC_k(u, v) = \begin{cases} \sum_{i=1}^A \sum_{j=1}^A \delta(i; u) \times \delta(j; v) & \text{if } v \neq u \\ \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \delta(i; u) \times \delta(j; u) & \text{if } v = u \end{cases}$$

where u and v refer to atom types, A is the number of graph vertices; $\delta(i; u)$ is a Kronecker delta function equal to 1 if the vertex i is of type u and zero otherwise; analogously, $\delta(j; v)$ is a Kronecker delta function equal to 1 if the vertex j is of type v and zero otherwise.

A special case of this matrix is obtained by considering direct graphs or ordered sequences of atom types such as amino acid sequences. In this case, only the occurrences of atom type v ,

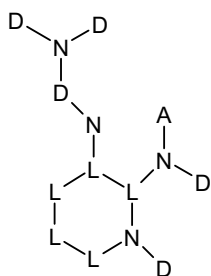
which follows atom type u at a given topological distance k , is considered. This kind of matrix was proposed by Randić for characterizing amino acid sequences and was called $\rightarrow$ *amino acid adjacency matrix*.

Each u th row sum of the atom-type autocorrelation matrix is the number of vertices of any type located at distance k from vertices of type u ; if all the graph vertices are assigned an atom type, then the $\rightarrow$ *Wiener-type index* of the matrix gives the $\rightarrow$ *graph distance count* of order k . Note that since diagonal elements of atom-type autocorrelation matrices can be different from zero, the $\rightarrow$ *Wiener operator* has to be applied as the following:

$$Wi(^k\text{ATAC}) = \sum_{u=1}^{N_{at}} \sum_{v=u}^{N_{at}} [^k\text{ATAC}]_{uv}$$

Example W31

Atom-type autocorrelation matrices $^1\text{ATAC}$, $^2\text{ATAC}$, and $^3\text{ATAC}$ of order 1, 2, and 3, respectively. Atoms are assigned the following atom types: A (acceptor), D (donor), L (lipophilic); type N is used for all atoms not assigned to defined types. Wi is the $\rightarrow$ *Wiener operator*, ^kf is the $\rightarrow$ *graph distance count* of k th order, and B is the number of edges.



$^1\text{ATAC}$

	A	D	L	N	VS_i
A	0	0	0	1	1
D	0	0	0	6	6
L	0	0	4	4	8
N	1	6	4	0	11

$$Wi(^1\text{ATAC}) = ^1f = B = 15$$

$^2\text{ATAC}$

	A	D	L	N	VS_i
A	0	1	1	0	2
D	1	3	4	0	8
L	1	4	4	5	14
N	0	0	5	2	7

$$Wi(^2\text{ATAC}) = 2f = 20$$

$^3\text{ATAC}$

	A	D	L	N	VS_i
A	0	0	1	1	2
D	0	0	5	4	9
L	1	5	2	5	13
N	1	4	5	2	12

$$Wi(^3\text{ATAC}) = ^3f = 20$$

Simple **unsymmetrical interaction geodesic matrices**, denoted by ^kUIM , are obtained from unsymmetrical interaction matrices UIM by setting $\alpha = \gamma = 0$ and applying the $\rightarrow$ *Hadamard matrix product* [Authors, This Book] as

$$^k\text{UIM}(w; 0, 0, 0, \lambda) = \text{UIM}(w; 0, 0, 0, \lambda) \otimes ^k\mathbf{B}$$

where $^k\mathbf{B}$ is the k th-order $\rightarrow$ *geodesic matrix*, whose elements are equal to 1 only for vertices v_i and v_j at topological distance k , and zero otherwise. Then, the elements of these matrices are

$$[{}^k\mathbf{UIM}(w; 0, 0, 0, \lambda)]_{ij} = \begin{cases} w_j^\lambda & \text{if } i \neq j \wedge d_{ij} = k \\ 0 & \text{otherwise} \end{cases}$$

Each i th row of this matrix contains information about properties of those vertices located at distance k from the i th vertex; therefore, the i th row sum is a local vertex invariant defined as the sum of the properties of vertices at distance k from the i th vertex:

$$VS_i({}^k\mathbf{UIM}(w; 0, 0, 0, \lambda)) = \sum_{j=1}^A w_j^\lambda \quad \text{if } i \neq j \wedge d_{ij} = k$$

For $\lambda = 1$ and by choosing the vertex degree δ as the weighting scheme, the row sum of ${}^k\mathbf{UIM}(\delta)$ reduces to the i - k element of the $\rightarrow$ *branching layer matrix*. Moreover, still for $\lambda = 1$, the row sum of the first-order unsymmetrical interaction matrix ${}^1\mathbf{UIM}(m)$ derived from relative atomic weights is the $\rightarrow$ *Madan chemical degree*.

Example W32

Unsymmetrical ${}^k\mathbf{UIM}$ and symmetric ${}^k\mathbf{IM}$ interaction geodesic matrices of 5-methyl-1,3,4-oxathiazol-2-one ($k = 1, 2$). VS_i is the matrix row sum and CS_j the matrix column sum. Calculation was performed setting $\lambda = 1$ and by means of the following relative atomic masses for carbon, oxygen, nitrogen, and sulfur, respectively, 1, 1.332, 1.166, and 2.670. Note that ${}^1\mathbf{UIM}(m; 0, 0, 0, 1)$ coincides with the $\rightarrow$ *chemical adjacency matrix* of the same molecule (see Example W12).

${}^1\mathbf{UIM}(m; 0, 0, 0, 1)$								${}^1\mathbf{IM}(m; 0, 0, 0, 1)$									
	1	2	3	4	5	6	7	VS_i		1	2	3	4	5	6	7	VS_i
1	0	1	0	0	1	0	0	2	1	0	1.332	0	0	1.332	0	0	2.664
2	1.332	0	2.670	0	0	1.332	0	5.334	2	1.332	0	2.670	0	0	1.332	0	4.002
3	0	1	0	1.166	0	0	0	2.166	3	0	2.670	0	3.113	0	0	0	5.783
4	0	0	2.670	0	1	0	0	3.670	4	0	0	3.113	0	1.166	0	0	4.279
5	1.332	0	0	1.166	0	0	1	3.498	5	1.332	0	0	1.166	0	0	1	3.498
6	0	1	0	0	0	0	0	1	6	0	1.332	0	0	0	0	0	1.332
7	0	0	0	0	1	0	0	1	7	0	0	0	0	1	0	0	1
CS_j	2.664	3	5.340	2.332	3	1.332	1	18.67	CS_j	2.664	5.334	5.783	4.279	3.498	1.332	1	23.89

$$ATS_1(m) = 11.95$$

${}^2\mathbf{UIM}(m; 0, 0, 0, 1)$									${}^2\mathbf{IM}(m; 0, 0, 0, 1)$								
	1	2	3	4	5	6	7	VS_i		1	2	3	4	5	6	7	VS_i
1	0	0	2.670	1.166	0	1.332	1	6.168	1	0	0	3.556	1.553	0	1.774	1.332	8.216
2	0	0	0	1.166	1	0	0	2.166	2	0	0	0	1.166	1	0	0	2.166
3	1.332	0	0	0	1	1.332	0	3.664	3	3.556	0	0	0	2.670	3.556	0	9.783
4	1.332	1	0	0	0	0	1	3.332	4	1.553	1.166	0	0	0	0	1.166	3.885
5	0	1	2.670	0	0	0	0	3.670	5	0	1	2.670	0	0	0	0	3.670
6	1.332	0	2.670	0	0	0	0	4.002	6	1.774	0	3.556	0	0	0	0	5.331
7	1.332	0	0	1.166	0	0	0	2.498	7	1.332	0	0	1.166	0	0	0	2.498
CS_j	5.328	2	8.01	3.498	2	2.664	2	25.50	CS_j	8.216	2.166	9.783	3.885	3.670	5.331	2.498	35.55

$$ATS_2(m) = 17.77$$

It should be noted that both interaction and perturbation matrices can be transformed into diagonal matrices by setting $\lambda = 0$ and the γ parameter at some large positive value (e.g., 50) in the third smoothing function f_3 . In effect, it can be shown that choosing $\gamma = 50$, the distance function would be $2^{50 \times 1}$ and then the matrix off-diagonal elements would result around 10^{-15} , which is enough small to be approximated to zero. Note that with an analogous parameter setting, but with $\alpha = 1$ and $\beta = 0$, the identity matrix is obtained.

From interaction matrices, perturbation matrices, interaction geodesic matrices, perturbation geodesic matrices, and atom-type autocorrelation matrices, several graph invariants can be calculated, such as $\rightarrow$ *spectral indices*, $\rightarrow$ *Wiener-type indices*, $\rightarrow$ *Balaban-like indices*, $\rightarrow$ *characteristic polynomial* and related descriptors, $\rightarrow$ *determinant-based descriptors*, and all those descriptors defined in terms of $\rightarrow$ *local vertex invariants* that, in this case, are calculated as the matrix row sums.

Some interaction and perturbation matrices for specific parameter combinations are listed in Tables W5 and W6.

Table W5 Some interaction graph matrices obtained by varying α , β , γ , and λ parameters and distance smoothing functions f ; w is the vertex weighting scheme.

M	α	β	γ	λ	f	w	Matrix
IM	0	–	–1	0	1	1	Distance matrix
IM	1	1	–1	0	1	w	Augmented distance matrix
IM	0	–	<0	0	1	1	Distance distribution moment matrices
IM	0	–	>0	0	1	1	Reciprocal distance matrices
IM	0	–	1	0	1	1	Harary matrix
IM	0	–	2	0	1	1	Reciprocal square distance matrix
IM	0	–	–1	1	1	δ	Vertex-distance–vertex-degree matrices
IM	0	–	0	–1/2	–	$val(w)$	XI matrix
IM	0	–	γ	0	4	1	Generalized molecular-graph matrix
^{k} IM	0	0	0	1	–	w	k th-order autocorrelation matrices
^{1} IM	0	0	0	0	–	1	Adjacency matrix
^{1} IM	0	0	0	–1/2	–	δ	χ matrix
^{1} IM	0	0	0	1	–	δ	Edge-Zagreb matrix
^{1} IM	0	0	0	–1	–	δ	Modified edge-Zagreb matrix
^{1} IM	0	0	0	–1/2	–	σ	Distance-sum–connectivity matrix

Table W6 Some unsymmetrical and diagonal interaction matrices obtained by varying α , β , γ , and λ parameters and distance smoothing functions f , which coincide with other well-known graph-theoretical matrices; w is the vertex weighting scheme.

M	α	β	γ	λ	f	w	Matrix
UIM	1	1	1	1	3	δ	Augmented vertex-degree matrix
^{k} UIM	0	0	0	1	–	δ	k th-order valence shell matrix
^{1} UIM	0	0	0	1	–	m_j/m_C	Chemical adjacency matrix
^{1} UIM	0	0	0	1	–	δ	Additive adjacency matrix

(Continued)

Table W6 (Continued)

M	α	β	γ	λ	f	w	Matrix
¹ UIM	0	0	0	1	–	δ^c	Additive chemical adjacency matrix
¹ UIM	0	0	0	–1	–	δ	Random walk Markov matrix
IM	1	1	High	0	3	w	Vertex-weighted diagonal matrix
IM	1	1	High	0	3	δ	Vertex degree matrix
IM	1	β	High	0	3	δ	Generalized vertex degree matrix
IM	1	2	High	0	3	δ	Vertex Zagreb matrix
IM	1	–2	High	0	3	δ	Modified vertex Zagreb matrix

- **weighted mean of k th power** → statistical indices (⊙ indices of central tendency)
- **weighted modified Wiener index** → Szeged index
- **weighted path counts** → path counts
- **weighted self-returning walk counts** → self-returning walk counts
- **weighted Tversky similarity coefficient** → similarity/diversity
- **weighted variance** → statistical indices (⊙ indices of dispersion)
- **weighted vertex degree** $\equiv$ *valency of a vertex* → vertex degree
- **weighted walk counts** → walk count
- **weighted walk degrees** → walk matrices
- **weighted walk numbers** → walk matrices

■ weighting schemes

Weighting schemes for graph vertices and edges are used to encode in a numerical form information about atom and bond properties that are not represented in a simple molecular graph.

Compounds containing heteroatoms and/or multiple bonds can be represented as a vertex- and/or edge-weighted molecular graph. The elements of the vertex and edge weight sets are computed by a defined weighting scheme w [Ivanciuc, Ivanciuc *et al.*, 2000d].

Graph invariants derived from weighted molecular graphs are sometimes referred to as topochemical indices to highlight the fact that they contain some chemical information about the molecule.

Weighting schemes are used to generate several → *weighted matrices*, from which topochemical indices are derived, and to directly calculate → *WHIM descriptors*, → *autocorrelation descriptors*, → *GETAWAY descriptors*, → *RDF descriptors*, → *3D-MoRSE descriptors*, → *BCUT descriptors*, → *TOMOCOMD descriptors*, and many others.

The most common weighting schemes for atoms are → *atomic properties* such as atomic mass, atomic electronegativity, atomic polarizabilities, and → *local vertex invariants* derived from molecular graphs. Weighting schemes to characterize bonds, that is, the molecular graph edges, can be based on quantities directly characterizing bonds, such as bond distances, bond orders, bond dipoles, or quantities derived from the weights associated with the two atoms forming a bond.

From vertex-weighted molecular graphs, edge weights can be derived from some combination of the weights of the two vertices incident to the edge:

$$w_{ij} = f(w_i, w_j)$$

where w_{ij} is the edge weight, w_i and w_j are the vertex weights, and f a generic function. The two most common functions to calculate edge weights are

$$w_{ij} = (w_i \cdot w_j)^\lambda \quad w_{ij} = \frac{w_i + w_j}{2}$$

where λ is a real parameter, usually equal to $-1/2$. The edge weighting scheme used in the definition of $\rightarrow$ *extended adjacency matrices* is

$$w_{ij} = \frac{1}{2} \cdot \left(\frac{w_i}{w_j} + \frac{w_j}{w_i} \right)$$

Moreover, to calculate the bilinear indices, which are among the $\rightarrow$ *TOMOCOMD descriptors*, the following edge weighting scheme was proposed to characterize each bond:

$$w_{ij} = \frac{w_i}{\delta_i^b} + \frac{w_j}{\delta_j^b}$$

where δ_i^b and δ_j^b are the $\rightarrow$ *bond vertex degrees* of each atom.

Derived from the $\rightarrow$ *Barysz distance matrix*, a general weighting scheme was proposed by Ivanciuc [Ivanciuc, Ivanciuc *et al.*, 1999a; Ivanciuc, 2000a, 2000i] in terms of the $\rightarrow$ *conventional bond order* π^* and any atomic property P_i . The vertex weight w_i associated with the vertex v_i was defined as

$$w_i = 1 - \frac{P_C}{P_i}$$

and the edge weight w_{ij} associated with the edge between vertices v_i and v_j was defined as

$$w_{ij}(P) = \frac{1}{\pi_{ij}^*} \cdot \frac{P_C^2}{P_i \cdot P_j}$$

where P_i is the atomic property of vertex v_i , P_j is the atomic property of vertex v_j , and P_C is the atomic property of the carbon atom.

Several **Ivanciuc weighting schemes** are obtained depending on the atomic property used to weight graph vertices. The **Z weighting scheme** is obtained when the atomic property P is the atomic number (i.e., the Barysz weighting scheme), the **A weighting scheme** when the property P is the atomic mass, the **P weighting scheme** when the property P is the atomic polarizability, the **E weighting scheme** when the property P is the atomic electronegativity, the **R weighting scheme** when the property P is the atomic radius. The **AH weighting scheme** is based on vertex and edge parameters corrected for the contributions from attached hydrogen atoms [Ivanciuc, 2000i]. According to this scheme, the vertex weight is defined as

$$w_i(AH) = 1 - \frac{m_C}{m_i + m_H \cdot n_H} = 1 - \frac{12.011}{m_i + 1.0079 \times n_H}$$

and the edge weight as

$$w_{ij}(AH) = \frac{1}{\pi_{ij}^*} \cdot \frac{m_C^2}{(m_i + m_H \cdot n_{Hi}) \cdot (m_j + m_H \cdot n_{Hj})} = \frac{1}{\pi_{ij}^*} \cdot \frac{144.26}{(m_i + 1.0079 \times n_{Hi}) \cdot (m_j + 1.0079 \times n_{Hj})}$$

where $m_C = 12.011$ is the atomic mass of carbon, $m_H = 1.0079$ the atomic mass of hydrogen, n_{H_i} the number of hydrogen atoms bonded to vertex v_i , and n_{H_j} the number of hydrogen atoms bonded to vertex v_j . Values of atomic and edge weights are provided in Tables W7 and W8, respectively.

In the **X weighting scheme**, the relative $\rightarrow$ atom electronegativity X_i is the atomic property P_i used to derive vertex and edge weights [Ivanciuc, Ivanciuc *et al.*, 1998b, 2000d]. The atomic electronegativity values S_i are derived from electronegativities recalculated by Sanderson on the Pauling scale with F, Na, and H atoms having values equal to 4.00, 0.56, and 2.592, respectively, by using a biparametric equation:

$$S_i = 1.1032 - 0.0204 \cdot Z_i + 0.4121 \cdot G_i$$

where Z denotes atomic numbers and G the group number of the Periodic System short form, respectively. Then, relative atomic electronegativities X_i are calculated by dividing the S_i values by the calculated value for carbon atom $S_C = 2.629$ as

$$X_i = 0.4196 - 0.0078 \cdot Z_i + 0.1567 \cdot G_i$$

In this way, the X value for carbon atom is equal to 1.

Finally, in the **Y weighting scheme**, relative covalent radius Y_i is used as the atomic property P_i . Relative covalent radii to the carbon atom are calculated by using a biparametric equation based on atomic number Z and group number G [Ivanciuc, Ivanciuc *et al.*, 1998b, 2000d]:

$$Y_i = 1.1191 + 0.0160 \cdot Z_i - 0.0537 \cdot G_i$$

The atomic radii Y_i are relative to a carbon covalent radius equal to 87.126 picometers ($Y_C = 1$). Relative atomic electronegativities X_i and covalent radii Y_i were also used as atomic weights to account for heteroatoms in the definition of the $\rightarrow$ Balaban modified distance connectivity indices, J^X and J^Y , respectively.

Based on these weighting schemes, several $\rightarrow$ Ivanciuc weighted distance matrices were proposed to represent vertex- and edge-weighted molecular graphs [Ivanciuc, 2000i].

Table W7 Vertex weightings derived from the atomic number Z , relative covalent radius Y , and relative electronegativity X for some atoms according to the Ivanciuc weighting schemes.

Atom	Z	Z-weight	X	X-weight	Y	Y-weight
B	5	-0.200	0.851	-0.175	1.038	0.037
C	6	0.000	1.000	0.000	1.000	0.000
N	7	0.143	1.149	0.130	0.963	-0.038
O	8	0.250	1.297	0.229	0.925	-0.081
Si	14	0.571	0.937	-0.067	1.128	0.113
P	15	0.600	1.086	0.079	1.091	0.083
S	16	0.625	1.235	0.190	1.053	0.050
F	9	0.333	1.446	0.308	0.887	-0.127
Cl	17	0.647	1.384	0.277	1.015	0.015
Br	35	0.829	1.244	0.196	1.303	0.233
I	53	0.887	1.103	0.093	1.591	0.371

Table W8 Bond weights defined according to Z, X, Y, A, P, R, and E weighting schemes.

Bond	w(Z)	X	Y	A	P	R	E
C–C	1.000	1.000	1.000	1.000	1.000	1.000	1.000
C=C	0.500	0.500	0.500	0.500	0.500	0.500	0.500
C≡C	0.333	0.333	0.333	0.333	0.333	0.333	0.333
C≈C	0.667	0.667	0.667	0.667	0.667	0.667	0.667
C–N	0.857	0.870	1.038	0.857	1.600	1.175	0.817
C=N	0.429	0.435	0.519	0.429	0.800	0.587	0.409
C≡N	0.286	0.290	0.346	0.286	0.533	0.392	0.272
C≈N	0.571	0.580	0.692	0.572	1.067	0.783	0.545
C–O	0.750	0.771	1.081	0.751	2.195	1.301	0.704
C=O	0.375	0.386	0.541	0.375	1.097	0.651	0.352
C–Si	0.429	1.067	0.887	0.428	0.327	0.691	1.364
C–P	0.400	0.921	0.917	0.388	0.485	0.786	1.149
C–S	0.375	0.810	0.950	0.375	0.607	0.846	1.024
C=S	0.188	0.405	0.475	0.187	0.303	0.423	0.512
C≈S	0.250	0.540	0.633	0.250	0.405	0.564	0.683
C–F	0.667	0.692	1.127	0.632	3.160	1.476	0.603
C–Cl	0.353	0.723	0.985	0.339	0.807	0.931	0.904
C–Br	0.171	0.804	0.767	0.150	0.577	0.834	0.996
C–I	0.113	0.907	0.629	0.095	0.374	0.720	1.123
N–N	0.735	0.757	1.078	0.735	2.560	1.380	0.668
N=N	0.367	0.379	0.539	0.368	1.280	0.690	0.334
N≈N	0.490	0.505	0.719	0.490	1.707	0.920	0.445
N–O	0.643	0.671	1.123	0.644	3.511	1.528	0.576
N=O	0.321	0.336	0.561	0.322	1.756	0.764	0.288
N≈O	0.423	0.447	0.748	0.429	2.341	1.019	0.384
O–S	0.281	0.624	1.027	0.281	1.332	1.101	0.721
O=S	0.141	0.312	0.513	0.141	0.666	0.550	0.361

➤ **Werckwerth solute descriptors** → Linear Solvation Energy Relationships

■ **WHIM descriptors** (≡ *Weighted Holistic Invariant Molecular descriptors*)

WHIM descriptors are molecular descriptors based on → *statistical indices* calculated on the projections of the atoms along principal axes [Todeschini, Lasagni *et al.*, 1994; Todeschini and Gramatica, 1997a].

WHIM descriptors are built in such a way as to capture relevant molecular 3D information regarding molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. The algorithm consists in performing a → *Principal Components Analysis* on the centered → *Cartesian coordinates* of a molecule (centered → *molecular matrix*) by using a weighted covariance matrix obtained from different → *weighting schemes* for the atoms:

$$s_{jk} = \frac{\sum_{i=1}^A w_i (q_{ij} - \bar{q}_j)(q_{ik} - \bar{q}_k)}{\sum_{i=1}^A w_i}$$

where s_{jk} is the weighted covariance between the j th and k th atomic coordinates, A is the number of atoms, w_i the weight of the i th atom, q_{ij} and q_{ik} represent the j th and k th coordinate ($j, k = x, y, z$) of the i th atom respectively, and $\bar{q}$ the corresponding average value.

Six different weighting schemes, providing **WHIM weighted covariance matrices (WWC matrices)**, are proposed: (1) the unweighted case u ($w_i = 1$ for all the atoms; (2) atomic mass m ; (3) the $\rightarrow$ *van der Waals volume* v ; (4) the Sanderson $\rightarrow$ *atomic electronegativity* e , (5) the $\rightarrow$ *atomic polarizability* p ; and (6) the $\rightarrow$ *electrotopological state indices* of Kier and Hall S . All the weights are also scaled with respect to the carbon atom, and their values can be found in $\rightarrow$ *atomic properties* (Table A3); moreover, as all the weights must be positive, the electrotopological indices are scaled thus:

$$S'_i = S_i + 7 \quad S'_i > 0$$

In this case, only the nonhydrogen atoms are considered, and the atomic electrotopological charge of each atom depends on its atom neighbor.

Depending on the kind of weighting scheme, different covariance matrices and different principal axes (i.e., principal components t_m) are obtained. For example, using atomic masses as the weighting scheme, the directions of the three principal axes are the directions of the $\rightarrow$ *principal inertia axes*. Thus, the WHIM approach can be viewed as a generalized search for the principal axes with respect to a defined atomic property ($\rightarrow$ *weighting schemes*).

Moreover, based on the WHIM approach applied to $\rightarrow$ *molecular interaction fields*, $\rightarrow$ *G-WHIM descriptors* were derived [Todeschini, Moro *et al.*, 1997].

For each weighting scheme, a set of statistical indices is calculated on the atoms projected onto each principal component t_m ($m = 1, 2, 3$), that is, the scores, as described below.

The invariance to translation of the calculated parameters is due to the centering of the atomic coordinates and the invariance to rotation is due to the uniqueness of the PCA solution.

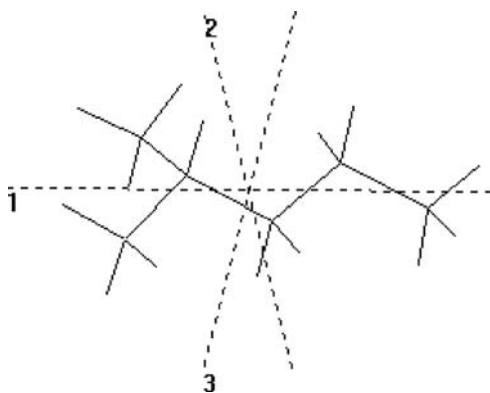


Figure W1 Principal axes calculated for 2-methylpentane, using the weighting scheme based on atomic masses.

• directional WHIM descriptors

These are molecular descriptors calculated as univariate statistical indices on the scores of each individual principal component t_m ($m = 1, 2, 3$).

The first group of descriptors are the **directional WHIM size descriptors** (or **d-WSIZ indices**) defined as the $\rightarrow$ *eigenvalues* λ_1 , λ_2 , and λ_3 of the weighted covariance matrix of the molecule atomic coordinates; they account for the molecular size along each principal direction.

The second group is constituted by the **directional WHIM shape** descriptors (or **d-WSHA indices**) ϑ_1 , ϑ_2 , and ϑ_3 , calculated as eigenvalue ratios and related to molecular shape:

$$\vartheta_m = \frac{\lambda_m}{\sum_m \lambda_m} \quad m = 1, 2, 3$$

Because of the closure condition ($\vartheta_1 + \vartheta_2 + \vartheta_3 = 1$), only two of them are independent.

The third group of descriptors consists of the **directional WHIM symmetry** descriptors (or **d-WSYM indices**) γ_1 , γ_2 , and γ_3 , calculated as $\rightarrow$ *mean information content* on the symmetry along each component with respect to the center of the scores:

$$\gamma'_m = - \left[\frac{n_s}{n} \cdot \log_2 \frac{n_s}{n} + n_a \cdot \left(\frac{1}{n} \cdot \log_2 \frac{1}{n} \right) \right] \quad \gamma_m = \frac{1}{1 + \gamma'_m} \quad 0 < \gamma_m \leq 1$$

where n_s , n_a , and n denote the number of central symmetric atoms (along the m th component), the number of unsymmetrical atoms, and the total number of atoms of the molecule, respectively.

Finally, the fourth group of descriptors consists of the inverse of the *kurtosis* κ_1 , κ_2 , and κ_3 , calculated from the fourth-order moments of the scores t_m , which are related to the atom distribution and density around the origin and along the principal axes:

$$\kappa_m = \frac{\sum_i t_i^4}{\lambda^2 \cdot n} \quad \eta_m = \frac{1}{\kappa_m} \quad m = 1, 2, 3$$

To avoid problems related to infinite (or very high) kurtosis values, obtained when along a principal axis all the atoms are projected in the center (or near the center, i.e., leptokurtic distribution), the inverse of the kurtosis is used.

Low values of the kurtosis are obtained when the data points (i.e., the atom projections) assume opposite values ($-t$ and t) with respect to the center of the scores. When an increasing number of data values are within the extreme values $\pm t$ along a principal axis, the kurtosis value increases (i.e., $\kappa = 1.8$ for a uniform distribution of points, $\kappa = 3.0$ for a normal distribution). When the kurtosis value tends to be infinite, the corresponding η value tends to zero.

Thus, the group of descriptors η_m can be related to the quantity of unfilled space per projected atom and have been called **directional WHIM density** descriptors (or **d-WDEN indices**, or *WHIM emptiness*): the greater the η_m values the more the projected unfilled space. The η_m descriptors are used in place of the kurtosis descriptors κ_m (previously proposed).

Therefore, for each weighting scheme w , 11 molecular directional WHIM descriptors (ϑ_3 is excluded) are obtained:

$$\{\lambda_1 \ \lambda_2 \ \lambda_3 \ \vartheta_1 \ \vartheta_2 \ \gamma_1 \ \gamma_2 \ \gamma_3 \ \eta_1 \ \eta_2 \ \eta_3\}_w$$

thus resulting in a total of 66 directional WHIM descriptors. For planar compounds, λ_3 , γ_3 , and η_3 are always equal to zero.

• global WHIM descriptors

These are molecular descriptors directly calculated as a combination of the directional WHIM descriptors, thus contemporarily accounting for the variation of molecular properties along with the three principal directions in the molecule. Thus, for nondirectional WHIM descriptors any information individually related to each principal axis disappears and the description is related only to a global view of the molecule.

In many cases, size descriptors can play, in modeling, a significant role independently of the measured directions, allowing more simple models. Thus, in view of the importance of this quantity, a group of descriptors of the total dimension of a molecule is considered in three different ways, based on the three eigenvalues defined above:

$$T = \lambda_1 + \lambda_2 + \lambda_3$$

$$A = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2 + \lambda_3$$

$$V = \prod_{m=1}^3 (1 + \lambda_m) - 1 = T + A + \lambda_1\lambda_2\lambda_3$$

where T and A are, respectively, related to linear and quadratic contributions, to the total molecular size. V is the complete expansion, also containing the third-order term. These molecular descriptors are called **WHIM size** descriptors (or **WSIZ indices**).

The shape of the molecule is represented by the **WHIM shape** (or **WSHA index**) defined as

$$K = \frac{\sum_m \left| \frac{\lambda_m}{\sum_m \lambda_m} - \frac{1}{3} \right|}{4/3} \quad m = 1, 2, 3 \quad 0 \leq K \leq 1$$

The term $4/3$ is the maximum value of the numerator term and is used to scale K between 0 and 1. This expression has also a more general meaning, K being the $\rightarrow$ *multivariate K correlation index* used to evaluate global correlation in data [Todeschini, 1997; Gramatica, 2006].

For example, for an ideal straight molecule both λ_2 and λ_3 are equal to zero and $K=1$; for an ideal spherical molecule all three eigenvalues are equal to $1/3$ and $K=0$. For all planar molecules, the third eigenvalue λ_3 is 0, there being no variance out of the molecular plane, and K ranges between 0.5 and 1, depending on the molecule linearity.

The K shape term definitely substitutes the acentric factor, it being more general than the previously proposed *acentric factor* ω .

The total molecular symmetry is accounted for by the **WHIM symmetry** (or **WSYM index**) defined as

$$G = (\gamma_1 \cdot \gamma_2 \cdot \gamma_3)^{1/3}$$

G is the geometric mean of the directional symmetries and is equal to 1 when the molecule shows a central symmetry along each axis and tends to be zero when there is a loss of symmetry along at least one axis. Different symmetry values are obtained only when unitary, mass, and electrotopological weights are used; for this reason, only three kinds of symmetry parameters are retained: G_u , G_m , and G_s .

The total density of the atoms within a molecule is accounted for by the **WHIM density** (or **WDEN index**) defined by the following expression:

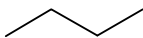
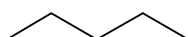
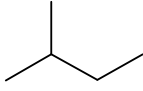
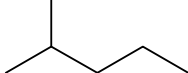
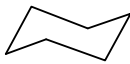
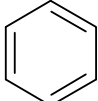
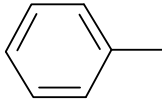
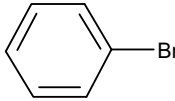
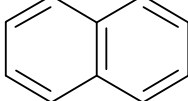
$$D = \eta_1 + \eta_2 + \eta_3$$

The molecular descriptors defined above are generalized molecular properties within each weighting scheme. The global WHIMs are five for each of the six proposed weighting schemes w :

$$\{T, A, V, K, D\}_w$$

plus the symmetry indices G_u , G_m , and G_s , giving a total number of 33 descriptors.

Table W9 Some WHIM descriptor values for simple organic molecules; the scaled atomic mass is the weighting scheme for atoms (see Table A3).

Molecule	λ_{1m}	λ_{2m}	λ_{3m}	T_m	A_m	V_m	K_m	D_m	G_m
	0.714	0.110	0.110	0.934	0.169	1.111	0.647	0.148	1.000
	1.211	0.287	0.112	1.610	0.515	2.165	0.629	0.140	0.608
	2.203	0.269	0.113	2.585	0.873	3.526	0.778	0.157	1.000
	3.366	0.309	0.114	3.789	1.458	5.366	0.833	0.172	0.581
	0.943	0.943	0.161	2.046	1.192	3.380	0.382	0.149	0.208
	1.870	0.785	0.170	2.825	1.919	4.994	0.493	0.161	0.197
	3.170	0.696	0.186	4.053	2.927	7.390	0.673	0.177	0.213
	1.199	1.199	0.178	2.575	1.864	4.695	0.396	0.169	1.000
	1.139	1.139	0.000	2.278	1.297	3.575	0.500	0.108	1.000
	2.180	0.983	0.018	3.181	2.200	5.419	0.528	0.126	0.511
	5.761	0.568	0.000	6.329	3.270	9.599	0.865	0.812	0.569
	3.195	1.265	0.000	4.460	4.041	8.500	0.575	0.169	1.000
	0.604	0.123	0.000	0.727	0.074	0.801	0.746	0.081	1.000
	1.241	0.229	0.039	1.510	0.342	1.863	0.733	0.099	0.386
	0.543	0.000	0.000	0.543	0.000	0.543	1.000	0.026	1.000

WHIM descriptors have been used to model toxicological indices [Todeschini, Bettiol *et al.*, 1996; Todeschini, Vighi *et al.*, 1996, 1997; Shapiro and Guggenheim, 1998a], several physico-chemical properties of polychlorobiphenyls [Gramatica, Navas *et al.*, 1998] and polycyclic aromatic hydrocarbons [Todeschini, Gramatica *et al.*, 1995], hydroxyl radical rate constants [Gramatica, Consonni *et al.*, 1999], and soil sorption partition coefficients [Gramatica, Corradi *et al.*, 2000].

Table W10 Some WHIM descriptors for phenetylamines data set (Appendix C – Set 2). The scaled atomic mass is the weighting scheme for atoms (see Table A3); u refers to unitary weighting scheme.

Molecule	X	Y	L _{1u}	L _{2u}	L _{3u}	L _{1m}	L _{2m}	L _{3m}	T _u	T _m	K _u	K _m	D _u	D _m
1	H	H	8.038	1.465	0.909	4.678	2.735	0.472	10.411	7.885	0.658	0.410	0.501	0.602
2	H	F	8.085	1.716	0.665	6.501	2.408	0.572	10.466	9.482	0.659	0.528	0.426	0.804
3	H	Cl	8.225	1.719	0.677	8.447	2.264	0.543	10.621	11.255	0.662	0.626	0.420	0.788
4	H	Br	8.301	1.725	0.674	12.468	1.934	0.471	10.699	14.873	0.664	0.757	0.417	0.846
5	H	I	8.308	1.700	0.719	16.081	1.716	0.403	10.728	18.200	0.662	0.825	0.416	0.988
6	H	Me	10.551	1.611	0.691	6.391	2.350	0.587	12.853	9.327	0.731	0.528	0.425	0.521
7	F	H	8.040	1.751	0.657	5.843	2.402	0.789	10.448	9.034	0.654	0.470	0.426	0.758
8	Cl	H	8.104	1.794	0.677	7.170	2.264	0.915	10.575	10.349	0.650	0.539	0.423	0.704
9	Br	H	8.129	1.812	0.687	10.116	1.946	0.935	10.628	12.997	0.647	0.667	0.421	0.708
10	I	H	8.155	1.836	0.697	12.912	1.701	0.882	10.688	15.494	0.645	0.750	0.417	0.824
11	Me	H	9.335	2.053	0.718	5.789	2.472	0.776	12.106	9.038	0.657	0.461	0.426	0.547
12	Cl	F	8.138	1.820	0.676	7.885	2.199	1.129	10.634	11.214	0.648	0.555	0.411	0.887
13	Br	F	8.163	1.856	0.671	9.935	1.991	1.301	10.690	13.226	0.645	0.627	0.405	0.865
14	Me	F	8.989	2.232	0.684	6.670	2.171	0.877	11.905	9.719	0.633	0.529	0.436	0.591
15	Cl	Cl	8.278	1.819	0.693	9.447	2.121	1.098	10.790	12.667	0.651	0.619	0.405	0.887
16	Br	Cl	8.320	1.871	0.672	11.006	2.022	1.334	10.863	14.362	0.649	0.650	0.396	0.863
17	Me	Cl	9.136	2.260	0.669	8.222	2.057	0.840	12.064	11.119	0.636	0.609	0.431	0.625
18	Cl	Br	8.357	1.836	0.679	12.877	1.844	1.008	10.873	15.728	0.653	0.728	0.399	0.916
19	Br	Br	8.381	1.880	0.669	13.632	1.943	1.269	10.930	16.844	0.650	0.714	0.392	0.948
20	Me	Br	9.195	2.220	0.709	11.412	1.763	0.721	12.124	13.895	0.638	0.732	0.434	0.657
21	Me	Me	10.698	2.108	0.741	6.843	2.164	0.877	13.547	9.884	0.685	0.539	0.437	0.415
22	Br	Me	10.553	1.704	0.744	8.461	2.377	1.226	13.001	12.065	0.718	0.552	0.410	0.684

 Additional references are listed in the thematic bibliography (see Introduction).

- **WHIM density** → WHIM descriptors (⊙ global WHIM descriptors)
- **WHIM shape** → WHIM descriptors (⊙ global WHIM descriptors)
- **WHIM size** → WHIM descriptors (⊙ global WHIM descriptors)
- **WHIM symmetry** → WHIM descriptors (⊙ global WHIM descriptors)
- **WHIM weighted covariance matrices** → WHIM descriptors
- **Wiberg index** $\equiv$ *bond index* → quantum-chemical descriptors
- **Wiener difference matrix** → Wiener matrix
- **Wiener–Hosoya index** → Wiener index

■ **Wiener index** (W) ($\equiv$ *Wiener number*, *Wiener path number*)

One of the first $\rightarrow$ *topological indices* introduced by H. Wiener in 1947 as a bond-additive index in which each bond gives a contribution equal to the product of the number of vertices on each side of the bond [Wiener, 1947a, 1947b, 1947c] and called it at the beginning *path number*. For acyclic graphs, the Wiener index is then calculated as

$$W = \sum_{b=1}^B N_{i,b} \cdot N_{j,b}$$

where $N_{i,b}$ and $N_{j,b}$ are the number of vertices on each side of the bond b , including vertices i and j , respectively, and B is the total number of graph edges.

Usually, the Wiener index is defined and calculated as the sum of all topological distances in the $\rightarrow$ *H-depleted molecular graph* [Hosoya, 1971]. It is obtained from the $\rightarrow$ *distance matrix* $\mathbf{D}$ as

$$W \equiv Wi(\mathbf{D}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij} = \sum_{k=1}^D {}^k f \cdot k$$

where Wi is the $\rightarrow$ *Wiener operator*, A is the number of graph vertices, d_{ij} is the $\rightarrow$ *topological distance* between vertices v_i and v_j , D is the maximum $\rightarrow$ *topological distance* in the graph ($\rightarrow$ *topological diameter*), and ${}^k f$ is the number of distances in the graph equal to k , that is, the k th order $\rightarrow$ *graph distance count*. Because $\mathbf{D}$ is a symmetric matrix, the factor $1/2$ avoids counting the distances twice. Note that the Wiener index was also independently proposed by Harary in the context of sociometry, with the name *total status of a graph* and denoted by $ts(\mathcal{G})$, as [Harary, 1959]:

$$ts(\mathcal{G}) = \frac{1}{2} \cdot \sum_{i=1}^A \sigma_i$$

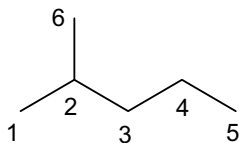
where σ_i is the sum of distances from the i th vertex to all other vertices in the graph (i.e., the $\rightarrow$ *distance degree*).

By exploiting this last formula for the calculation of the Wiener index, the **conditional Wiener index** was proposed [Rodriguez and Sigarreta, 2005], considering in the summation only those vertices satisfying a predefined condition, such as $\delta_i \geq \alpha$, where α is a parameter and δ the $\rightarrow$ *vertex degree*.

Interesting properties of the Wiener index and its connection to other graph invariants are reported in Refs [Polanski and Bonchev, 1990; Gutman, Yeh *et al.*, 1993; Plavšić, Nikolić *et al.*, 1993a; Gutman, 1994a, 1997, 2002b; Marković, Gutman *et al.*, 1995; Gutman and Klavžar, 1998; Bonchev and Klein, 2002; Vukicević and Trinajstić, 2004; Walikar, Shigehalli *et al.*, 2004; Zhou and Gutman, 2004b; Yan and Yeh, 2006]. Other papers and reviews about the Wiener index are listed in the thematic bibliography.

Example W35

Distance matrix **D** and Wiener index for 2-methylpentane; N_i and N_j represent the number of vertices on the side of the i th vertex and j th vertex, respectively.

**Wiener definition**

Vertex pair (i, j)	N_i	N_j	$N_i \cdot N_j$
(1, 2)	1	5	5
(2, 3)	3	3	9
(2, 6)	5	1	5
(3, 4)	4	2	8
(4, 5)	5	1	5

$$W = N_1 \cdot N_2 + N_2 \cdot N_3 + N_2 \cdot N_6 + N_3 \cdot N_4 + N_4 \cdot N_5 = 5 + 9 + 5 + 8 + 5 = 32$$

Hosoya definition

Atom	1	2	3	4	5	6
1	0	1	2	3	4	2
2	1	0	1	2	3	1
3	2	1	0	1	2	2
4	3	2	1	0	1	3
5	4	3	2	1	0	4
6	2	1	2	3	4	0

D =

$$W \equiv Wi(\mathbf{D}) = \frac{1}{2} \cdot \sum_{i=1}^6 \sum_{j=1}^6 d_{ij} = 32$$

The **mean Wiener index** is defined as

$$\overline{W} = \frac{2 \cdot W}{A \cdot (A-1)}$$

Twice the mean Wiener index was proposed as a molecular descriptor called $\rightarrow$ *compactness*, while twice the Wiener index is the $\rightarrow$ *Rouvray index*.

For acyclic graphs the Wiener index, according to its original definition, can also be obtained from the $\rightarrow$ *Wiener matrix* **W** by summing all of the entries corresponding to pairs of adjacent vertices above the main diagonal:

$$W = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{W} \otimes \mathbf{A}]_{ij}$$

where $\mathbf{W}_e = [\mathbf{W} \otimes \mathbf{A}]$ is the edge-Wiener matrix obtained by the $\rightarrow$ *Hadamard matrix product* of the Wiener matrix **W** and the $\rightarrow$ *adjacency matrix* **A**.

For acyclic graphs, the Wiener index can also be obtained from the symmetric $\rightarrow$ *edge-Cluj-distance matrix* **CJD_e**, $\rightarrow$ *edge-Cluj-detour matrix* **CJΔ_e** and $\rightarrow$ *edge-Szeged matrix* **SZ_e** by applying the $\rightarrow$ *Wiener operator* or from the corresponding unsymmetrical matrices by applying the $\rightarrow$ *orthogonal Wiener operator*.

The $\rightarrow$ *Szeged index* was proposed as generalization of the original Wiener index, valid both for acyclic and cyclic graphs [Khadikar, Deshpande *et al.*, 1995; Gutman, Khadikar *et al.*, 1995]. The $\rightarrow$ *weighted modified Wiener index* was later proposed as an extension of the Szeged index to weighted molecular graphs [Gutman and Žerovnik, 2002].

According to the original definition of the bond-additive index, the Wiener index can be calculated by summing the bond contributions w_b^* as the following:

$$W = \sum_{b=1}^B w_b^* = \sum_{b=1}^B \left(\sum_{i < j} \frac{\min P_{ij}^b}{\min P_{ij}} \right)$$

where B is the number of bonds, $\min P_{ij}$ is the number of shortest paths between vertices v_i and v_j , and $\min P_{ij}^b$ is the number of those shortest paths between vertices v_i and v_j that contain the bond b [Lukovits, 1990b, 1992, 1995c; Juvan and Mohar, 1995]. The first summation is performed for all bonds b and for each b bond all pairs of vertices i and j are considered. Contrary to the original Wiener definition, this formula holds for any graphs. For acyclic graphs, the bond contribution w_b^* coincides with the product of the number of vertices on each side of the considered bond, that is, the number of external paths including the bond considered. Moreover, this formula allows the study of the effect of different types of bonds on the global behavior of the molecule and thus interpretation of the structural meaning of the Wiener index. This led to the splitting of the Wiener index into **partial Wiener indices** related to single, double, triple, and aromatic bonds [Lukovits, 1990b]:

$$W = W_S + W_D + W_T + W_A = \sum_{b \in S} w_b^* + \sum_{b \in D} w_b^* + \sum_{b \in T} w_b^* + \sum_{b \in A} w_b^*$$

where W_S is obtained by adding the contributions of all single bonds, W_D by adding the contributions of all double bonds, W_T by adding the contributions of all triple bonds, and W_A by adding the contributions of all aromatic bonds. These partial Wiener indices can be used separately as $\rightarrow$ *multiple bond descriptors*.

The same method based on the bond contribution w_b^* for the Wiener index calculation was proposed by Pisanski–Žerovnik [Pisanski and Žerovnik, 1994] together with a more general edge weight w_b based on paths, defined as

$$w_b = \sum_{i < j} \frac{P_{ij}^b}{P_{ij}}$$

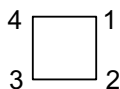
where P_{ij} is the number of paths of any length between vertices v_i and v_j and P_{ij}^b the number of those paths that include the b bond; the sum runs over all pairs of vertices in the graph. By adding all the edge weights w_b a global index, called **Pisanski–Žerovnik index** Ω , is obtained:

$$\Omega = \sum_{b=1}^B w_b$$

where B is the number of graph edges. This index coincides with the Wiener index for any acyclic graph; moreover, the relation $\Omega \geq W$ holds for any graph.

Example W36

Wiener index bond contributions of cyclobutane.



Vertex pair	Paths	$\min p_{ij}^b / \min p_{ij}$				p_{ij}^b / p_{ij}			
		(1, 2)	(2, 3)	(3, 4)	(1, 4)	(1, 2)	(2, 3)	(3, 4)	(1, 4)
(1, 2)	{1, 2}, {1, 4, 3, 2}	1	—	—	—	0.5	0.5	0.5	0.5
(1, 3)	{1, 2, 3}, {1, 4, 3}	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
(1, 4)	{1, 2, 3, 4}, {1, 4}	—	—	—	1	0.5	0.5	0.5	0.5
(2, 3)	{2, 3}, {2, 1, 4, 3}	—	1	—	—	0.5	0.5	0.5	0.5
(2, 4)	{2, 3, 4}, {2, 1, 4}	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
(3, 4)	{3, 4}, {3, 2, 1, 4}	—	—	1	—	0.5	0.5	0.5	0.5
	w_b^*	2.0	2.0	2.0	2.0	3.0	3.0	3.0	3.0

$$\text{Wiener index : } W = \sum_{b=1}^4 w_b^* = 4 \times 2 = 8$$

$$\text{Pisanski–Zerovnik index : } \Omega = \sum_{b=1}^4 w_b = 4 \times 3 = 12$$

The Wiener index is closely related to the $\rightarrow$ *Altenburg polynomial* of a graph; graphs with the same Altenburg polynomial always have the same Wiener number.

The Wiener index increases with the number of atoms (i.e., the molecular size) and, for a constant number of atoms, reaches a maximum for linear structure and a minimum for the most branched and cyclic structures. For this reason, it was suggested as a measure of $\rightarrow$ *molecular branching*. It is insensitive to atom type. Moreover, the Wiener index seems to be related to the molecular surface area, thus reflecting molecular compactness and, in some way, the intermolecular forces [Gutman and Körtévlyesi, 1995].

For all graphs connected, the Wiener index is between

$$\frac{A \cdot (A-1)}{2} \leq W \leq \frac{A \cdot (A^2-1)}{6}$$

where the lower limit refers to a linear $\rightarrow$ *path graph* and the upper to a $\rightarrow$ *complete graph*.

Sometimes a **normalized Wiener index** can be found, defined as

$$W_N = \frac{W}{A^2} \quad \text{or} \quad W_N = \frac{W}{(A+1)^2}$$

where A is the number of graph vertices [Zhang, Liu *et al.*, 1997].

The **root mean square Wiener index** is defined as

$$W_{RMS} = \frac{1}{\sqrt{A(A-1)}} \cdot \left(\sum_{i=1}^A \sum_{j=1}^A d_{ij}^2 \right)^{1/2}$$

where the summation is performed on the square topological distances. The cubic root of the Wiener index was also suggested as a measure of the mean distance among molecule atoms [Platt, 1952; Morales and Araujo, 1993].

Topological indices related to the Wiener index are the $\rightarrow$ *Kirchhoff number*, the $\rightarrow$ *quasi-Wiener index*, the $\rightarrow$ *detour index*, the $\rightarrow$ *expanded Wiener number*, and the $\rightarrow$ *all-path Wiener index*. Moreover, a number of topological indices have been expressed as functions of the Wiener index or as its extensions or modifications [Zhu, Klein *et al.*, 1996]; for examples, $\rightarrow$ *detour/Wiener index*, $\rightarrow$ *hyper-Wiener index*, $\rightarrow$ *Harary index*, $\rightarrow$ *total information content on the distance magnitude*, $\rightarrow$ *mean information content on the distance magnitude*, and $\rightarrow$ *mean information content on the distance degree magnitude*.

A generalization of the Wiener index to account for heteroatoms and multiple bonds was proposed based on the $\rightarrow$ *Barysz distance matrix*; likewise, various $\rightarrow$ *Wiener-type indices* are calculated from different $\rightarrow$ *weighted distance matrices*.

Even/odd distance indices are calculated by a partition of the Wiener index based on counts of even and odd molecular graph distances as [Ivanciuc, Ivanciuc *et al.*, 2001e]:

$$SumE(\lambda_E) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A d_{ij}^{\lambda_E} \quad \text{if } d_{ij} \text{ is even} \quad SumO(\lambda_O) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A d_{ij}^{\lambda_O} \quad \text{if } d_{ij} \text{ is odd}$$

where A is the number of graph vertices, and λ_E and λ_O are exponent parameters for the even-restricted sum and the odd-restricted sum, respectively; these parameters can be optimized in QSAR/QSPR modeling and can be considered among the $\rightarrow$ *variable descriptors*. If $\lambda_E = 1$ and $\lambda_O = 1$, the following relationship with the Wiener index holds:

$$W = SumE(1) + SumO(1)$$

Moreover, from the two partitions of the Wiener index, different molecular descriptors were defined as [Ivanciuc, Ivanciuc *et al.*, 2001e]:

$$W_{O/E}(\lambda_O, \lambda_E) = \frac{SumO(\lambda_O)}{SumE(\lambda_E)} \quad W_{E/O}(\lambda_E, \lambda_O) = \frac{SumE(\lambda_E)}{SumO(\lambda_O)}$$

$$W_{E \times O}(\lambda_E, \lambda_O) = SumE(\lambda_E) \times SumO(\lambda_O)$$

The **Wiener polynomial** (or **Hosoya–Wiener polynomial**) of a graph is defined as [Hosoya, 1988; Sagan, Yeh *et al.*, 1996]

$$H_G(x) = \sum_{k=1}^D {}^k f \cdot x^k = \sum_{i=1}^{A-1} \sum_{j=i+1}^A x^{d_{ij}}$$

where x is a scalar variable, A the total number of vertices, ${}^k f$ is the number of pairs of vertices located at a topological distance equal to k , and d_{ij} is the distance between vertices v_i and v_j . D is the diameter of the graph, which coincides with the Wiener polynomial degree. The first derivative of the Wiener polynomial at $x = 1$ gives the Wiener index, while the higher derivatives give the **extended Wiener indices**, denoted as ${}^m W$ [Estrada, Ivanciuc *et al.*, 1998; Gutman, Estrada *et al.*,

1999; Konstantinova and Diudea, 2000]. The extended Wiener indices are calculated as

$${}^2W = \sum_{k=2}^D k \cdot (k-1) \cdot {}^k f \quad {}^3W = \sum_{k=3}^D k \cdot (k-1) \cdot (k-2) \cdot {}^k f$$

$${}^mW = \sum_{k=m}^D k \cdot (k-1) \cdot (k-2) \cdot \dots \cdot [k-(m-1)] \cdot {}^k f$$

Even/odd Wiener polynomial descriptors were proposed by partitioning the Wiener polynomial based on counts of even and odd distances. The sum of Wiener polynomial terms corresponding to even graph distances and the sum of the terms corresponding to odd graph distances were defined as

$$WiPolE(x) = \sum_{k, \text{even}} {}^k f \cdot x^k \quad WiPolO(y) = \sum_{k, \text{odd}} {}^k f \cdot y^k$$

where ${}^k f$ is the number of pairs of vertices located at a topological distance equal to k , and the summations go up to the maximal distance in the graph; x and y are two independent variable parameters optimized during the modeling procedure.

Another decomposition of the Wiener index was proposed to define topological substituent indices representing the effects of substituents on a parent molecule and the substituent interactions [Lukovits, 1988]. Let M be the parent structure with two substituents denoted by A and B (Figure W2), the Wiener index of the molecule can be decomposed thus:

$$W = W_M + S_A + S_B + S_{AB}$$

where W_M is the Wiener index of the parent structure, that is, the sum of all distances in the parent structure. S_A and S_B are the topological substituent indices defined as

$$S_A = W_A + n_A \cdot s_a + n_A \cdot n_M + n_M \cdot s_c \quad \text{and} \quad S_B = W_B + n_B \cdot s_b + n_B \cdot n_M + n_M \cdot s_e$$

where W_A and W_B are the Wiener indices of the two substituents; n_A , n_B , and n_M denote the number of atoms in the substituents A and B , and in the parent structure M ; note that the terms $n_A \cdot n_M$ and $n_B \cdot n_M$ represent the numbers of times the bond linking A and M as well as B and M has to be traversed; a and b denote the substitution sites, that is, the link vertex, in the parent structure for the substituents A and B , respectively, and c and e denote the link vertices in the substituent A and B , respectively.

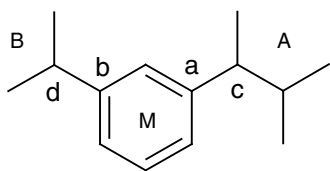


Figure W2 Disubstituted benzene derivative.

The term s is the **connectedness index** characterizing the connectedness of a substitution site defined as [Seybold, 1983b, 1983a]:

$$s_k = \sum_{i \in M} d_{ik}$$

where the summation runs over all atoms of the parent molecule and k denotes a generic substitution site in M , d_{ik} is the topological distance of the i th vertex from the substitution site; the same formula holds for the connectedness of the link vertex in a substituent.

The topological substituent index S_{AB} represents the interaction between the substituents A and B defined as

$$S_{AB} = n_A \cdot s_e + n_B \cdot n_c + n_A \cdot n_B \cdot (2 + d_{ab})$$

where d_{ab} is the topological distance between the two substitution sites in the parent structure.

The topological substituent indices S_A , S_B , and S_{AB} have been used to model the biological activity of a series of congeneric compounds; in the models, the Wiener index W_M of the parent structure can be neglected as it represents the constant term due to the main bulk of the molecules. These substituent indices, unlike the $\rightarrow$ *substituent constants*, can be calculated easily; moreover, they depend on the site at which the substitution takes place and the interaction index S_{AB} overcomes the additive scheme of the substituent constants.

The **Nikolić–Trinajstić–Randić index**, denoted by ${}^mW$ and called by the authors of this book **modified Wiener index**, is defined according to the original Wiener definition but using the reciprocal of the number of vertices on each side of the bond [Nikolić, Trinajstić *et al.*, 2001b; Randić and Zupan, 2001]:

$${}^mW \equiv H_{W_e} = \sum_{b=1}^B (N_{i,b})^{-1} \cdot (N_{j,b})^{-1}$$

where $N_{i,b}$ and $N_{j,b}$ denote the number of vertices on each side of the edge b , including vertices i and j , respectively, and B is the total number of graph edges. It should be noted that this index can be calculated only for acyclic graphs and that it coincides with the $\rightarrow$ *Harary Wiener index* H_{W_e} proposed by Diudea some years before [Diudea, 1997a].

This index was considered desirable from a structural interpretation point of view because it enhances the weights of more external bonds, which are associated with the larger part of the molecular surface and consequently more responsible for the reactive behavior of molecules. Mathematical properties of this modified Wiener index ${}^mW$ were extensively studied [Gutman and Žerovnik, 2002]; in particular, it was shown to possess the basic properties required by a topological index to be acceptable as a measure of the extent of $\rightarrow$ *molecular branching* for acyclic graphs.

There are different classes of modified Wiener indices, based on different ways of generalization of the original Wiener index in terms of a variable parameter λ . These are sometimes called *variable Wiener indices* as, varying the value of the parameter λ , different molecular descriptors are derived. Here, the name **generalized Wiener indices** is suggested in place of **variable Wiener indices** or **modified Wiener indices** to distinguish them from those $\rightarrow$ *variable descriptors* containing variable parameters evaluated by some optimization procedure during the search for the best correlation with a given molecular property.

The **W_λ indices** are a generalization of the Wiener index defined as [Gutman, 1997; Gutman, Vidović *et al.*, 1998]

$$W_\lambda \equiv Wi(\mathbf{D}^\lambda) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij}^\lambda$$

where d_{ij} is the topological distance raised to the variable parameter λ and the summation goes over all the pairs of vertices; D^λ is a $\rightarrow$ *generalized distance matrix* and W the $\rightarrow$ *Wiener operator*. Of course, for $\lambda = 1$, W_1 is the Wiener index, for $\lambda = -1$, the $\rightarrow$ *Harary index* (or *RDSUM index*), and for $\lambda = -2$ the $\rightarrow$ *Harary number*. Other λ values, namely, $\lambda = 1/3$ and $\lambda = -1/3$, were used to model boiling points of C8 alkane isomers [Lučić, Miličević *et al.*, 2003]. Note that for positive integer λ values, W_λ indices coincide with the $\rightarrow$ *distance distribution moments* D_λ [Klein and Gutman, 1999].

Moreover, a modified Wiener index, was proposed in 1992 by Bemis as the sum of the square topological distances ($\lambda = 2$) in molecular subgraphs consisting of three vertices [Bemis and Kuntz, 1992].

The $^\lambda W$ **indices** are another class of generalized Wiener indices, based on the original formulation of the Wiener index, defined as [Gutman, Vukicević *et al.*, 2004; Vukicević and Gutman, 2003; Vukicević and Graovac, 2004a]

$$^\lambda W = \sum_{b=1}^B (N_{i,b})^\lambda \cdot (N_{j,b})^\lambda$$

where $N_{i,b}$ and $N_{j,b}$ are the number of vertices on each side of the edge b , including vertices i and j , respectively, and the summation goes over all the graph edges. For $\lambda = +1$ and $\lambda = -1$, the generalized index reduces to the ordinary Wiener index W and the Nikolić–Trinajstić–Randić index $^m W$, respectively.

It was demonstrated that, for $\lambda > 0$, the two types of generalization of the Wiener index, W_λ and $^\lambda W$, are suitable measures of branching; moreover, their mathematical properties were extensively studied [Gutman, 1997; Vukicević, 2003; Gutman, Vukicević *et al.*, 2004; Gorše and Žerovnik, 2004].

Linear combinations of two different generalized $^\lambda W$ indices constitute a class of $\rightarrow$ *variable descriptors* [Vukicević and Žerovnik, 2003]:

$$W_{\lambda_1, \lambda_2; \alpha_1, \alpha_2} = \alpha_1 \cdot {}^{\lambda_1} W + \alpha_2 \cdot {}^{\lambda_2} W \quad \alpha_1, \alpha_2 > 0, \quad \lambda_1, \lambda_2 \in [-1, 0]$$

where α are parameters to be optimized. These combinations give rise to molecular descriptors that are essentially different from the other single Wiener indices.

Another class of generalized Wiener indices is further derived, only for acyclic graphs, from the $^\lambda W$ indices as [Vukicević and Žerovnik, 2005b]

$$_\lambda W = \frac{1}{2} \cdot \sum_{b=1}^B [A^\lambda - (N_{i,b})^\lambda - (N_{j,b})^\lambda] \quad \lambda \neq 0, 1$$

where $N_{i,b}$ and $N_{j,b}$ are the number of vertices on each side of the edge b , and their sum is always equal to the number A of graph vertices. For $\lambda = +2$, this generalized index reduces to the ordinary Wiener index W .

Finally, another class of generalized Wiener indices, called **altered Wiener indices**, is defined as [Vukicević and Žerovnik, 2005a]:

$$W_{\min, \lambda} = \sum_{b=1}^B [A^\lambda \cdot \min(N_{i,b}, N_{j,b})^\lambda - \min(N_{i,b}, N_{j,b})^{2 \cdot \lambda}] \quad \lambda \neq 0$$

where A is the number of graph vertices and $N_{i,b}$ and $N_{j,b}$ denote the numbers of vertices on each side of the edge b ; the summation goes over all graph edges and λ is some real number. For $\lambda = +1$, this generalized index reduces to the ordinary Wiener index W .

Of course, the altered Wiener indices are valid only for acyclic graphs. These indices, for $\lambda < 0$ and for $\lambda \geq 1$, were demonstrated to be a measure of $\rightarrow$ molecular branching.

To extend all these classes of generalized Wiener indices to cyclic graphs, these indices can be calculated by using the Szeged index algorithm, as modified by Randić in defining the $\rightarrow$ revised Wiener index, that is, counting 1/2 all the vertices that are at equal distance from both the vertices on each side of the bond.

The **Wiener–Hosoya index**, denoted as $W-H$, was defined as [Randić, 2004b]:

$$W-H = W + W_{ee}$$

where W is the Wiener index and W_{ee} is a Wiener-type index calculated by summing the Wiener indices relative to the subgraphs obtained with deletion of each edge and all incident edges to it, following an analogous approach to the $\rightarrow$ Hosoya Z index calculation.

A set of **extended Wiener–Hosoya indices** ${}^mV_t$ [Li, 2002; Li, Li *et al.*, 2003] was proposed by combining the original definition of the Wiener index as the sum of the products $N_{i,b} \times N_{j,b}$ with the definition of $\rightarrow$ Kier–Hall connectivity indices and the approach for the $\rightarrow$ Hosoya Z index calculation:

$${}^mV_t = \frac{1}{A^2} \cdot \sum_{k=1}^K \cdot \left(\prod_{i=1}^n N_{i,t} \right)_k$$

where the summation runs over all of the m th-order subgraphs constituted by n atoms and m edges; K is the total number of m th-order subgraphs present in the molecular graph. The subscript t refers to the type of molecular subgraph and is *ch* for chain or ring, *pc* for path-cluster, *c* for cluster, and *p* for path. The product is over all the vertices of each subgraph. $N_{i,t}$ is the number of vertices on the side of the i th vertex (vertex i included) when all the bonds of the subgraph t considered are contemporarily deleted from the molecular graph. For $m = 0$, $N_{i,t}$ is equal 1, and ${}^0V_t = 1/A$, where A is the number of nonhydrogen atoms.

The **multiplicative Wiener index** was defined as [Gutman, Linert *et al.*, 2000b; Gutman, Linert *et al.*, 2000a]:

$$\pi^W = \prod_{i=1}^{A-1} \prod_{j=i+1}^A d_{ij} = \prod_{k=1}^D k^{k f}$$

where $k f$ is the number of distances in the graph equal to k and D is the largest distance in the graph. In contrast to the Wiener index, the multiplicative Wiener index reflects only long-distance structural features of molecules, since distances equal to 1, corresponding to pairs of bonded atoms, have no effect on the multiplicative Wiener index.

Due to the very large numbers that are often reached by π^W , its logarithmic version seems to be more appropriate in searching for QSAR/QSPR models:

$$\ln \pi^W = \sum_{k=1}^D k f \cdot \ln(k)$$

An empirical modification of the Wiener index, combining a mass term M with the Wiener index was proposed to study chloroalkane properties [Lin, 2004]. This is defined as

$$W_m = M^\lambda \cdot W$$

where λ is a variable parameter and M is defined as

$$M = \frac{\alpha \cdot N_{\text{Cl}} + N_{\text{C}}}{N_{\text{Cl}} + N_{\text{C}}} \quad \alpha = \frac{m_{\text{Cl}}}{m_{\text{C}} + 3 \cdot m_{\text{H}}}$$

where N_{Cl} and N_{C} denote the number of chlorine and carbon atoms, respectively; m_{Cl} , m_{C} , and m_{H} are the atomic masses of chlorine, carbon, and hydrogen atoms, respectively. The optimal value of λ is evaluated during the search for the best regression of the property of interest.

 Additional references are listed in the thematic bibliography (see Introduction).

■ Wiener matrix ($\mathbf{W}$)

By a generalization of the original definition of the $\rightarrow$ *Wiener index* W [Wiener, 1947c], the matrix $\mathbf{W}$ is a square symmetric $A \times A$ matrix, A being the number of graph vertices, derived from the $\rightarrow$ *H-depleted molecular graph* $\mathcal{G}$ and defined only for acyclic graphs [Randić, 1993c]. Each off-diagonal entry of the Wiener matrix corresponds to the number of external paths in the graph that contains the path p_{ij} from vertex v_i to vertex v_j and is calculated as the product of the numbers of vertices on each side of the path p_{ij} , namely, $N_{i,p}$ and $N_{j,p}$, including both vertices i and j . This matrix, which is a **dense Wiener matrix**, is usually called **path-Wiener matrix**, denoted as $\mathbf{W}_p$, and its elements are formally defined as

$$[\mathbf{W}_p]_{ij} = \begin{cases} N_{i,p_{ij}} \cdot N_{j,p_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where

$$N_{i,p_{ij}} = |\{v | v \in V(\mathcal{G}); \quad d_{iv} < d_{jv}; \quad p_{iv} \cap p_{ij} = \{i\}\}|$$

$$N_{j,p_{ij}} = |\{v | v \in V(\mathcal{G}); \quad d_{jv} < d_{iv}; \quad p_{jv} \cap p_{ij} = \{j\}\}|$$

where $V(\mathcal{G})$ is the set of graph vertices and p_{iv} and p_{jv} are the paths connecting the vertex v to vertices i and j , respectively.

The **edge-Wiener matrix**, denoted as $\mathbf{W}_e$, is a sparse matrix, whose elements different from zero are only those corresponding to pairs of adjacent vertices (i.e., edges); this matrix can thus be considered a $\rightarrow$ *weighted adjacency matrix*. The edge-Wiener matrix is formally defined as

$$[\mathbf{W}_e]_{ij} = \begin{cases} N_{i,e_{ij}} \cdot N_{j,e_{ij}} & \text{if } i \neq j \quad \wedge \quad e_{ij} \in E(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases}$$

where the nonvanishing matrix elements are the product of the numbers $N_{i,e}$ and $N_{j,e}$ of vertices on each side of an edge e_{ij} and $E(\mathcal{G})$ represents the set of edges of the graph.

The i th row sum of the Wiener matrix is a local vertex invariant called **Wiener matrix degree** ρ_i :

$$\rho_i \equiv VS_i(\mathbf{W}_{e/p}) = \sum_{j=1}^A [\mathbf{W}_{e/p}]_{ij}$$

where the symbol VS standing for vertex sum is the $\rightarrow$ row sum operator and $\mathbf{W}_{e/p}$ can be either the edge- or the path-Wiener matrix.

Some interesting properties of the path-Wiener matrix are (a) each matrix entry gives the number of paths of which the path p_{ij} is a subgraph; (b) entries corresponding to paths between $\rightarrow$ terminal vertices are necessarily equal to 1; (c) rows of the matrix corresponding to terminal vertices all have entries less than A (i.e., the number of vertices); (d) an entry with value $(A - 1)$ shows a terminal bond because only terminal bonds divide the vertices into $1 + (A - 1)$ partition; and (e) terminal vertices have a smaller row sum (i.e., the Wiener matrix degree) than the centrally located ones associated with larger row sums.

For acyclic graphs, the Wiener matrix $\mathbf{W}_{e/p}$ coincides with the symmetric $\rightarrow$ Cluj distance matrix $\mathbf{SCJD}_{e/p}$ that is derived by symmetrization of the $\mathbf{UCJD}_{e/p}$ as

$$\mathbf{W}_{e/p} \equiv \mathbf{SCJD}_{e/p} = \mathbf{UCJD}_{e/p} \otimes \mathbf{UCJD}_{e/p}^T$$

where the symbol $\otimes$ indicates the $\rightarrow$ Hadamard matrix product.

The **sparse Wiener matrix** of m th order ${}^m\mathbf{W}$ is derived from the path-Wiener matrix setting to zero all entries except those corresponding to paths ${}^m p_{ij}$ of length m . This matrix can be calculated by the Hadamard product of the Wiener matrix and the $\rightarrow$ geodesic matrix ${}^m\mathbf{B}$ whose elements corresponding to paths of length m are equal to 1, or else zero:

$${}^m\mathbf{W} = \mathbf{W}_p \otimes {}^m\mathbf{B}$$

The edge-Wiener matrix $\mathbf{W}_e$ is thus a sparse Wiener matrix of order 1, $\mathbf{W}_e \equiv {}^1\mathbf{W}$.

Several Wiener matrix invariants [Randić, Guo *et al.*, 1993; Randić, Guo *et al.*, 1994] can be calculated from all the Wiener matrices defined above.

The $\rightarrow$ Wiener index W can be obtained from the path-Wiener matrix $\mathbf{W}_p$ by summing all the entries corresponding to edges above the main diagonal or simply by applying the $\rightarrow$ Wiener operator Wi to the edge-Wiener matrix $\mathbf{W}_e$ as

$$W \equiv Wi(\mathbf{W}_e) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{W}_e]_{ij}$$

Higher-order Wiener numbers ${}^m W$ are obtained by summing contributions of paths of a same length m [Randić, Guo *et al.*, 1993]. They are derived from sparse Wiener matrices of order m by using the Wiener operator or from the path-Wiener matrix by summing, above the main diagonal, all entries corresponding to paths of the length m considered. Thus, a sequence of Wiener numbers of different order is obtained (${}^1 W, {}^2 W, {}^3 W, \dots, {}^D W$), the sequence length D being determined by the maximum length of the paths in the graph, that is, the $\rightarrow$ topological diameter [Randić, Guo *et al.*, 1993]:

$${}^m W \equiv Wi({}^m\mathbf{W}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [{}^m\mathbf{W}]_{ij}$$

where ${}^mW$ is the sparse Wiener matrix of m th order and Wi indicates the Wiener operator. The Wiener number of first order obviously coincides with the Wiener index, $W \equiv {}^1W$.

The **hyper-Wiener index** WW is calculated from the path-Wiener matrix W_p as [Randić, 1993c]

$$WW \equiv Wi(W_p) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [W_p]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A VS_i(W_p) = \sum_{m=1}^D {}^mW$$

where Wi is the Wiener operator and VS is the vertex sum operator. The hyper-Wiener index can also be calculated as the sum of all possible Wiener numbers ${}^mW$ from order 1 to the maximum order D .

The hyper-Wiener index can be considered among the $\rightarrow ID$ numbers; it measures the “expansiveness” of a graph, weighting expansive graphs even more than does the Wiener index [Klein, Lukovits *et al.*, 1995].

As the Wiener matrix is defined only for acyclic graphs, the hyper-Wiener index is restricted to that kind of graph too [Lukovits, 1994, 1995b; Lukovits, 1995a; Gutman, Linert *et al.*, 1997]; however, some extensions of the hyper-Wiener index to cycle-containing structures have been proposed [Randić, Guo *et al.*, 1993; Lukovits and Linert, 1994].

For instance, for any graph, the hyper-Wiener index can be calculated by applying the Wiener operator to the symmetric path-Cluj distance matrix $SCJD_p$ or the $\rightarrow$ orthogonal Wiener operator to the unsymmetrical path-Cluj distance matrix $UCJD_p$.

Moreover, the $\rightarrow$ hyper-path-distance index D_p has been proposed [Klein, Lukovits *et al.*, 1995; Klein and Gutman, 1999] as a general formula to calculate the hyper-Wiener index WW for any graph in terms of topological distances d_{ij} :

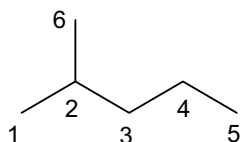
$$\begin{aligned} WW \equiv HyWi(D) \equiv D_p &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{d_{ij} \cdot (d_{ij} + 1)}{2} \\ &= \frac{1}{2} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A (d_{ij}^2 + d_{ij}) = \frac{1}{2} \cdot (D_2 + W) = \frac{1}{2} \cdot [tr(D^2)/2 + W] \end{aligned}$$

where $HyWi$ is the $\rightarrow$ hyper-Wiener operator applied to $\rightarrow$ distance matrix D , D_2 is the second-order $\rightarrow$ distance distribution moment, that is, the sum of the square distances in the graph, and W is the Wiener index. The index D_2 was demonstrated to be equal to half the trace of the distance matrix D raised to the second power [Diudea, 1996a; Diudea and Gutman, 1998]; moreover, it is closely related to the Balaban $\rightarrow$ mean square distance index. If distance distribution moments D_λ of order λ are used instead of the second moment of distances D_2 , a whole sequence of **generalized hyper-Wiener indices** is obtained:

$${}^\lambda WW = \frac{1}{2} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A (d_{ij}^\lambda + d_{ij}) = \frac{1}{2} \cdot (D_\lambda + W)$$

Example W37

Values of some indices derived from Wiener matrices $\mathbf{W}_e$ and $\mathbf{W}_p$ for the H-depleted molecular graph of 2-methylpentane; $\mathbf{D}$ is the topological distance matrix.



$$\mathbf{D} = \begin{array}{c|cccccc} \text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 2 & 3 & 4 & 2 \\ 2 & 1 & 0 & 1 & 2 & 3 & 1 \\ 3 & 2 & 1 & 0 & 1 & 2 & 2 \\ 4 & 3 & 2 & 1 & 0 & 1 & 3 \\ 5 & 4 & 3 & 2 & 1 & 0 & 4 \\ 6 & 2 & 1 & 2 & 3 & 4 & 0 \end{array}$$

$$\mathbf{W}_e = \begin{array}{c|cccccc|c} \text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 & VS_i \\ \hline 1 & 0 & 5 & 0 & 0 & 0 & 0 & 5 \\ 2 & 5 & 0 & 9 & 0 & 0 & 5 & 19 \\ 3 & 0 & 9 & 0 & 8 & 0 & 0 & 17 \\ 4 & 0 & 0 & 8 & 0 & 5 & 0 & 13 \\ 5 & 0 & 0 & 0 & 5 & 0 & 0 & 5 \\ 6 & 0 & 5 & 0 & 0 & 0 & 0 & 5 \end{array}$$

$$\mathbf{W}_p = \begin{array}{c|cccccc|c} \text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 & VS_i \\ \hline 1 & 0 & 5 & 3 & 2 & 1 & 1 & 12 \\ 2 & 5 & 0 & 9 & 6 & 3 & 5 & 28 \\ 3 & 3 & 9 & 0 & 8 & 4 & 3 & 27 \\ 4 & 2 & 6 & 8 & 0 & 5 & 2 & 23 \\ 5 & 1 & 3 & 4 & 5 & 0 & 1 & 14 \\ 6 & 1 & 5 & 3 & 2 & 1 & 0 & 12 \end{array}$$

$${}^1W = [W_p]_{12} + [W_p]_{23} + [W_p]_{34} + [W_p]_{45} + [W_p]_{26} = 32$$

$${}^2W = [W_p]_{13} + [W_p]_{24} + [W_p]_{35} + [W_p]_{36} + [W_p]_{16} = 17$$

$${}^3W = [W_p]_{14} + [W_p]_{25} + [W_p]_{46} = 7$$

$${}^4W = [W_p]_{15} + [W_p]_{56} = 2$$

$$WW = \frac{1}{2} \cdot \sum_{i=1}^A VS_i(\mathbf{W}_p) = \frac{1}{2} \cdot (12 + 28 + 27 + 23 + 14 + 12) = 116/2 = 58$$

$$WW \equiv HyWi(\mathbf{D}) = 5 \cdot \frac{1^2 + 1}{2} + 5 \cdot \frac{2^2 + 2}{2} + 3 \cdot \frac{3^2 + 3}{2} + 2 \cdot \frac{4^2 + 4}{2} = 5 + 15 + 18 + 20 = 58$$

$$WW = W + {}^2W + {}^3W + {}^4W = 32 + 17 + 7 + 2 = 58$$

$$JJ_p^1 = 2 \cdot (12 \cdot 28)^{-1/2} + (28 \cdot 27)^{-1/2} + (27 \cdot 23)^{-1/2} + (23 \cdot 14)^{-1/2} = 0.2413$$

$$JJ_p^2 = 2 \cdot (12 \cdot 28 \cdot 27)^{-1/2} + (28 \cdot 27 \cdot 23)^{-1/2} + (27 \cdot 23 \cdot 14)^{-1/2} = 0.0393$$

The **resistance distance hyper-Wiener index** R' has been proposed [Klein, Lukovits *et al.*, 1995] based on the same general formula as for the hyper-Wiener index, but the topological distance d_{ij} is replaced by the $\rightarrow$ *resistance distance* Ω_{ij} :

$$R' \equiv HyWi(\mathbf{\Omega}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{\Omega_{ij} \cdot (\Omega_{ij} + 1)}{2} = \frac{1}{2} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A (\Omega_{ij}^2 + \Omega_{ij})$$

It should be noted that the hyperdistance–path index and the resistance distance hyper-Wiener index are equivalent to the hyper-Wiener index as defined by Randić for acyclic graphs, but they can also be applied to any cycle-containing connected graphs.

By analogy with the Kier–Hall $\rightarrow$ *connectivity indices* χ , **JJ indices** [Randić, Guo *et al.*, 1994] are derived from the edge- or path-Wiener matrix by using the Wiener matrix degrees ρ_i in place of the $\rightarrow$ *vertex degrees* δ_i :

$${}^1\text{JJ}_{e/p} = \sum_b (\rho_i \cdot \rho_j)_b^{-1/2} \quad {}^2\text{JJ}_{e/p} = \sum_{k=1}^{2p} (\rho_i \cdot \rho_l \cdot \rho_j)_k^{-1/2} \quad {}^m\text{JJ}_{e/p} = \sum_{k=1}^K \left(\prod_{a=1}^n \rho_a \right)_k^{-1/2}$$

where in the formula for ${}^1\text{JJ}$ the summation goes over all the edges b and i and j indicate the vertices forming the edge b ; in ${}^2\text{JJ}$ formula, the summation goes over all paths of length 2, while in the general formula ${}^m\text{JJ}$, the summation goes over all of the m th-order molecular subgraphs constituted by n vertices and m edges; K is the total number of m th-order subgraphs, and in the case of path subgraphs equals the m th-order path count ${}^mP$. These indices would have high discrimination power.

The eigenvalues of the Wiener matrix are other Wiener matrix invariants proposed to model molecular properties. They belong to the class of $\rightarrow$ *spectral indices*.

The **reciprocal Wiener matrix**, denoted by $\mathbf{W}^{-1}$, is the matrix whose elements are the reciprocal of the corresponding Wiener matrix elements [Diudea, 1997a]:

$$[\mathbf{W}_e^{-1}]_{ij} = \frac{1}{[\mathbf{W}_e]_{ij}} \quad [\mathbf{W}_p^{-1}]_{ij} = \frac{1}{[\mathbf{W}_p]_{ij}}$$

All elements equal to zero are left unchanged in the reciprocal matrix. Note that the reciprocal edge-Wiener matrix $\mathbf{W}_e^{-1}$ was later called by other authors **modified edge-Wiener matrix** and denoted as ${}^{me}\mathbf{W}$ [Nikolić, Trinajstić *et al.*, 2001b; Janežič, Miličević *et al.*, 2007]. $\rightarrow$ *Harary Wiener indices* H_{W_e} and H_{W_p} are derived by applying the Wiener operator to the reciprocal edge- and path-Wiener matrix, respectively.

Moreover, the **generalized Wiener matrix**, denoted by $\mathbf{W}^\lambda$, was proposed as a generalization of the reciprocal Wiener matrix as

$$[\mathbf{W}_e^\lambda]_{ij} = [\mathbf{W}_e]_{ij}^\lambda \quad [\mathbf{W}_p^\lambda]_{ij} = [\mathbf{W}_p]_{ij}^\lambda$$

where λ is a real parameter. Applying the Wiener operator to generalized edge-Wiener matrices $\mathbf{W}_e^\lambda$, a class of $\rightarrow$ *generalized Wiener indices* was proposed [Gutman, Vukicević *et al.*, 2004].

Moreover, the **Wiener difference matrix** $\mathbf{W}_\Delta$ was also proposed as $\mathbf{W}_\Delta = \mathbf{W}_p - \mathbf{W}_e$, whose off-diagonal elements are based on path contributions calculated only on paths larger than 1 [Diudea, 1996a, 1996b; Ivanciuc, Ivanciuc *et al.*, 1997]. By applying the Wiener operator to the Wiener difference matrix, Wiener-type indices are obtained that are based on counts of paths larger than 1.

📖 [Diudea and Pârv, 1995; Linert, Renz *et al.*, 1995; Linert, Kleestorfer *et al.*, 1995; Diudea, 1996b; Diudea, Katona *et al.*, 1997; Linert and Lukovits, 1997; Dobrynin, Gutman *et al.*, 1999; Plavšić, 1999; Klavžar, Žigert *et al.*, 2000; Plavšić, Lers *et al.*, 2000; Žigert, Klavžar *et al.*, 2000; Aringhieri, Hansen *et al.*, 2001; Trinajstić, Nikolić *et al.*, 2001; Cash, 2002b; Cash, Klavžar *et al.*, 2002; Gutman, 2002b, 2003b; Gutman, Furtula *et al.*, 2003; Gutman and Furtula, 2003; Li and Jalbout, 2003; Zhou and Gutman, 2004b]

- **Wiener matrix degree** → Wiener matrix
- **Wiener matrix eigenvalues** → spectral indices
- **Wiener matrix leading eigenvalue** → spectral indices ($\odot$ Wiener matrix eigenvalues)
- **Wiener number** $\equiv$ *Wiener index*
- **Wiener operator** → Wiener-type indices
- **Wiener polarity number** $\equiv$ *polarity number* → distance matrix
- **Wiener polymer index** → polymer descriptors
- **Wiener polynomial** → Wiener index
- **Wiener sum D/Ω index** $\equiv$ *D/Ω index* → resistance matrix
- **Wiener sum D/Δ index** $\equiv$ *D/Δ index* → detour matrix
- **Wiener topochemical index** → weighted matrices ($\odot$ weighted distance matrices)

■ Wiener-type indices

These are molecular descriptors calculated from a $\rightarrow$ *graph-theoretical matrix* **M** by analogy with the $\rightarrow$ *Wiener index*. For some, the term “Wiener-type indices” is restricted to those variants of the Wiener index that coincide with the index itself for acyclic structures but yield different values for cycle-containing structures [Lukovits and Linert, 1998].

For any square symmetric graph-theoretical matrix **M**, whose diagonal elements are equal to zero, the Wiener-type indices, denoted by Wi , are defined as half sum of the entries of the matrix:

$$Wi(\mathbf{M}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{M}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A VS_i(\mathbf{M})$$

where A is the number of graph vertices and the symbol VS (vertex sum) stands for the matrix row sum. The formula for the calculation of the Wiener-type indices was called by Ivanciuc **Wiener operator** [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc, 2001c, 2000i].

If **M** is the $\rightarrow$ *distance matrix*, the classical Wiener index is obtained, $Wi(\mathbf{D}) = W$; other Wiener-type indices are $\rightarrow$ *detour index* $\rightarrow$ *Kirchhoff number* or $\rightarrow$ *quasi-Wiener index*, $\rightarrow$ *Szeged index*, $\rightarrow$ *all-path Wiener index*, $\rightarrow$ *edge Wiener index*, $\rightarrow$ *hyper-Wiener index*, $\rightarrow$ *Lu index*, $\rightarrow$ *Cluj-distance index*, $\rightarrow$ *hyper-Cluj-distance index*, $\rightarrow$ *Cluj-detour index*, $\rightarrow$ *hyper-Cluj-detour index*, $\rightarrow$ *Harary indices*, $\rightarrow$ *hyper-Harary indices*, $\rightarrow$ *detour/Wiener index*, $\rightarrow$ *hyper-path-distance index*, $\rightarrow$ *Wiener topochemical index*, $\rightarrow$ *bond length-weighted Wiener index*, $\rightarrow$ *3D-Wiener index*, $\rightarrow$ *bond order-weighted Wiener index*.

For any square symmetric matrix **SM**(w), representing a vertex- and edge-weighted molecular graph, the Wiener-type indices are calculated according to the following [Ivanciuc, 2000i]:

$$Wi(\mathbf{SM}, w) = \sum_{i=1}^A \sum_{j=i}^A [\mathbf{SM}(w)]_{ij} = \sum_{i=1}^A w_i + \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{SM}(w)]_{ij}$$

where w is a weighting scheme and w_i the vertex weight.

Moreover, for any unsymmetrical matrix **UM**, Wiener-type indices $Wi^\perp$ are calculated as

$$Wi^\perp(\mathbf{UM}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{UM}]_{ij} \cdot [\mathbf{UM}]_{ji}$$

The formula for the calculation of the Wiener-type indices from unsymmetrical matrices was called by Ivanciuc **orthogonal Wiener operator** [Ivanciuc, Ivanciuc *et al.*, 1997].

The orthogonal Wiener operator applied to the unsymmetrical matrix **UM** gives the same result as the application of the Wiener operator to the symmetric matrix **SM** obtained as

$$\mathbf{SM} = \mathbf{UM} \otimes \mathbf{UM}^T$$

where the symbol $\otimes$ indicates the Hadamard matrix product.

Moreover, for any unsymmetrical matrix **UM** **matrix sum indices MS** were defined by analogy with Wiener-type indices as [Ivanciuc, 1999c]

$$MS(\mathbf{UM}, w) = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{UM}(w)]_{ij}$$

where w is a weighting scheme and the sums are over all the matrix elements.

📖 [Lukovits, 2001b; Furtula, Gutman *et al.*, 2002; Ivanciuc and Klein, 2002a, 2002b; Gutman, Vidović *et al.*, 2003d; Klavžar and Gutman, 2003; Klein, 2003b]

- **Wiener-Wiener number** → double invariants
- **Wilcox resonance energy** → delocalization degree indices
- **Wilson solute descriptors** → Linear Solvation Energy Relationships
- **W^λ indices** → Wiener index
- **W_λ indices** → Wiener index
- **$^\lambda W$ indices** → Wiener index
- **$_\lambda W$ indices** → Wiener index
- **Wiswesser Line Notation system** → molecular descriptors
- **WSHA index** $\equiv$ *WHIM shape* → WHIM descriptors (⊙ global WHIM descriptors)
- **WSIZ index** $\equiv$ *WHIM size* → WHIM descriptors (⊙ global WHIM descriptors)
- **WSYM index** $\equiv$ *WHIM symmetry* → WHIM descriptors (⊙ global WHIM descriptors)
- **WTPT index** → path counts
- **W'/W index** → bond order indices (⊙ graphical bond order)
- **WW'/WW index** → bond order indices (⊙ graphical bond order)
- **WWC matrices** $\equiv$ *WHIM weighted covariance matrices* → WHIM descriptors

X

- **XI matrix** → weighted matrices (⊙ weighted distance matrices)
- **X index** → topological information indices (⊙ extended local information on distances)
- **X-like indices** → topological information indices (⊙ Balaban-like information indices)
- **XLOGP** → lipophilicity descriptors

■ Xu index

It is a topological molecular descriptor derived from the → *adjacency matrix A* and → *distance matrix D* in a similar way as the → *Schultz molecular topological index MTI*; it is defined for a → *H-depleted molecular graph* as [Ren, 1999]

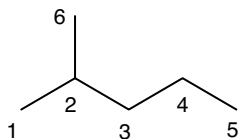
$$Xu = \sqrt{A} \cdot \log(L) = \sqrt{A} \cdot \log\left(\frac{\sum_{i=1}^A \delta_i \cdot \sigma_i^2}{\sum_{i=1}^A \delta_i \cdot \sigma_i}\right) = \sqrt{A} \cdot \log\left(\frac{\sum_{i=1}^A \delta_i \cdot \sigma_i^2}{S}\right)$$

where *A* is the number of graph vertices, δ the → *vertex degree*, σ the → *vertex distance degree*, and *S* is the → *S index*, which is a part of the Schultz molecular topological index. The term *L* represents the valence average topological distance in a graph, that is, a sort of valence weighted average distance degree.

It was proposed as a particularly high discriminant molecular descriptor accounting for molecular size and branching; however, no discrimination can be achieved for heteroatom-containing molecules. Based on a similar approach are the → *Sh indices* that unlike the *Xu* index are based on the → *valence vertex degree*.

Example X1

Calculation of the *Xu* index for the H-depleted molecular graph of 2-methylpentane. δ and σ are the valence vertex degrees and the distance sums, respectively.



Atom	1	2	3	4	5	6
δ	1	3	2	2	1	1
σ	12	8	8	10	14	12

$$\begin{aligned}Xu &= \sqrt{6} \times \log \left[\frac{1 \times 12^2 + 3 \times 8^2 + 2 \times 8^2 + 2 \times 10^2 + 1 \times 14^2 + 1 \times 12^2}{1 \times 12 + 3 \times 8 + 2 \times 8 + 2 \times 10 + 1 \times 14 + 1 \times 12} \right] \\&= \sqrt{6} \times \log(10.245) = 2.475\end{aligned}$$

- **Xu-Stevenson drug-like index** → scoring functions
- **X weighting scheme** → weighting schemes

Y

- **Yang connectivity index** → connectivity indices
- **Yang's electronegativity force gauge** → vertex degree
- **Yang vertex degree** → vertex degree
- **Y index** → topological information indices (⊙ mean extended local information on distances)
- **Ylike indices** → topological information indices (⊙ Balaban-like information indices)
- **Ypolarity scale** → Linear Solvation Energy Relationships (⊙ dipolarity/polarizability term)
- **γ -randomization test** $\equiv \gamma$ -scrambling → validation techniques
- **γ -scrambling** → validation techniques
- **Yukawa–Tsuno equation** → electronic substituent constants (⊙ resonance electronic constants)
- **Yule characteristic** → statistical indices (⊙ concentration indices)
- **Yule similarity coefficient** → similarity/diversity (⊙ Table S9)
- **Y weighting scheme** → weighting schemes

Z

■ Zagreb indices (M_n)

→ *Topological indices* based on the → *vertex degree* δ of the atoms in the → *H-depleted molecular graph* and called **first Zagreb index**, denoted as M_1 , and **second Zagreb index**, denoted as M_2 . They are defined as [Gutman and Trinajstić, 1972; Gutman, Rusčić *et al.*, 1975; Nikolić, Kovacevic *et al.*, 2003; Gutman and Das, 2004]

$$M_1 = \sum_{i=1}^A \delta_i^2 = \sum_g g^2 \cdot {}^gF \quad M_2 = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i \cdot \delta_j) = \sum_{b=1}^B (\delta_{b(1)} \cdot \delta_{b(2)})_b$$

where, in M_1 , the summation runs over the A nonhydrogen atoms of the molecule and δ_i is the → *vertex degree* of the i th atom; ${}^gF$ is the → *vertex degree count*, that is, the number of atoms with the same vertex degree, g is the value of the considered vertex degree. In M_2 , the first summation goes over all the pairs of vertices v_i and v_j in the molecular graph, but only contributions from pairs of adjacent vertices are accounted for, a_{ij} being the elements of the → *adjacency matrix*; the second summation goes over all the edges in the molecular graph. A and B are the total number of vertices and edges in the graph, respectively; δ_i and δ_j are the vertex degrees of the vertices v_i and v_j ; the subscripts $b(1)$ and $b(2)$ represent the two vertices connected by the edge b .

For isomeric series, the Zagreb indices are related to the → *molecular branching*.

The two Zagreb indices are strictly related to zero-order ${}^0\chi$ and first-order ${}^1\chi$ → *connectivity indices*, respectively. The first Zagreb index M_1 (also called **Gutman index**) is also related to the → *Platt number* F , the → *connection number* N_2 , the molecular walk count of order 2 ($mwc^{(2)}$), and the molecular self-returning walk count of order 4 ($srw^{(4)}$) by the relationship

$$M_1 = F + 2(A-1) = 2 \cdot (N_2 + A-1) = mwc^{(2)} = \frac{srw^{(4)} + 2 \cdot B}{2}$$

where A is the number of atoms. It has been pointed out that M_1 is also related to the → *molecular complexity*. The second Zagreb index M_2 is a part of the → *Schultz molecular topological index*. Other interesting properties of Zagreb indices and their connection to other graph invariants are reported in [Nikolić, Kovacevic *et al.*, 2003; Gutman and Das, 2004].

The **modified Zagreb indices** are defined as [Bonchev, 2001a; Bonchev and Trinajstić, 2001; Golbraikh, Bonchev *et al.*, 2001b; Nikolić, Kovacevic *et al.*, 2003]

$${}^mM_1 = \sum_{i=1}^A (\delta_i^2)^{-1} \quad {}^mM_2 \equiv {}^1\text{ON} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i \cdot \delta_j)^{-1}$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to one only for adjacent vertices, and zero otherwise; ${}^1\text{ON}$ is the first-order modified $\rightarrow$ *overall Zagreb index*.

A generalization of the original Zagreb indices are the **variable Zagreb indices** defined as [Li and Zhao, 2004; Miličević, Nikolić *et al.*, 2004; Miličević and Nikolić, 2004]

$${}^vM_1 = \sum_{i=1}^A (\delta_i^2)^\lambda \quad {}^vM_2 = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i \cdot \delta_j)^\lambda$$

where λ is a real number whose different values lead to well-known molecular descriptors and a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to one only for adjacent vertices, and zero otherwise. For $\lambda = 1/2$, ${}^vM_1$ reduces to the $\rightarrow$ *total adjacency index* A_V that is the sum of all the vertex degrees in the molecular graph and twice the number B of graph edges:

$$A_V = \sum_{i=1}^A \delta_i = 2 \cdot B$$

This index was proposed as the first and simplest measure of graph connectivity. Defined in the same way as the total adjacency index, but using different vertex degrees to take into account multiple bonds and heteroatoms, three other simple molecular descriptors were defined [Pogliani, 1992a]:

$$D^v = \sum_{i=1}^A \delta_i^v \quad D^b = \sum_{i=1}^A \delta_i^b \quad D^Z = \sum_{i=1}^A \delta_i^Z$$

where δ^v is the $\rightarrow$ *valence vertex degree*, δ^b the $\rightarrow$ *bond vertex degree*, and δ^Z is the $\rightarrow$ *Z-delta number*. The third index D^Z is called **Pogliani index** [Pogliani, 1996c, 1997b; Nikolić and Raos, 2001]. Obviously, analogous descriptors can be obtained by summing up other differently defined $\rightarrow$ *vertex degrees*; these molecular descriptors, coinciding with $\rightarrow$ *overall connectivity indices* of zero-order, are here called **sum-delta connectivity indices**.

Since the total adjacency is a highly degenerate graph invariant, assuming the same value for all molecules having the same number of bonds, other functions of vertex degrees were used to describe molecular branching, such as Zagreb indices ($\lambda = 1$) and the $\rightarrow$ *Randić connectivity index* that coincides with the second variable Zagreb index for $\lambda = -1/2$. Later, Bonchev [Bonchev, 2001a] proposed the modified Zagreb index ${}^1\text{ON}$ by using $\lambda = -1$, which gives an intermediate function between Zagreb index M_2 and Randić connectivity index. Analogously, Nikolić proposed the **modified total adjacency index** as [Nikolić, Kovacevic *et al.*, 2003]

$${}^mA = \sum_{i=1}^A \delta_i^{-1}$$

Zagreb indices can be also calculated by using the $\rightarrow$ *valence vertex degree* δ^v [DRAGON – Taletė s.r.l., 2007] and the $\rightarrow$ *edge degree* ϵ [Miličević, Nikolić *et al.*, 2004; Miličević and Nikolić, 2004] in place of the simple vertex degree; the resulting indices are the **valence Zagreb indices**

and **edge-degree Zagreb indices**, respectively:

$$M_1^v = \sum_{i=1}^A (\delta_i^v)^2 \quad M_2^v = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i^v \cdot \delta_j^v)$$

$${}^eM_1 = \sum_{b=1}^B (\epsilon_b^2) \quad {}^eM_2 = \sum_{i=1}^{B-1} \sum_{j=i+1}^B [E]_{ij} \cdot (\epsilon_i \cdot \epsilon_j)$$

where a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to one only for pairs of adjacent vertices, and $[E]_{ij}$ are the elements of the $\rightarrow$ *edge adjacency matrix* equal to one only for pairs of adjacent edges. A and B are the number of vertices and edges, respectively.

The **Zagreb topochemical indices** [Bajaj, Sami *et al.*, 2005] are modifications of the original Zagreb indices, which account both for the presence and relative position of heteroatoms in a H-depleted molecular graph. They are calculated from the $\rightarrow$ *chemical adjacency matrix* based on relative atomic weights as

$$M_1^c = \sum_{i=1}^A (\delta_i^c)^2 \quad M_2^c = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i^c \cdot \delta_j^c)$$

where δ_i^c is the $\rightarrow$ *Madan chemical degree* of the i th atom, calculated by summing up the relative atomic weights of all its adjacent atoms; a_{ij} are the elements of the $\rightarrow$ *adjacency matrix* equal to one only for pairs of adjacent vertices.

A modification of the Zagreb indices was also proposed to define $\rightarrow$ *chirality descriptors* [Golbraikh, Bonchev *et al.*, 2001a, 2001b].

By normalization (i.e., imposing a lower bound equal to zero for linear graphs) of the M_1 index, the **quadratic index**, denoted by Q and also called **normalized quadratic index**, was defined as [Balaban, 1979]

$$Q = \frac{\sum_g [(g^2 - 2g) \cdot {}^gF + 2]}{2} = 3 - 2 \cdot A + \frac{M_1}{2}$$

where g are the different vertex degree values and ${}^gF$ is the vertex degree count. It can be demonstrated that the quadratic index is equal to the difference between the M_1 index of the considered graph and the M_1 index of the corresponding $\rightarrow$ *linear graph*. Moreover, for acyclic molecules, it can be simply calculated by

$$Q = 3 \cdot {}^4F + {}^3F$$

where ${}^3F$ and ${}^4F$ are the vertex degree counts of the three and four orders, respectively.

The **binormalized quadratic index** Q' is obtained by binormalization of M_1 , that is, imposing a lower bound equal to zero for linear graphs and an upper bound equal to one for star graphs [Balaban, 1979]. In practice, it is calculated by dividing the quadratic index Q by its respective value for the $\rightarrow$ *star graph*:

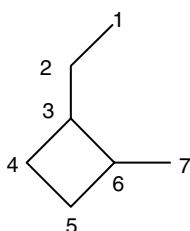
$$Q' = \frac{2 \cdot Q}{(A-2) \cdot (A-3)}$$

where A is the number of graph vertices.

The binormalized quadratic index was proposed, together with the $\rightarrow$ *binormalized centric index* C' , to provide information on the topological shape of trees, that is, the trade-off between linear and star graphs.

Example Z1

Zagreb indices and some modified or related indices for the H-depleted molecular graph of 1-ethyl-2-methyl-cyclobutane. δ_i is the vertex degree, ${}^{\#}F$ is the vertex degree count.



$$\begin{aligned}
 A &= 7 & B &= 7 & \delta_1 &= \delta_7 = 1 & \delta_2 &= \delta_4 = \delta_5 = 2 & \delta_3 &= \delta_6 = 3 \\
 {}^1F &= 2 & {}^2F &= 3 & {}^3F &= 2 \\
 M_1 &= 1^2 + 2^2 + 3^2 + 2^2 + 2^2 + 3^2 + 1^2 = 32 \\
 M_2 &= 1 \times 2 + 2 \times 3 + 3 \times 2 + 2 \times 2 + 2 \times 3 + 3 \times 3 + 3 \times 1 = 36 \\
 {}^m M_1 &= 1 + 2^{-2} + 3^{-2} + 2^{-2} + 2^{-2} + 3^{-2} + 1 = 2.97 \\
 {}^m M_2 &= (1 \times 2)^{-1} + (2 \times 3)^{-1} + (3 \times 2)^{-1} + (2 \times 2)^{-1} + (2 \times 3)^{-1} \\
 &\quad + (3 \times 3)^{-1} + (3 \times 1)^{-1} = 1.69 \\
 Q &= 3 - 2 \times 7 + \frac{M_1}{2} = 5.0 \\
 Q' &= \frac{2 \times 5.0}{(7-2) \times (7-3)} = 0.5
 \end{aligned}$$

📖 [Nikolić, Tolić *et al.*, 2000; Golbraikh, Bonchev *et al.*, 2001a; Bonchev and Trinajstić, 2001; Sardana and Madan, 2002a; Vukicević and Trinajstić, 2003; Das and Gutman, 2004; Nikolić, Miličević *et al.*, 2004; Peng, Fang *et al.*, 2004; Vukicević and Graovac, 2004c; Zhou, 2004b, 2007; Zhou and Gutman, 2004b, 2005; Braun, Kerber *et al.*, 2005; Gutman, Furtula *et al.*, 2005; Liu and Gutman, 2006, 2007; Zhou and Stevanović, 2006; Hansen and Vukicević, 2007; Vukicević, 2007]

- **Zagreb matrices** $\rightarrow$ weighted matrices ($\odot$ weighted adjacency matrices)
- **Zagreb topochemical indices** $\rightarrow$ Zagreb indices
- **Z-counting polynomial** $\rightarrow$ Hosoya Z index
- **Z-delta number** $\rightarrow$ vertex degree

■ Zenkevich index

A molecular descriptor providing a reliable estimate of the internal molecular energy of alkanes, related to their vibrational energies and defined as [Zenkevich, 1998, 1999]

$$U = \sum_c \sqrt{\frac{(m_C + 2 \cdot m_H) \cdot n_C + 2 \cdot m_H}{[(m_C + 2 \cdot m_H) \cdot n_1] \cdot [(m_C + 2 \cdot m_H) \cdot n_2 + m_H]}}$$

where the summation goes over all the C–C bonds, m denotes the atomic masses, n_C the number of carbon atoms and n_1 and n_2 are the number of carbon atoms on the two sides of the C–C bond ($n_1 + n_2 = n_C$). This index is linearly related to the $\rightarrow$ *Wiener index* [Gutman, Furtula *et al.*, 2004b].

- **Z index** $\equiv$ Hosoya Z index
- **Z' index** $\equiv$ Hosoya Z' index $\rightarrow$ characteristic polynomial-based descriptors
- **Z* index** $\rightarrow$ Hosoya Z index
- **Z-Modified Information Content index** $\rightarrow$ indices of neighborhood symmetry
- **^mZ numbers** $\rightarrow$ Hosoya Z matrix
- **Z-matrix** $\rightarrow$ molecular geometry
- **ZMIC index** $\equiv$ Z-Modified Information Content index $\rightarrow$ indices of neighborhood symmetry
- **Z polarity scale** $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ dipolarity/polarizability term)
- **z-scales** $\equiv$ z-scores $\rightarrow$ Principal Component Analysis
- **z-scores** $\rightarrow$ Principal Component Analysis
- **Z weighting scheme** $\rightarrow$ weighting schemes
- **Z'/Z index** $\rightarrow$ Hosoya Z matrix

Greek Alphabet Entries

α A (alpha)

■ $^k\alpha$ descriptors

These are $\rightarrow$ *Wiener-type indices* derived from the powers of the $\rightarrow$ χ matrix, by summing up all the elements above the main diagonal:

$$^k\alpha = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\chi_e^k]_{ij}$$

where χ_e^k is the k th $\rightarrow$ *power matrix* ($k = 1, 2, 3, \dots$) of the χ matrix; A is the number of graph vertices [Randić, 1992d].

The first term $^1\alpha$ coincides with the $\rightarrow$ *Randić connectivity index* χ ; the remaining values show monotonically smaller values. In acyclic graphs, the even powers of χ_e^k matrices necessarily make zero contributions as it is not possible to reach an adjacent vertex by an even number of steps.

- α scale $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- α_m scale $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- α_2^H scale $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)

β B (beta)

- β scale $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- β_m scale $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)
- β_2^H scale $\rightarrow$ Linear Solvation Energy Relationships ($\odot$ hydrogen-bond parameters)

δ Δ (delta)

- Δ/D index $\rightarrow$ detour matrix
- $\Delta \log P$ $\rightarrow$ hydrogen-bonding descriptors ($\odot$ hydrogen-bonding ability constants)

η H (eta)

- η - χ diagram $\rightarrow$ quantum-chemical descriptors ($\odot$ hardness indices)

λ Λ (lambda)

- **λ matrix** → layer matrices (⊙ cardinality layer matrix)
- **$\lambda\lambda_1$ branching index** → spectral indices

 π Π (pi)

- **π -inductive effect** → electronic substituent constants
- **π^* polarity scale** → Linear Solvation Energy Relationships (⊙ dipolarity/polarizability term)

 σ Σ (sigma)

- **σ -electron descriptor** → vertex degree
- **σ electronic constants** $\equiv$ *electronic substituent constants*
- **σ^* electronic constant** $\equiv$ *Taft σ^* constant* → electronic substituent constants (⊙ inductive electronic constants)
- **σ_r electronic constants** → electronic substituent constants (⊙ field/resonance effect separation)
- **σ -inductive effect** → electronic substituent constants
- **σ_α^* radical substituent constants** → electronic substituent constants (⊙ field/resonance effect separation)

 τ T (tau)

- **τ matrix** $\equiv$ *path-sequence matrix* → sequence matrices

 χ X (chi)

- **χ matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **χ^E matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **χ'/χ index** → bond order indices (⊙ graphical bond order)

 φ Ψ (psi)

- **φ_0 index** → chromatographic descriptors (⊙ retention time)

 ω Ω (omega)

- **Ω/D index** → resistance matrix

Numerical Entries

- **0D-descriptors** → molecular descriptors
- **1D-descriptors** → molecular descriptors
- **first-order sparse matrix** → algebraic operators ($\odot$ sparse matrices)
- **first Zagreb index** → Zagreb indices
- **2D-descriptors** → molecular descriptors
- **2D-QSAR** → structure/response correlations
- **second-grade structural parameters** → multiple bond descriptors
- **second-order submolecular polarity parameter** → charge descriptors ($\odot$ submolecular polarity parameter)
- **second path matrix** → Laplacian matrix
- **second Zagreb index** → Zagreb indices
- **3D-Balaban index** → Balaban distance connectivity index
- **3D-connectivity indices** → connectivity indices
- **3D-descriptors** → molecular descriptors
- **3D-double invariants** → double invariants
- **3D-HoVAIF descriptors** → 3D-VAIF descriptors

■ **3D-VAIF descriptors** ($\equiv$ *Three-Dimensional Vector of Atomic Interaction Field descriptors*)

These are → *vectorial descriptors* derived by an approach similar to that of → *MEDV-13 descriptor* and defined in terms of nonbonding interaction energies between pairs of atom types [Zhou, Zhou *et al.*, 2006]. Five atom types are defined on the basis of the chemical element of the most occurring atoms in organic compounds; these are (1) H; (2) C; (3) N or P; (4) O or S; (5) F, Cl, Br, or I.

There are a total of 15 (i.e., $(5 \times 6)/2$) atom-type pairs if both interactions between the atoms of the same type and cross-interactions are calculated. For each atom-type pair, two different nonbonding interactions (electrostatic and van der Waals potential) are considered, leading to a final molecular vector comprising of 30 interactions terms.

The electrostatic interaction for each atom-type pair (u, v) is calculated according to the → *Coulomb potential energy function* defined as

$$E_{el}(u, v) = \sum_{i \in u} \sum_{j \in v} \frac{e^2 \cdot q_i \cdot q_j}{4\pi \cdot \epsilon_0 \cdot r_{ij}}$$

where the first summation is over all the atoms of type u and the second over all the atoms of type v ; r_{ij} is the interatomic geometric distance (Å) between the i th atom and the j th atom, q_i and q_j are the corresponding partial atomic charges. The constant e ($1.602\,1892 \times 10^{-19}$ C) represents the elementary charge, while ϵ_0 ($8.854\,187\,82 \times 10^{-12}$ C²/J m) is the $\rightarrow$ dielectric constant in vacuum.

The van der Waals interaction for each atom-type pair (u, v) is calculated according to the general formula of $\rightarrow$ Lennard-Jones 6-12 potential function as

$$E_{vdw}(u, v) = \sum_{i \in u} \sum_{j \in v} \epsilon_{ij} \cdot \left[\left(\frac{R_{ij}^*}{r_{ij}} \right)^{12} - 2 \cdot \left(\frac{R_{ij}^*}{r_{ij}} \right)^6 \right]$$

where the first summation runs over all the atoms of type u and the second over all the atoms of type v ; r_{ij} is the interatomic geometric distance (Å) between the i th atom and the j th atom; $\epsilon_{ij} = \sqrt{\epsilon_i \cdot \epsilon_j}$ is the well depth and $R_{ij}^* = (R_i^* + R_j^*)/2$ the average van der Waals radius for atom pairs.

3D-HoVAIF descriptors (or *Three-Dimensional Holographic Vector of Atom Interacting Field descriptors*) are vectorial descriptors derived as an extension of 3D-VAIF descriptors accounting for a large number of atom types distinguished by atom hybridization [Ren, Chen *et al.*, 2008]. Ten atom types are defined on the basis of the chemical element and the hybridization state; these are (1) H; (2) C(sp³); (3) C(sp²); (4) C(sp); (5) N(sp³) or P(sp³); (6) N(sp²) or P(sp²); (7) N(sp) or P(sp); (8) O(sp³) or S(sp³); (9) O(sp²) or S(sp²); (10) F, Cl, Br, or I. Then, a total of 55 (i.e., $(10 \times 11)/2$) atom-type pairs derive; moreover, for each atom-type pair, three different non-bonding interactions (electrostatic, van der Waals, and hydrophobic) are considered, leading to a final molecular vector comprising of 165 interactions terms.

In defining the Lennard-Jones 6-12 potential function, a modified van der Waals radius for atom pairs is calculated as

$$R_{ij}^* = (k_{hi} \cdot R_i^* + k_{hj} \cdot R_j^*)/2$$

where k_h are correction factors for hybridization: $k_h(\text{sp}^3) = 1.00$, $k_h(\text{sp}^2) = 0.95$, and $k_h(\text{sp}) = 0.90$.

The hydrophobic interaction for each atom-type pair (u, v) is calculated according to the general formula of $\rightarrow$ Kellogg and Abraham interaction field:

$$E_{hy}(u, v) = \sum_{i \in u} \sum_{j \in v} (SA_i \cdot h_i \cdot SA_j \cdot h_j \cdot e^{-r_{ij}} \cdot T_{ij})$$

where the first summation runs over all the atoms of type u and the second over all the atoms of type v ; r_{ij} is the interatomic geometric distance (Å) between the i th atom and the j th atom; SA is the atomic $\rightarrow$ solvent-accessible surface area, h the atomic hydrophobic constant, and T_{ij} a discriminant function accounting for entropic changing orientation caused by different interatomic interactions.

- **3D molecular autocorrelation** $\rightarrow$ autocorrelation descriptors
- **3D-MoRSE descriptors** $\rightarrow$ molecular transforms
- **3D-MTI' index** $\rightarrow$ Schultz molecular topological index
- **3D-PP** $\equiv$ 3D principal properties $\rightarrow$ principal component analysis

- **3D principal properties** → principal component analysis
- **3D-QSAR** → structure/response correlations
- **3D-Schultz index** → Schultz molecular topological index
- **3D-TDB descriptors** $\equiv$ 3D-topological distance based descriptors → autocorrelation descriptors
- **3D-topological distance based descriptors** → autocorrelation descriptors
- **3D-Wiener index** → molecular geometry
- **4D-Absolute Molecular Similarity Analysis** → 4D-Molecular Similarity Analysis
- **4D-descriptors** → molecular descriptors
- **4D-fingerprints** → 4D-Molecular Similarity Analysis

■ 4D-Molecular Similarity Analysis ($\equiv$ 4D MS Analysis)

This is a methodology aimed at measuring similarity/diversity between molecules, both on a relative and an absolute basis [Duca and Hopfinger, 2001].

Relative similarity is dependent on an alignment constraint (an external reference frame) while the absolute similarity is alignment independent. This methodology accounts for the thermodynamic distribution of conformer states of a molecule and is able to provide measures of similarity with respect to the whole molecule as well as pharmacophoric features of the molecule.

To generate the molecular descriptors, by which molecular similarity is evaluated, first, the **Conformation Energy Profile** (CEP) of each molecule is estimated. This is indicated by the Boltzmann distribution plot of the number of conformers $N(\Delta E)$ at energy ΔE . CEP was first defined and used in the framework of → 4D-QSAR Analysis.

Then, the **Main Distance-Dependent Matrix** (MDDM), for each pair of **Interaction Pharmacophore Elements** (IPEs) is estimated. The IPEs are specific and independent groups representing molecule functionality (Table Numerical Entries1). The seventh IPE type (HS) encodes information about the overall molecular shape since all the nonhydrogen atoms are considered. From the MDDM estimated for the HS-HS pair, a similarity measure with respect to the whole molecule is obtained.

Table Numerical Entries1 Interaction Pharmacophore Elements (IPEs) used in 4D-Molecular Similarity Analysis.

IPE code	Symbol	Definition
0	ALL	all atoms in the molecule
1	NP	non-polar atoms
2	P+	polar atoms with positive charge
3	P-	polar atoms with negative charge
4	HBA	hydrogen-bond acceptor atoms
5	HBD	hydrogen-bond donor atoms
6	ARO	aromatic atoms
7	HS	non-hydrogen atoms

In **4D-Absolute Molecular Similarity Analysis**, the elements of the MDDM for each pair (u, v) of IPE types are defined as

$$[\text{MDDM}(u, v; \mathfrak{g})]_{ij} = e^{-(\mathfrak{g} \cdot \bar{r}_{ij})}$$

where ϑ is a constant commonly settled to 0.25, the value that maximizes the ranges in the molecular similarity measures, $\bar{r}_{ij}$ is the average distance between the i th atom of IPE type u and the j th atom of IPE type v

$$\bar{r}_{ij} = \sum_k r_{ij}(k) \cdot p(k)$$

where the summation runs over the conformer states, $r_{ij}(k)$ is the distance between the i th atom of IPE type u and the j th atom of IPE type v for the k th conformer state and $p(k)$ the thermodynamic probability of the k th conformer state.

In **4D-Relative Molecular Similarity Analysis**, to construct the MDDM for each pair of IPEs, the $\rightarrow$ *Grid Cell Occupancy Descriptors* (GCODs) need first to be calculated by performing a partial 4D-QSAR analysis. A GCOD is defined as the probability that a given IPE type will occupy a specific grid cell in a given molecule.

The elements of the **MDDM** for each IPE pair (u, v) are then defined as

$$[\mathbf{MDDM}(u, v; \vartheta)]_{ij} = \text{GCOD}_i \cdot \text{GCOD}_j \cdot e^{-(\vartheta \cdot r_{ij})}$$

where ϑ is a constant as above, r_{ij} is the geometric distance between the center of the i th grid cell of IPE type u and the center of the j th grid cell of IPE type v . The product of grid cell occupancy descriptors corresponds to the joint probability of an IPE of type u being at grid cell i and an IPE of type v being at grid cell j .

The **MDDM** for each pair of IPEs that are the same (i.e., $u = v$) is a square symmetric matrix. Its eigenvalues, normalized and sorted in descending order, constitute a set of molecular descriptors. Normalization is obtained using the rank of the **MDDM** as a weighting factor:

$$\lambda_m(u, u) = \frac{\lambda'_m(u, u)}{\text{rank}(u, u)}$$

The **MDDM** for pairs of IPE that are not the same (i.e., $u \neq v$) are in general not square matrices; therefore, to calculate its eigenvalues, the matrix is transformed into a square symmetric matrix as the following:

$$\mathbf{MDDM}(u, u) = \mathbf{MDDM}(u, v) \cdot \mathbf{MDDM}^T(u, v)$$

$$\mathbf{MDDM}(v, v) = \mathbf{MDDM}(v, u) \cdot \mathbf{MDDM}^T(v, u)$$

The nonzero eigenvalues of both matrices **MDDM**(u, u) and **MDDM**(v, v) coincide. Therefore, the final eigenvalues are calculated as the square root of the eigenvalues of one of these matrices.

4D-MS descriptors are $\rightarrow$ *spectral indices* containing all of the eigenvalues (36) of the MDDMs constructed for all combinations of two IPE types. Eigenvalues smaller than a given threshold (e.g., 0.002) are usually not included in the final descriptor. These descriptors encode information on molecular size, conformational flexibility, chemical structure, and pharmacophore information. Moreover, descriptors calculated in the framework of 4D-Absolute Molecular Similarity Analysis are alignment independent and also referred to as **4D-fingerprints**.

The similarity between molecules s and t is finally calculated as

$$s_{st} = \left(1 - \sum_m |\lambda_{ms} - \lambda_{mt}|\right) \cdot \left(1 - \frac{|\text{rank}_s - \text{rank}_t|}{\text{rank}_s + \text{rank}_t}\right)$$

where the sum is the $\rightarrow$ *Manhattan distance* and the second term is a correction factor accounting for different molecular sizes, *rank* corresponding to the dimensionality of the MDDMs of the two molecules.

📖 [Senese, Duca *et al.*, 2004; Liu, Yang *et al.*, 2006; Iyer and Hopfinger, 2007; Iyer, Zheng *et al.*, 2007]

- **4D-MS Analysis** $\equiv$ *4D-Molecular Similarity Analysis*
- **4D-MS descriptors** $\rightarrow$ 4D-Molecular Similarity Analysis
- **4D-QSAR Analysis** $\rightarrow$ grid-based QSAR techniques
- **4-Potential Pharmacophore Point keys** $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **4-PPP keys** $\equiv$ *4-Potential Pharmacophore Point keys* $\rightarrow$ substructure descriptors ($\odot$ pharmacophore-based descriptors)
- **4D-QSAR** $\equiv$ *dynamic QSAR* $\rightarrow$ structure/response correlations
- **4D-Relative Molecular Similarity Analysis** $\rightarrow$ 4D Molecular Similarity Analysis

Bibliography

- A-Razzak, M. and Glen, R.C. (1992) Application of rule-induction in the derivation of quantitative structure–activity relationships. *J. Comput. Aid. Mol. Des.*, **6**, 349–383.
- Abe, I., Tatsumoto, H. and Hirashima, T. (1986) Prediction of activated carbon adsorption by adsorbability index (AI). *Suishitsu Odaku Kenkyu*, **9**, 153–161.
- Åberg, K.M. and Jacobsson, S.P. (2001) Pre-processing of three-way data by pulse-coupled neural networks – an imaging approach. *Chemom. Intell. Lab. Syst.*, **57**, 25–36.
- Aboushaaban, R.R., Alkhamees, H.A., Abouauda, H.S. and Simonelli, A.P. (1996) Atom level electrotopological state indexes in QSAR designing and testing antithyroid agents. *Pharm. Res.*, **13**, 129–136.
- Abraham, D.J. and Kellogg, G.E. (1993) Hydrophobic fields, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 506–520.
- Abraham, D.J. and Leo, A. (1987) Amino acid scale: hydrophobicity ($\Delta G_{1/2}$ cal). *Prot. Struct. Funct. Gen.*, **2**, 130–152.
- Abraham, M.H. (1993a) Application of solvation equations to chemical and biochemical processes. *Pure & Appl. Chem.*, **65**, 2503–2512.
- Abraham, M.H. (1993b) Hydrogen bonding. Part 31. Construction of a scale of solute effective or summation hydrogen bond basicity. *J. Phys. Org. Chem.*, **6**, 660–684.
- Abraham, M.H. (1993c) Physico-chemical and biological processes. *Chem. Soc. Rev.*, **22**, 73–83.
- Abraham, M.H. (1993d) Scales of solute hydrogen-bonding: their construction and application to physico-chemical and biochemical processes. *Chem. Soc. Rev.*, **22**, 73–83.
- Abraham, M.H., Andonian-Haftvan, J., Cometto Muniz, J.E. and Cain, W.S. (1996) An analysis of nasal irritation thresholds using a new solvation equation. *Fund. Appl. Toxicol.*, **31**, 71–76.
- Abraham, M.H., Andonian-Haftvan, J., Whiting, G.S., Leo, A. and Taft, R.S. (1994) Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. *J. Chem. Soc. Perkin Trans. 2*, 1777–1791.
- Abraham, M.H., Chadha, H.S., Dixon, J.P. and Leo, A.J. (1994a) Hydrogen bonding. 39. The partition of solutes between water and various alcohols. *J. Phys. Org. Chem.*, **7**, 712–716.
- Abraham, M.H., Chadha, H.S., Dixon, J.P., Rafols, C. and Treiner, C. (1995a) Hydrogen bonding. Part 40. Factors that influence the distribution of solutes between water and sodium dodecylsulfate micelles. *J. Chem. Soc. Perkin Trans. 2*, 887–894.
- Abraham, M.H., Chadha, H.S. and Mitchell, R.C. (1994b) Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. *J. Pharm. Sci.*, **83**, 1257–1268.
- Abraham, M.H., Chadha, H.S. and Mitchell, R.C. (1995b) Hydrogen bonding. Part 36. Determination of blood brain distribution using octanol–water partition coefficients. *Drug Design & Discovery*, **13**, 123–131.
- Abraham, M.H., Duce, P.P., Prior, D.V., Barratt, D.G., Morris, J.J. and Taylor, P.J. (1989) Hydrogen bonding. Part 9. Solute proton donor and proton acceptor scales for use in drug design. *J. Chem. Soc. Perkin Trans. 2*, 1355–1375.
- Abraham, M.H., Green, C.E. and Acree, W.E., Jr (2000) Correlation and prediction of the solubility of buckminsterfullerene in organic solvents; estimation of some physico-chemical properties. *J. Chem. Soc. Perkin Trans. 2*, 281–286.
- Abraham, M.H., Green, C.E., Acree, W.E., Jr, Hernandez, C.E. and Roy, L.E. (1998) Descriptors

- for solutes from the solubility of solids: *trans*-stilbene as an example. *J. Chem. Soc. Perkin Trans. 2*, 2677–2681.
- Abraham, M.H., Grellier, P.L., Hamerton, I., McGill, R.A., Prior, D.V. and Whiting, G.S. (1988) Solvation of gaseous non-electrolytes. *Faraday Discuss. Chem. Soc.*, **85**, 107–115.
- Abraham, M.H., Grellier, P.L. and McGill, R.A. (1987) Determination of olive oil–gas and hexadecane–gas partition coefficients, and calculation of the corresponding olive oil–water and hexadecane–water partition coefficients. *J. Chem. Soc. Perkin Trans. 2*, 797–803.
- Abraham, M.H., Grellier, P.L., Prior, D.V., Duce, P.P., Morris, J.J. and Taylor, P.J. (1989) Hydrogen bonding. Part 7. A scale of solute hydrogen-bond acidity based on log *K* values for complexation in tetrachloromethane. *J. Chem. Soc. Perkin Trans. 2*, 699–711.
- Abraham, M.H., Grellier, P.L., Prior, D.V., Morris, J.J. and Taylor, P.J. (1990) Hydrogen bonding. Part 10. A scale of solute hydrogen-bond basicity using log *K* values for complexation in tetrachloromethane. *J. Chem. Soc. Perkin Trans. 2*, 521–529.
- Abraham, M.H., Ibrahim, A. and Acree, W.E., Jr (2005) Air to blood distribution of volatile organic compounds: a linear free energy analysis. *Chem. Res. Toxicol.*, **18**, 904–911.
- Abraham, M.H., Ibrahim, A. and Zissimos, A.M. (2004) Determination of sets of solute descriptors from chromatographic measurements. *J. Chromat.*, **1037**, 29–47.
- Abraham, M.H., Ibrahim, A., Zissimos, A.M., Zhao, Y.H., Comer, J. and Reynolds, D.P. (2002) Application of hydrogen bonding calculations in property based drug design. *Drug Discov. Today*, **7**, 1056–1063.
- Abraham, M.H. and Le, J. (1999) The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.*, **88**, 868–880.
- Abraham, M.H., Lieb, W.R. and Franks, N.P. (1991) Role of hydrogen bonding in general anesthesia. *J. Pharm. Sci.*, **80**, 719–724.
- Abraham, M.H., Martins, F. and Mitchell, R.C. (1997) Algorithms for skin permeability using hydrogen bond descriptors: the problem of steroids. *J. Pharm. Pharmacol.*, **49**, 858–865.
- Abraham, M.H. and McGowan, J.C. (1987) The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography. *Chromatographia*, **23**, 243–246.
- Abraham, M.H. and Platts, J.A. (2001) Hydrogen bond structural group constants. *J. Org. Chem.*, **66**, 3484–3491.
- Abraham, M.H. and Rafols, C. (1995) Factors that influence tadpole narcosis. An LFER analysis. *J. Chem. Soc. Perkin Trans. 2*, 1843–1851.
- Abraham, M.H. and Whiting, G.S. (1992) Hydrogen bonding. XXI. Solvation parameters for alkylaromatic hydrocarbons from gas–liquid chromatographic data. *J. Chromat.*, **594**, 229–241.
- Abraham, M.H., Whiting, G.S., Alarie, Y., Morris, J.J., Taylor, P.J., Doherty, R.M., Taft, R.W. and Nielsen, G.D. (1990a) Hydrogen bonding. Part 12. A new QSAR for upper respiratory tract irritation by airborne chemicals in mice. *Quant. Struct. -Act. Relat.*, **9**, 6–10.
- Abraham, M.H., Whiting, G.S., Carr, P.W. and Ouyang, H. (1998) Hydrogen bonding. Part 45. The solubility of gases and vapours in methanol at 298 K: an LFER analysis. *J. Chem. Soc. Perkin Trans. 2*, 1385–1390.
- Abraham, M.H., Whiting, G.S., Doherty, R.M. and Shuely, W.J. (1990b) Hydrogen bonding. Part 13. A new method for the characterisation of GLC stationary phases – the Laffort data set. *J. Chem. Soc. Perkin Trans. 2*, 1451–1460.
- Abraham, M.H., Whiting, G.S., Doherty, R.M. and Shuely, W.J. (1991a) Hydrogen bonding. XVI. A new solute solvation parameter, π_2^H , from gas chromatographic data. *J. Chromat.*, **587**, 213–228.
- Abraham, M.H., Whiting, G.S., Doherty, R.M. and Shuely, W.J. (1991b) Hydrogen bonding. XVII. The characterisation of 24 gas–liquid chromatographic stationary phases studied by Poole and co-workers, including molten salts, and evaluation of solute–stationary phase interactions. *J. Chromat.*, **587**, 229–236.
- Abraham, R.J. and Smith, P.E. (1988) Charge calculations in molecular mechanics. IV. A general method for conjugated systems. *J. Comput. Chem.*, **9**, 288–297.
- Abrahamian, E., Fox, P.C., Nærum, L., Christensen, I. T., Thøgersen, H. and Clark, R.D. (2003) Efficient generation, storage, and manipulation of fully flexible pharmacophore multiplets and their use in 3D similarity searching. *J. Chem. Inf. Comput. Sci.*, **43**, 458–468.
- Abramowitz, R. and Yalkowsky, S.H. (1990) Estimation of aqueous solubility and melting point of PCB congeners. *Chemosphere*, **21**, 1221–1229.
- Absalan, G., Hemmateenejad, B., Soleimani, M., Akhond, M. and Miri, R. (2004) Quantitative structure–micellization relationship study of

- gemini surfactants using genetic-PLS and genetic-MLR. *QSAR Comb. Sci.*, **23**, 416–425.
- Acevedo-Martínez, J., Escalona-Arranz, J.C., Villar-Rojas, A., Téllez-Palmero, F., Pérez-Rosés, R., González, L. and Carrasco-Velaz, R. (2006) Quantitative study of the structure–retention index relationship in the imine family. *J. Chromat.*, **1102**, 238–244.
- Adams, N. and Schubert, U.S. (2004) From data to knowledge: chemical data management, data mining, and modeling in polymer science. *J. Comb. Chem.*, **6**, 12–23.
- Adamson, G.W., Lynch, M.F. and Town, W.G. (1971) Analysis of structural characteristics of chemical compounds in a large computer-based file. Part II. Atom-centred fragments. *J. Chem. Soc., C*, 3702–3706.
- ADAPT, Jurs, P.C., Pennsylvania State University, PA, <http://research.chem.psu.edu/pcjgroup/adapt.html>.
- Afantitis, A., Melagraki, G., Makridima, K., Alexandridis, A., Sarimveis, H. and Iglessi-Markopoulou, O. (2005) Prediction of high weight polymers glass transition temperature using RBF neural networks. *J. Mol. Struct. (Theochem)*, **716**, 193–198.
- Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P.A., Markopoulos, J. and Iglessi-Markopoulou, O. (2006) A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis. *Mol. Div.*, **10**, 405–414.
- Affolter, C., Baumann, K., Clerc, J.T., Schriber, H. and Pretsch, E. (1997) Automatic interpretation of infrared spectra. *Mikrochim. Acta*, **14**, 143–147.
- Afzelius, L., Masimirembwa, C.M., Karlén, A., Andersson, T.B. and Zamora, I. (2002) Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. *J. Comput. Aid. Mol. Des.*, **16**, 443–458.
- Agarwal, A., Pearson, P.P., Taylor, E.W., Li, H.B., Dahlgren, T., Herslof, M., Yang, Y.H., Lambert, G., Nelson, D.L., Regan, J.W. and Martin, A.R. (1993) Three dimensional quantitative structure–activity relationships of 5-HT receptor binding data for tetrahydropyridinylindole derivatives. A comparison of the Hansch and CoMFA methods. *J. Med. Chem.*, **36**, 4006–4014.
- Agarwal, K.K. (1998) An algorithm for computing the automorphism group of organic structures with stereochemistry and a measure of its efficiency. *J. Chem. Inf. Comput. Sci.*, **38**, 402–404.
- Agatonovic-Kustrin, S., Beresford, R. and Yusof, A.P.M. (2001) Theoretically derived molecular descriptors important in human intestinal absorption. *J. Pharm. Biomed. Anal.*, **25**, 227–237.
- Agrafiotis, D.K. (1997) On the use of information theory for assessing molecular diversity. *J. Chem. Inf. Comput. Sci.*, **37**, 576–580.
- Agrafiotis, D.K., Bandyopadhyay, D., Wegner, J.K. and van Vlijmen, H. (2007) Recent advances in chemoinformatics. *J. Chem. Inf. Model.*, **47**, 1279–1293.
- Agrafiotis, D.K., Cedeño, W. and Lobanov, V.S. (2002) On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.*, **42**, 903–911.
- Agrafiotis, D.K. and Lobanov, V.S. (1999) An efficient implementation of distance-based diversity measures based on *k-d* trees. *J. Chem. Inf. Comput. Sci.*, **39**, 51–58.
- Agrafiotis, D.K. and Rassokhin, D.N. (2002) A fractal approach for selecting an appropriate bin size for cell-based diversity estimation. *J. Chem. Inf. Comput. Sci.*, **42**, 117–122.
- Agrafiotis, D.K. and Xu, H. (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.*, **43**, 475–484.
- Agrawal, V.K., Bano, S. and Khadikar, P.V. (2003a) QSAR study on 5-lipoxygenase inhibitors using distance-based topological indices. *Bioorg. Med. Chem.*, **11**, 5519–5527.
- Agrawal, V.K., Bano, S. and Khadikar, P.V. (2003b) Topological approach to quantifying molecular lipophilicity of heterogeneous set of organic compounds. *Bioorg. Med. Chem.*, **11**, 4039–4047.
- Agrawal, V.K., Chaturvedi, S.C., Abraham, M.H. and Khadikar, P.V. (2003) QSAR study on tadpole narcosis. *Bioorg. Med. Chem.*, **11**, 4523–4533.
- Agrawal, V.K., Gupta, M., Singh, J. and Khadikar, P.V. (2005) A novel method of estimation of lipophilicity using distance-based topological indices: dominating role of equalized electronegativity. *Bioorg. Med. Chem.*, **13**, 2109–2120.
- Agrawal, V.K., Karmarkar, S. and Khadikar, P.V. (2002) QSAR study on competition binding of rodenticides (PATs) to H₁ receptor in rat and guinea pig brain. *Bioorg. Med. Chem.*, **10**, 2913–2918.
- Agrawal, V.K., Karmarkar, S., Khadikar, P.V. and Shrivastava, S. (2003) Use of distance-based topological indices in modeling antihypertensive activity: case of 2-aryl-imino-imidazolines. *Indian J. Chem.*, **42**, 1426–1435.
- Agrawal, V.K. and Khadikar, P.V. (2001) QSAR prediction of toxicity of nitrobenzenes. *Bioorg. Med. Chem.*, **9**, 3035–3040.

- Agrawal, V.K. and Khadikar, P.V. (2002) QSAR study on narcotic mechanism of action and toxicity: a molecular connectivity approach to *Vibrio fischeri* toxicity testing. *Bioorg. Med. Chem.*, **10**, 3517–3522.
- Agrawal, V.K., Sharma, R. and Khadikar, P.V. (2002) QSAR studies on carbonic anhydrase inhibitors: a case of ureido and thioureido derivatives of aromatic/heterocyclic sulfonamides. *Bioorg. Med. Chem.*, **10**, 2993–2999.
- Agrawal, V.K., Singh, J. and Khadikar, P.V. (2002) On the topological evidences for modeling lipophilicity. *Bioorg. Med. Chem.*, **10**, 3981–3996.
- Agrawal, V.K., Singh, J., Louis, B., Joshi, S., Joshi, A. and Khadikar, P.V. (2006) The topology of molecule and its lipophilicity. *Curr. Comput. -Aided Drug Des.*, **2**, 369–403.
- Agrawal, V.K., Singh, K. and Khadikar, P.V. (2004) QSAR studies on adenosine kinase inhibitors. *Med. Chem. Res.*, **13**, 479–496.
- Agrawal, V.K., Sohgaure, R. and Khadikar, P.V. (2002) QSAR studies on biological activity of piritrexim analogues against *pc* DHFR. *Bioorg. Med. Chem.*, **10**, 2919–2926.
- Agrawal, V.K., Srivastava, S. and Khadikar, P.V. (2004) QSAR study on phosphoramidothioate (Ace) toxicities in housefly. *Mol. Div.*, **8**, 413–419.
- Ahmad, P., Fyfe, C.A. and Mellors, A. (1975) Parachors in drug design. *Biochem. Pharmacol.*, **24**, 1103–1110.
- Ai, N., DeLisle, R.K., Yu, S.J. and Welsh, W.J. (2003) Computational models for predicting the binding affinities of ligands for the wild-type androgen receptor and a mutated variant associated with human prostate cancer. *Chem. Res. Toxicol.*, **16**, 1652–1660.
- Aihara, J. (1976) A generalized total π -energy index for a conjugated hydrocarbon. *J. Org. Chem.*, **41**, 2488–2490.
- Aihara, J. (1977a) Aromatic sextets and aromaticity in benzenoid hydrocarbons. *Bull. Chem. Soc. Jap.*, **50**, 2010–2012.
- Aihara, J. (1977b) Resonance energies of benzenoid hydrocarbons. *J. Am. Chem. Soc.*, **99**, 2048–2053.
- Aihara, J. (1978) Resonance energies of nonbenzenoid hydrocarbons. *Bull. Chem. Soc. Jap.*, **51**, 3540–3543.
- Aires-de-Sousa, J. (2003) Representation of molecular chirality, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1062–1078.
- Aires-de-Sousa, J. and Gasteiger, J. (2001) New description of molecular chirality and its application to the prediction of the preferred enantiomer in stereoselective reactions. *J. Chem. Inf. Comput. Sci.*, **41**, 369–375.
- Aires-de-Sousa, J. and Gasteiger, J. (2002) Prediction of enantiomeric selectivity in chromatography. Application of conformation-dependent and conformation-independent descriptors of molecular chirality. *J. Mol. Graph. Model.*, **20**, 373–388.
- Ajay, Walters, W.P. and Murcko, M.A. (1998) Can we learn to distinguish between “drug-like” and “nondrug-like” molecules? *J. Med. Chem.*, **41**, 3314–3324.
- Akagi, T., Mitani, S., Komyoji, T. and Nagatani, K. (1995) Quantitative structure–activity relationships of fluazinam and related fungicidal N-phenylpyridinamines preventive activity against *Botrytis cinerea*. *J. Pestic. Sci.*, **20**, 279–290.
- Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transaction of Automatic Control*, **AC-19**, 716–723.
- Al-Fahemi, J.H., Cooper, D.L. and Allan, N.L. (2005) The use of momentum–space descriptors for predicting octanol–water partition coefficients. *J. Mol. Struct. (Theochem)*, **727**, 57–61.
- Albahri, T.A. and George, R.S. (2003) Artificial neural network investigation of the structural group contribution method for predicting pure components auto ignition temperature. *Ind. Eng. Chem. Res.*, **42**, 5708–5714.
- Albert, R., Jeong, H. and Barabási, A.-L. (1999) Diameter of the World Wide Web. *Nature*, **401**, 130.
- Albuquerque, M.G., Hopfinger, A.J., Barreiro, E.J. and de Alencastro, R.B. (1998) Four-dimensional quantitative structure–activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A_2 receptor antagonists. *J. Chem. Inf. Comput. Sci.*, **38**, 925–938.
- Alifrangis, L.H., Christensen, I.T., Berglund, A., Sandberg, M., Hovgaard, L. and Frokjaer, S. (2000) Structure–property model for membrane partitioning of oligopeptides. *J. Med. Chem.*, **43**, 103–113.
- Alkhanidi, S. and Takahashi, Y. (2006) New molecular fragmental descriptors and their application to the prediction of fish toxicity. *MATCH Commun. Math. Comput. Chem.*, **55**, 205–232.
- Alkorta, I., Rozas, I. and Elguero, J. (1998) Bond length–electron density relationships: from covalent bonds to hydrogen bond interactions. *Struct. Chem.*, **9**, 243–247.
- Allen, B.C.P., Grant, G.H. and Richards, W.G. (2001) Similarity calculations using two-dimensional

- molecular representations. *J. Chem. Inf. Comput. Sci.*, **41**, 330–337.
- Allen, B.C.P., Grant, G.H. and Richards, W.G. (2003) Calculation of protein domain structural similarity using two-dimensional representations. *J. Chem. Inf. Comput. Sci.*, **43**, 134–143.
- Allen, D.M. (1971) Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469–475.
- Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.
- Allen, F.H., Bath, P.A. and Willett, P. (1995) Angular spectroscopy: rapid visualization of three-dimensional substructure dissimilarity using valence angle or torsional descriptors. *J. Chem. Inf. Comput. Sci.*, **35**, 261–271.
- Allerhand, A. and Schleyer, P.V.R. (1963a) Nitriles and isonitriles as proton acceptors in hydrogen bonding: correlation of $\Delta\nu_{\text{OH}}$ with acceptor structure. *J. Am. Chem. Soc.*, **85**, 866–870.
- Allerhand, A. and Schleyer, P.V.R. (1963b) Solvent effects in infrared spectroscopic studies of hydrogen bonding. *J. Am. Chem. Soc.*, **85**, 371–380.
- Allred, A. and Rochow, E.G. (1958) A scale of electronegativity based on electronic forces. *J. Inorg. Nuc. Chem.*, **5**, 264–268.
- Allred, A. and Rochow, E.G. (1961) Electronegativity values from thermochemical data. *J. Inorg. Nuc. Chem.*, **17**, 215–221.
- Almerico, A.M., Lauria, A., Tutone, M., Diana, P., Barraja, P., Montalbano, A., Cirrincione, G. and Dattolo, G. (2003) A multivariate analysis on non-nucleoside HIV-1 reverse transcriptase inhibitors and resistance induced by mutation. *QSAR Comb. Sci.*, **22**, 984–996.
- ALMOND, Ver. 2.0, Multivariate Infometric Analysis s.r.l., Viale dei Castagni 16, Perugia, Italy.
- Alsberg, B.K. (1990) Molecular reference (MOLREF), a new method in quantitative structure–activity relationships (QSAR). *Chemom. Intell. Lab. Syst.*, **8**, 173–181.
- Alsberg, B.K. (2000) Parsimonious multiscale classification models. *J. Chemom.*, **14**, 529–539.
- Alsberg, B.K., Woodward, A.M., Winson, M.K., Rowland, J.J. and Kell, D.B. (1998) Variable selection in wavelet regression models. *Anal. Chim. Acta*, **368**, 29–44.
- Altenburg, K. (1961) Zur Berechnung des Radius Verweigter Moleküle. *Kolloid Zeitschr.*, **178**, 112–117.
- Altenburg, K. (1980) Eine Bemerkung zu dem Randicschen “Molekularen Bindungs-Index”. *Z. Phys. Chemie (German)*, **261**, 389–393.
- Altomare, C., Carotti, A., Trapani, G. and Liso, G. (1997) Estimation of partitioning parameters of nonionic surfactants using calculated descriptors of molecular size, polarity, and hydrogen bonding. *J. Pharm. Sci.*, **86**, 1417–1425.
- Altomare, C., Carrupt, P.-A., Gaillard, P., El Tayar, N., Testa, B. and Carotti, A. (1992) Quantitative structure–metabolism relationship analyses of MAO-mediated toxication of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine and analogues. *Chem. Res. Toxicol.*, **5**, 366–375.
- Altomare, C., Cellamare, S., Carotti, A. and Ferappi, M. (1993) Linear solvation energy relationships in reversed-phase liquid chromatography. Examination in deltabond C₈ as stationary phase for measuring lipophilicity parameters. *Quant. Struct. -Act. Relat.*, **12**, 261–268.
- Altun, A., Kumru, M. and Dimoglo, A. (2001) The role of conformational and electronic parameters of thiosemicarbazone and thiosemicarbazide derivatives for their dermal toxicity. *J. Mol. Struct. (Theochem)*, **572**, 121–134.
- Alunni, S., Clementi, S., Edlund, U., Johnels, D., Hellberg, S., Sjöström, M. and Wold, S. (1983) Multivariate data analysis of substituent descriptors. *Acta Chem. Scand.*, **37**, 47–53.
- Alvarez-Ginarte, Y.M., Crespo, R., Montero-Cabrera, L.A., Ruiz-García, J.A., Marrero-Ponce, Y., Santana, R., Pardillo-Fontdevila, E. and Alonso-Becerra, E. (2005) A novel *in-silico* approach for QSAR studies of anabolic and androgenic activities in the 17 β -hydroxy-5 α -androstane steroid family. *QSAR Comb. Sci.*, **24**, 218–226.
- Alvarez-Ginarte, Y.M., Marrero-Ponce, Y., Ruiz-García, J.A., Montero-Cabrera, L.A., De La Vega, J.M.G., Marin, P.N., Crespo-Otero, R., Zaragoza, F.T. and García-Domenech, R. (2007) Applying pattern recognition methods plus quantum and physico-chemical molecular descriptors to analyze the anabolic activity of structurally diverse steroids. *J. Comput. Chem.*, **29**, 317–333.
- Alves, C.N., Pinheiro, J.C., Camargo, A.J., Ferreira, M.M.C., Romero, R.A.F. and da Silva, A.B.F. (2001) A multiple linear regression and partial least squares study of flavonoid compounds with anti-HIV activity. *J. Mol. Struct. (Theochem)*, **541**, 81–88.
- Amat, L., Besalú, E. and Carbó-Dorca, R. (2001) Identification of active molecular sites using quantum-self-similarity measures. *J. Chem. Inf. Comput. Sci.*, **41**, 978–991.
- Amat, L., Robert, D., Besalú, E. and Carbó-Dorca, R. (1998) Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.*, **38**, 624–631.

- Amboni, D., de, M.C., da Silva Junkes, B., Heinzen, V.E.F. and Yunes, R.A. (2002a) Semi-empirical topological method for prediction of the chromatographic retention esters. *J. Mol. Struct. (Theochem)*, **579**, 53–62.
- Amboni, D., de, M.C., da Silva Junkes, B., Yunes, R.A. and Heinzen, V.E.F. (2000) Quantitative structure–odor relationships of aliphatic esters using topological indices. *J. Agr. Food Chem.*, **48**, 3517–3521.
- Amboni, D., de, M.C., da Silva Junkes, B., Yunes, R.A. and Heinzen, V.E.F. (2002b) Quantitative structure–property relationship study of chromatographic retention indices and normal boiling point for oxo compounds using the semi-empirical topological method. *J. Mol. Struct. (Theochem)*, **586**, 71–80.
- Amić, D., Basak, S.C., Lučić, B., Nikolić, S. and Trinajstić, N. (2002) Structure–water solubility modeling of aliphatic alcohols using the weighted path numbers. *SAR & QSAR Environ. Res.*, **13**, 281–295.
- Amić, D., Beslo, D., Lučić, B., Nikolić, S. and Trinajstić, N. (1998) The vertex-connectivity index revisited. *J. Chem. Inf. Comput. Sci.*, **38**, 819–822.
- Amić, D., Davidović-Amić, D., Beslo, D., Lučić, B. and Trinajstić, N. (1997) The use of the ordered orthogonalized multivariate linear regression in a structure–activity study of coumarin and flavonoid derivatives as inhibitors of aldose reductase. *J. Chem. Inf. Comput. Sci.*, **37**, 581–586.
- Amić, D., Davidović-Amić, D., Beslo, D., Lučić, B. and Trinajstić, N. (1998) QSAR of flavylum salts as inhibitors of xanthine oxidase. *J. Chem. Inf. Comput. Sci.*, **38**, 815–818.
- Amić, D., Davidović-Amić, D., Beslo, D., Lučić, B. and Trinajstić, N. (1999) Prediction of pK values, half-lives, and electronic spectra of flavylum salts from molecular structure. *J. Chem. Inf. Comput. Sci.*, **39**, 967–973.
- Amić, D., Davidović-Amić, D., Bešlo, D., Rastija, V., Lučić, B. and Trinajstić, N. (2007) SAR and QSAR of the antioxidant activity of flavonoids. *Curr. Med. Chem.*, **14**, 827–845.
- Amić, D., Davidović-Amić, D., Beslo, D. and Trinajstić, N. (2003) Structure–radical scavenging activity relationships of flavonoids. *Croat. Chem. Acta*, **76**, 55–61.
- Amić, D., Davidović-Amić, D., Jurić, A., Lučić, B. and Trinajstić, N. (1995a) Structure–activity correlation of flavone derivatives for inhibition of cAMP phosphodiesterase. *J. Chem. Inf. Comput. Sci.*, **35**, 1034–1038.
- Amić, D., Davidović-Amić, D. and Trinajstić, N. (1995b) Calculation of retention times of anthocyanins with orthogonalized topological indices. *J. Chem. Inf. Comput. Sci.*, **35**, 136–139.
- Amić, D., Lučić, B., Nikolić, S. and Trinajstić, N. (2001) Predicting inhibition of microsomal p-hydroxylation of aniline by aliphatic alcohols: a QSAR approach based on the weighted path numbers. *Croat. Chem. Acta*, **74**, 237–250.
- Amić, D. and Trinajstić, N. (1995) On the Detour matrix. *Croat. Chem. Acta*, **68**, 53–62.
- Amidon, G.L., Yalkowsky, S.H., Anik, S.T. and Valvani, S.C. (1975) Solubility of nonelectrolytes in polar solvents. V. Estimation of the solubility of aliphatic monofunctional compounds in water using a molecular surface area approach. *J. Phys. Chem.*, **79**, 2239–2246.
- Amini, A., Muggleton, S.H., Lodhi, H. and Sternberg, M.J.E. (2007) A novel logic-based approach for quantitative toxicology prediction. *J. Chem. Inf. Model.*, **47**, 998–1006.
- Amoore, J.E. (1964) Current status of the steric theory of odor. *Ann. N. Y. Acad. Sci.*, **116**, 457–476.
- AMPAC, Semichem, Inc., Kansas City, MO.
- Andersson, P., Haglund, P., Rappe, C. and Tysklind, M. (1996) Ultraviolet absorption characteristic and calculated semi-empirical parameters as chemical descriptors in multivariate modelling of polychlorinated biphenyls. *J. Chemom.*, **10**, 171–185.
- Andersson, P.L., Maran, U., Fara, D., Karelson, M. and Hermens, J.L.M. (2002) General and class specific models for prediction of soil sorption using various physico-chemical descriptors. *J. Chem. Inf. Comput. Sci.*, **42**, 1450–1459.
- Andersson, P.M. and Lundstedt, T. (2002) Hierarchical experimental design exemplified by QSAR evaluation of a chemical library directed towards the melanocortin 4 receptor. *J. Chemom.*, **16**, 490–496.
- Andersson, P.M., Sjöström, M. and Lundstedt, T. (1998) Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemom. Intell. Lab. Syst.*, **42**, 41–50.
- Andersson, P.M., Sjöström, M., Wold, S. and Lundstedt, T. (2000) Comparison between physico-chemical and calculated molecular descriptors. *J. Chemom.*, **14**, 629–642.
- Andre, V., Boissart, C., Sichel, F., Gauduchon, P., Le Talaer, J.Y., Lancelot, J.C., Mercier, C., Chemtob, S., Raoult, E. and Tallec, A. (1997) Mutagenicity of nitro- and amino-substituted carbazoles in *Salmonella typhimurium*. III. Methylated derivatives of 9H-carbazole. *Mut. Res.*, **389**, 247–260.

- Andrea, T.A. (1995) Novel structure–activity insights from neural network models. *ACS Symp. Ser.*, **606**, 282–287.
- Andrea, T.A. and Kalayeh, H. (1991) Applications of neural networks in quantitative structure–activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem.*, **34**, 2824–2836.
- Andreazza Costa, M.C., Gaudio, A.C. and Takahata, Y. (1997) A comparative study of principal component and linear multiple regression analysis in SAR and QSAR applied to 1,4-dihydropyridine calcium channel antagonists (nifedipine analogues). *J. Mol. Struct. (Theochem)*, **394**, 291–300.
- Andreazza Costa, M.C., Soares Barata, L.E. and Takahata, Y. (1995) SAR analysis of synthetic neolignans and related compounds which are anti-leishmaniasis active compounds using pattern recognition methods. *J. Mol. Struct. (Theochem)*, **340**, 185–192.
- Andrews, D.F. (1972) Plots of high-dimensional data. *Biometrics*, **28**, 125–136.
- Andrews, P.R. (1993) Drug–receptor interactions, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 13–40.
- Andrews, P.R., Craik, D.J. and Martin, J.L. (1984) Functional group contributions to drug–receptor interactions. *J. Med. Chem.*, **27**, 1648–1657.
- Anker, L.S., Jurs, P.C. and Edwards, P.A. (1990) Quantitative structure–retention relationship studies of odor-active aliphatic compounds with oxygen-containing functional groups. *Anal. Chem.*, **62**, 2676–2684.
- Anwair, M.A., Károlyházy, L., Szabó, D., Balogh, B., Kövesdi, I., Harmat, V., Krenyác, J., Gellért, A., Takács-Novák, K. and Mátyus, P. (2003) Lipophilicity of aminopyridazinone regioisomers. *J. Agr. Food Chem.*, **51**, 5262–5270.
- Anzali, S., Barnickel, G., Cezanne, B., Krug, M., Filimonov, D. and Poroikov, V.V. (2001) Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS). *J. Med. Chem.*, **44**, 2432–2437.
- Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M., Gasteiger, J. and Polanski, J. (1996) The comparison of geometric and electronic properties of molecular surfaces by neural networks: application to the analysis of corticosteroid binding globulin activity of steroids. *J. Comput. Aid. Mol. Des.*, **10**, 521–534.
- Anzali, S., Barnickel, G., Krug, M., Wagener, M. and Gasteiger, J. (1997) Kohonen neural network: a novel approach to search for bioisosteric groups, in *Computer-Assisted Lead Finding and Optimization* (eds H. van de Waterbeemd, B. Testa and G. Folkers), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 95–106.
- Anzali, S., Gasteiger, J., Holzgrabe, U., Polanski, J., Sadowski, J., Teckentrup, A. and Wagener, M. (1998a) The use of self-organizing neural networks in drug design, in *3D QSAR in Drug Design*, Vol. 2 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 273–299.
- Anzali, S., Gasteiger, J., Holzgrabe, U., Polanski, J., Teckentrup, A. and Wagener, M. (1998b) The use of self-organizing neural networks in drug design. *Persp. Drug Disc. Des.*, **9/10/11**, 273–299.
- Anzini, M., Cappelli, A., Vomero, S., Seeber, M., Menziani, M.C., Langer, T., Hagen, B., Manzoni, C. and Bourguignon, J.-J. (2001) Mapping and fitting the peripheral benzodiazepine receptor binding site by carboxamide derivatives. Comparison of different approaches to quantitative ligand–receptor interaction modeling. *J. Med. Chem.*, **44**, 1134–1150.
- Aoyama, T., Suzuki, Y. and Ichikawa, H. (1990a) Neural networks applied to quantitative structure–activity relationships analysis. *J. Med. Chem.*, **33**, 2583–2590.
- Aoyama, T., Suzuki, Y. and Ichikawa, H. (1990b) Neural networks applied to structure–activity relationships. *J. Med. Chem.*, **33**, 905–908.
- Aptula, A.O., Kühne, R., Ebert, R.-U., Cronin, M.T.D., Netzeva, T.I. and Schüürmann, G. (2003) Modeling discrimination between antibacterial and non-antibacterial activity based on 3D molecular descriptors. *QSAR Comb. Sci.*, **22**, 113–128.
- Aptula, A.O., Netzeva, T.I., Valkova, I.V., Cronin, M.T.D., Schultz, T.W., Kühne, R. and Schüürmann, G. (2002) Multivariate discrimination between modes of toxic action of phenols. *Quant. Struct. -Act. Relat.*, **21**, 12–22.
- Aptula, A.O., Patlewicz, G. and Roberts, D.W. (2005) Skin sensitization: reaction mechanistic applicability domains for structure–activity relationships. *Chem. Res. Toxicol.*, **18**, 1420–1426.
- Aptula, A.O., Roberts, D.W. and Cronin, M.T.D. (2005a) From experiment to theory: molecular orbital parameters to interpret the skin sensitization potential of 5-chloro-2-methylisothiazol-3-one and 2-methylisothiazol-3-one. *Chem. Res. Toxicol.*, **18**, 324–329.
- Aptula, A.O., Roberts, D.W., Cronin, M.T.D. and Schultz, T.W. (2005b) Chemistry–toxicity relationships for the effects of di- and trihydroxybenzenes to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.*, **18**, 844–854.

- Åqvist, J. and Tapia, O. (1987) Surface fractality as a guide for studying protein–protein interactions. *J. Mol. Graph.*, **5**, 30–34.
- Arakawa, M., Hasegawa, K. and Funatsu, K. (2006) QSAR study of anti-HIV HEPTanalogues based on multi-objective genetic programming and counter-propagation neural network. *Chemom. Intell. Lab. Syst.*, **83**, 91–98.
- Araujo, O. and De La Peña, J.A. (1998) Some bounds for the connectivity index of a chemical graph. *J. Chem. Inf. Comput. Sci.*, **38**, 827–831.
- Araujo, O. and Morales, D.A. (1996a) A theorem about the algebraic structure underlying orthogonal graph invariants. *J. Chem. Inf. Comput. Sci.*, **36**, 1051–1053.
- Araujo, O. and Morales, D.A. (1996b) An alternative approach to orthogonal graph theoretical invariants. *Chem. Phys. Lett.*, **257**, 393–396.
- Araujo, O. and Morales, D.A. (1998) Properties of new orthogonal graph theoretical invariants in structure–property correlations. *J. Chem. Inf. Comput. Sci.*, **38**, 1031–1037.
- Araujo, O. and Rada, J. (2000) Randić index and lexicographic order. *J. Math. Chem.*, **27**, 201–212.
- Arcos, J.C. (1987) Structure–activity relationships: criteria for predicting carcinogenic activity of chemical compounds. *Environ. Sci. Technol.*, **21**, 743–745.
- Ardelan, M., Katona, G., Hopartean, I. and Diudea, M.V. (2001) Cluj and Szeged indices in property modeling. *Studia Univ. Babes-Bolyai*, **45**, 81–95.
- Argese, E., Bettiol, C., Giurin, G. and Miana, P. (1999) Quantitative structure–activity relationships for the toxicity of chlorophenols to mammalian submitochondrial particles. *Chemosphere*, **38**, 2281–2292.
- Argese, E., Bettiol, C., Volpi Ghirardini, A., Fasolo, M., Giurin, G. and Ghetti, F. (1998) Comparison of *in vitro* submitochondrial particle and Microtox® assays for determining the toxicity of organotin compounds. *Environ. Toxicol. Chem.*, **17**, 1005–1012.
- Ariëns, E.J. (1979) *Drug Design*, Vol. VII, Academic Press, New York.
- Ariëns, E.J. (1992) QSAR conceptions and misconceptions. *Quant. Struct. -Act. Relat.*, **11**, 190–194.
- Arimoto, S., Spivakovsky, M., Ohno, H., Zizler, P., Taylor, K.F., Yamabe, T. and Mezey, P.G. (2001) Structural analysis of certain linear operators representing chemical network systems via the existence and uniqueness theorems of spectral resolution. VI. *Int. J. Quant. Chem.*, **84**, 389–400.
- Aringhieri, R., Hansen, P. and Malucelli, F. (2001) A linear algorithm for the hyper-Wiener index of chemical trees. *J. Chem. Inf. Comput. Sci.*, **41**, 958–963.
- Armstrong, D.R., Perkins, P.G. and Stewart, J.J.P. (1973) Bond indices and valency. *J. C. S. Dalton Trans.*, 838–840.
- Arnaud, L., Taillandier, G., Kaouadji, M., Ravel, P. and Tissut, M. (1994) Photosynthesis inhibition by phenylureas: a QSAR approach. *Ecotox. Environ. Safety*, **28**, 121–133.
- Arnold, S.F. (1990) *Mathematical Statistics*, Prentice-Hall, Englewood Cliffs, NJ, p. 636.
- Arroio, A., Honório, K.M. and da Silva, A.B.F. (2004) A theoretical study on the analgesic activity of cannabinoid compounds. *J. Mol. Struct. (Theochem)*, **709**, 223–229.
- Arteca, G.A. (1991) Molecular shape descriptors, in *Reviews in Computational Chemistry*, Vol. 9 (eds K.B. Lipkowitz and D. Boyd), VCH Publishers, New York, pp. 191–253.
- Arteca, G.A. (1996) Different molecular size scaling regimes for inner and outer regions of proteins. *Phys. Rev. A*, **54**, 3044–3047.
- Arteca, G.A. (1999) Path-integral calculation of the mean number of overcrossings in an entangled polymer network. *J. Chem. Inf. Comput. Sci.*, **39**, 550–557.
- Arteca, G.A. (2003a) A measure of folding complexity for *D*-dimensional polymers. *J. Chem. Inf. Comput. Sci.*, **43**, 63–67.
- Arteca, G.A. (2003b) Analysis of shape transitions using molecular size descriptors associated with inner and outer regions of a polymer chain. *J. Mol. Struct. (Theochem)*, **630**, 113–123.
- Arteca, G.A., Jammal, V.B. and Mezey, P.G. (1988a) Shape group studies of molecular similarity and regioselectivity in chemical reactions. *J. Comput. Chem.*, **9**, 608–619.
- Arteca, G.A., Jammal, V.B., Mezey, P.G., Yadav, J.S., Hermsmeier, M.A. and Gund, T.M. (1988b) Shape group studies of molecular similarity: relative shapes of van der Waals and electrostatic potential surfaces of nicotinic agonists. *J. Mol. Graph.*, **6**, 45–53.
- Arteca, G.A. and Mezey, P.G. (1987) A method for the characterization of molecular conformations. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **14**, 133–147.
- Arteca, G.A. and Mezey, P.G. (1988a) Molecular conformations and molecular shape: a discrete characterization of continua of van der Waals surfaces. *Int. J. Quant. Chem.*, **34**, 517–526.
- Arteca, G.A. and Mezey, P.G. (1988b) Shape characterization of some molecular model surfaces. *J. Comput. Chem.*, **9**, 554–563.
- Arteca, G.A. and Mezey, P.G. (1989) Shape group theory of van der Waals surfaces. *J. Math. Chem.*, **3**, 43.

- Arteca, G.A. and Mezey, P.G. (1990) A method for the characterization of foldings in protein ribbon models. *J. Mol. Graph.*, **8**, 66–80.
- Artemenko, N.V., Baskin, I.I., Palyulin, V.A. and Zefirov, N.S. (2003) Artificial neural network and fragmental approach in prediction of physico-chemical properties of organic compounds. *Russ. Chem. Bull.*, **52**, 20–29.
- Artemi, C. and Balaban, A.T. (1987) Mathematical modeling of polymers. Part II. Irreducible sequences in *n*-ary copolymers. *MATCH Commun. Math. Comput. Chem.*, **22**, 33–66.
- Arulmozhiraja, S. and Morita, M. (2004) Structure–activity relationships for the toxicity of polychlorinated dibenzofurans: approach through density functional theory-based descriptors. *Chem. Res. Toxicol.*, **17**, 348–356.
- Arupiyoti, S. and Iragavarapu, S. (1998) New electrotopological descriptor for prediction of boiling points of alkanes and aliphatic alcohols through artificial neural network and multiple linear regression analysis. *Computers Chem.*, **22**, 515–522.
- Aschi, M., D'Archivio, A.A., Maggi, M.A., Mazzeo, P. and Ruggieri, F. (2007) Quantitative structure–retention relationships of pesticides in reversed-phase high-performance liquid chromatography. *Anal. Chim. Acta*, **582**, 235–242.
- Ash, J.E., Warr, W.A. and Willett, P. (eds) (1991) *Chemical Structure Systems*, Ellis Horwood, Chichester, UK.
- Ashby, J. (1985) Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Envir. Mutag.*, **7**, 919–921.
- Ashby, J. and Tennant, R.W. (1988) Chemical structure, *Salmonella* mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the US NCI/NTP. *Mut. Res.*, **204**, 17–115.
- Ashrafi, A.R. and Hamadani, M. (2005) Symmetry properties of some chemical graphs. *Croat. Chem. Acta*, **78**, 159–163.
- Ashrafi, A.R. and Manoochehrian, B. (2006) On the PI polynomial of a graph. *Utilitas Mathematica*, **71**, 97–108.
- Ashton, M., Barnard, J., Casset, F., Charlton, M., Downs, G., Gorse, D., Holliday, J.D., Lahana, R. and Willett, P. (2002) Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quant. Struct. Act. Relat.*, **21**, 598–604.
- Ashton, M.J., Jaye, M.C. and Mason, J.S. (1996) New perspectives in lead generation II: evaluating molecular diversity. *Drug Discov. Today*, **1**, 71–78.
- Asikainen, A.H., Ruuskanen, J. and Tuppurainen, K.A. (2004) Consensus kNN QSAR: a versatile method for predicting the estrogenic activity of organic compounds *in silico*. A comparative study with five estrogen receptors, and a large, diverse set of ligands. *Environ. Sci. Technol.*, **38**, 6724–6729.
- Asikainen, A.H., Ruuskanen, J. and Tuppurainen, K.A. (2005) Alternative QSAR models for selected estradiol and cytochrome P450 ligands: comparison between classical, spectroscopic, CoMFA and GRID/GOLPE methods. *SAR & QSAR Environ. Res.*, **16**, 555–565.
- Assefa, H., Kamath, S. and Buolamwini, J.K. (2003) 3D-QSAR and docking studies on 4-anilinoquinazoline and 4-anilinoquinoline epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors. *J. Comput. Aid. Mol. Des.*, **17**, 475–493.
- Atkinson, A.C. (1985) *Plots, Transformations, and Regression*, Clarendon Press, Oxford, UK, pp. 282.
- Atkinson, R. (1987) A structure–activity relationships for the estimation of rate constants for the gas-phase reactions of OH radicals with organic compounds. *Int. J. Chem. Kinet.*, **19**, 799–828.
- Atkinson, R. (1988) Estimation of gas-phase hydroxyl radical rate constants for organic compounds. *Environ. Toxicol. Chem.*, **7**, 435–442.
- Attias, R. and Dubois, J.-E. (1990) Substructure systems: concepts and classifications. *J. Chem. Inf. Comput. Sci.*, **30**, 2–7.
- Attias, R. and Petitjean, M. (1993) Statistical analysis of atom topological neighborhoods and multivariate representations of a large chemical file. *J. Chem. Inf. Comput. Sci.*, **33**, 649–656.
- Audry, E., Dallet, Ph., Langlois, M.H., Colleter, J.C. and Dubost, J.P. (1989) Quantitative structure–affinity relationships in a series of α_2 adrenergic amines using the molecular lipophilicity potential. *Proc. Clin. Biol. Res.*, **291**, 63.
- Audry, E., Dubost, J.P., Colleter, J.C. and Dallet, Ph. (1986) Une Nouvelle Approche des Relations Structure–Activité: le “Potentiel de Lipophilie Moléculaire”. *Eur. J. Med. Chem.*, **21**, 71–72.
- Audry, E., Dubost, J.P., Langlois, M.H., Croizet, F., Braquet, P., Dallet, Ph. and Colleter, J.C. (1992) Use of molecular lipophilicity potential in QSAR, in *QSAR Design of Bioactive Compounds* (ed. M. Kuchar), Prous Science, Barcelona, Spain, pp. 249–268.
- Augen, J. (2002) The evolving role of information technology in the drug discovery process. *Drug Discov. Today*, **7**, 315–323.
- Aureli, L., Cruciani, G., Cesta, M.C., Anacardio, R., De Simone, L. and Moriconi, A. (2005) Predicting human serum albumin affinity of interleukin-8

- (CXCL8) inhibitors by 3D-QSPR approach. *J. Med. Chem.*, **48**, 2469–2479.
- Aurenhammer, F. (1991) Voronoi diagrams – a survey of fundamental geometric data structure. *ACM Comp. Serv.*, **23**, 345–405.
- Austel, V. (1983) Features and problems of practical drug design, in *Steric Effects in Drug Design, Topics in Current Chemistry*, Vol. 114 (eds M. Charton and I. Motoc), Springer-Verlag, Berlin, Germany, pp. 7–19.
- Austel, V. (1995) Experimental design in synthesis planning and structure–property correlations, in *Experimental Design. Chemometrics Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 49–62.
- Austel, V., Kutter, E. and Kalbfleisch, W. (1979) A new easily accessible steric parameter for structure–activity relationships. *Arzneim. Forsch. (German)*, **29**, 585–587.
- Awad, H.M., Boersma, M.G., Boeren, S., van Bladeren, P.J., Vervoort, J. and Rietjens, I.M.C.M. (2001) Structure–activity study on the quinone/quinone methide chemistry of flavonoids. *Chem. Res. Toxicol.*, **14**, 398–408.
- Azzaoui, K. and Morinallory, L. (1995) Quantitative structure–retention relationships for the investigation of the retention mechanism in high performance liquid chromatography using apolar eluent with a very low content of polar modifiers. *Chromatographia*, **40**, 690–696.
- Babic, D., Balaban, A.T. and Klein, D.J. (1995) Nomenclature and coding of fullerenes. *J. Chem. Inf. Comput. Sci.*, **35**, 515–526.
- Babic, D., Brinkmann, G. and Dress, A. (1997) Topological resonance energy of fullerenes. *J. Chem. Inf. Comput. Sci.*, **37**, 920–923.
- Babic, D., Graovac, A. and Gutman, I. (1991) On a resonance energy model based on expansion in terms of acyclic moments: exact results. *Theor. Chim. Acta*, **79**, 403–411.
- Babic, D., Klein, D.J., Lukovits, I., Nikolić, S. and Trinajstić, N. (2002) Resistance–distance matrix: a computational algorithm and its application. *Int. J. Quant. Chem.*, **90**, 166–176.
- Babu, M.A., Shakya, N., Prathipati, P., Kaskhedikar, S.G. and Saxena, A.K. (2002) Development of 3D-QSAR models for 5-lipoxygenase antagonists: chalcones. *Bioorg. Med. Chem.*, **10**, 4035–4041.
- Bacha, P.A., Gruver, H.S., Den Hartog, B.K., Tamura, S.Y. and Nutt, R.F. (2002) Rule extraction from a mutagenicity data set using adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.*, **42**, 1104–1111.
- Baczek, T. and Kaliszan, R. (2002) Combination of linear solvent strength model and quantitative structure–retention relationships as a comprehensive procedure of approximate prediction of retention in gradient liquid chromatography. *J. Chromat.*, **962**, 41–55.
- Baczek, T. and Kaliszan, R. (2003) Predictive approaches to gradient retention based on analyte structural descriptors from calculation chemistry. *J. Chromat.*, **987**, 29–37.
- Baczek, T., Kaliszan, R., Novotná, K. and Jandera, P. (2005) Comparative characteristics of HPLC columns based on quantitative structure–retention relationships (QSRR) and hydrophobic-subtraction model. *J. Chromat.*, **1075**, 109–115.
- Bader, R.F.W. (1990) *Atoms in Molecules: A Quantum Theory*, Oxford University Press, Oxford, UK, p. 458.
- Bader, R.F.W. (2003) Letter to the editor: quantum mechanics, or orbitals? *Int. J. Quant. Chem.*, **94**, 173–177.
- Bader, R.F.W., Nguyendang, T.T. and Tal, Y. (1981) A topological theory of molecular structure. *Rep. Prog. Phys.*, **44**, 893–948.
- Badertscher, M., Bsichofberger, K., Mink, M.E. and Pretsch, E. (2001) A novel formalism to characterize the degree of unsaturation of organic molecules. *J. Chem. Inf. Comput. Sci.*, **41**, 889–893.
- Badreddin Abolmaali, S.F., Wegner, J.K. and Zell, A. (2003) The compressed feature matrix – a fast method for feature based substructure search. *J. Mol. Model.*, **9**, 235–241.
- Bagchi, M.C., Maiti, B.C. and Bose, S. (2004a) QSAR of antituberculosis drugs of INH type using graphical invariants. *J. Mol. Struct. (Theochem)*, **679**, 179–186.
- Bagchi, M.C., Maiti, B.C., Mills, D. and Basak, S.C. (2004b) Usefulness of graphical invariants in quantitative structure–activity correlations of tuberculostatic drugs of the isonicotinic acid hydrazide type. *J. Mol. Model.*, **10**, 102–111.
- Bahnick, D.A. and Doucette, W.J. (1988) Use of molecular connectivity indices to estimate soil sorption for organic chemicals. *Chemosphere*, **17**, 1703–1715.
- Bai, F. and Wang, T. (2005) A 2-D graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.*, **413**, 458–462.
- Baird, N.C. (1969) Calculation of “Dewar” resonance energies in conjugated organic molecules. *Can. J. Chem.*, **47**, 3535–3538.
- Baird, N.C. (1971) Dewar resonance energy. *J. Chem. Educ.*, **48**, 509–514.

- Bajaj, S., Sambi, S.S., Gupta, S. and Madan, A.K. (2006a) Model for prediction of anti-HIV activity of 2-pyridinone derivatives using novel topological descriptors. *QSAR Comb. Sci.*, **25**, 813–823.
- Bajaj, S., Sambi, S.S. and Madan, A.K. (2004a) Predicting anti-HIV activity of phenethylthiazolethiourea (PETT) analogs: computational approach using Wiener's topochemical index. *J. Mol. Struct. (Theochem)*, **684**, 197–203.
- Bajaj, S., Sambi, S.S. and Madan, A.K. (2004b) Prediction of carbonic anhydrase activation by tri-/tetrasubstituted-pyridinium-azole compounds: a computational approach using novel topochemical descriptor. *QSAR Comb. Sci.*, **23**, 506–514.
- Bajaj, S., Sambi, S.S. and Madan, A.K. (2005) Prediction of anti-inflammatory activity of N-arylanthranilic acids: computational approach using refined Zagreb indices. *Croat. Chem. Acta*, **78**, 165–174.
- Bajaj, S., Sambi, S.S. and Madan, A.K. (2006b) Models for prediction of anti-neoplastic activity of 1,2-bis(sulfonyl)-1-methylhydrazines: computational approach using Wiener's indices. *MATCH Commun. Math. Comput. Chem.*, **55**, 193–204.
- Bajorath, J. (2001) Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.*, **41**, 233–245.
- Bajzer, Ž., Randić, M., Plavšić, D. and Basak, S.C. (2003) Novel map descriptors for characterization of toxic effects in proteomics maps. *J. Mol. Graph. Model.*, **22**, 1–9.
- Baker, F.W., Parish, R.C. and Stock, L.M. (1967) Dissociation constants of bicyclo[2.2.2]oct-2-ene-1-carboxylic acids, dibenzobicyclo[2.2.2]octa-2,5-diene-1-carboxylic acids, and cubanecarboxylic acids. *J. Am. Chem. Soc.*, **89**, 5677–5685.
- Baker, J.K. (1979) Estimation of high pressure liquid chromatographic retention indices. *Anal. Chem.*, **51**, 1693–1697.
- Baker, J.R., Mihelčić, J.R. and Sabljčić, A. (2001) Reliable QSAR for estimating K_{oc} from persistent organic pollutants: correlation with molecular connectivity indices. *Chemosphere*, **45**, 213–221.
- Bakken, G.A. and Jurs, P.C. (1999a) Prediction of hydroxyl radical rate constants from molecular structure. *J. Chem. Inf. Comput. Sci.*, **39**, 1064–1075.
- Bakken, G.A. and Jurs, P.C. (1999b) Prediction of methyl radical addition rate constants from molecular structure. *J. Chem. Inf. Comput. Sci.*, **39**, 508–514.
- Bakken, G.A. and Jurs, P.C. (2001) QSARs for 6-azasteroids as inhibitors of human type 15 α -reductase: prediction of binding affinity and selectivity relative to 3-BHSD. *J. Chem. Inf. Comput. Sci.*, **41**, 1255–1265.
- Balaban, A.T. (1969) Chemical graphs. VII. Proposed nomenclature of branched cata-condensed benzenoid polycyclic hydrocarbons. *Tetrahedron*, **25**, 2949–2956.
- Balaban, A.T. (1970a) Chemical graphs. X (Aromaticity. VIII). Resonance energies of cata-condensed benzenoid polycyclic hydrocarbons. *Rev. Roum. Chim.*, **15**, 1243–1250.
- Balaban, A.T. (1970b) Chemical graphs. XI (Aromaticity. IX). Isomerism and topology of non-branched cata-condensed polycyclic conjugated non-benzenoid hydrocarbons. *Rev. Roum. Chim.*, **15**, 1251–1262.
- Balaban, A.T. (1971) Chemical graphs. XII. Configurations of annulenes. *Tetrahedron*, **27**, 6115–6131.
- Balaban, A.T. (1972) Chemical graphs. XVII (Aromaticity. X). Cata-condensed polycyclic hydrocarbons which fulfil the Hückel rule but lack closed electronic shells. *Rev. Roum. Chim.*, **17**, 1531–1543.
- Balaban, A.T. (1975) Some applications of graph theory. *MATCH Commun. Math. Comput. Chem.*, **1**, 33–60.
- Balaban, A.T. (ed.) (1976a) *Chemical Applications of Graph Theory*, Academic Press, New York, p. 390.
- Balaban, A.T. (1976b) Chemical graphs. XXVI. Codes for configurations of conjugated polyenes. *Rev. Roum. Chim.*, **21**, 1045–1047.
- Balaban, A.T. (1976c) Chemical graphs. XXVII. Enumeration and codification of staggered conformations of alkanes. *Rev. Roum. Chim.*, **21**, 1049–1071.
- Balaban, A.T. (1976d) Enumeration of cyclic graphs, in *Chemical Applications of Graph Theory* (ed. A.T. Balaban), Academic Press, London, UK, pp. 63–105.
- Balaban, A.T. (1977) Chemical graphs. XXVIII. A new topological index for catafusenes: L-transform of their three-digit codes. *Rev. Roum. Chim.*, **22**, 45–47.
- Balaban, A.T. (1978a) Chemical graphs. XXXII. Constitutional and steric isomers of substituted cycloalkanes. *Croat. Chem. Acta*, **51**, 35–42.
- Balaban, A.T. (1978b) Mathematical principles in chemistry. *Noesis*, **4**, 15–22.
- Balaban, A.T. (1979) Chemical graphs. XXXIV. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta*, **53**, 355–375.

- Balaban, A.T. (1982) Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **89**, 399–404.
- Balaban, A.T. (1983a) Topological indices based on topological distances in molecular graphs. *Pure & Appl. Chem.*, **55**, 199–206.
- Balaban, A.T. (1983b) Topological indices: what they are and what they can do. Proceedings of the Second Balkan Chemistry Days, Varna.
- Balaban, A.T. (1984) Numerical and non-numerical methods in chemistry: present and future. *ACM SIGSAM Bull.*, **18**, 29–30.
- Balaban, A.T. (1985a) Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.*, **25**, 334–343.
- Balaban, A.T. (1985b) Graph theory and theoretical chemistry. *J. Mol. Struct. (Theochem)*, **120**, 117–142.
- Balaban, A.T. (1985c) *Symbolic Computation and Chemistry. EUROCAL-85* (ed. B. Buchberger), Springer, Berlin, Germany, pp. 68–79.
- Balaban, A.T. (1986a) Chemical graphs. 48. Topological index *J* for heteroatom-containing molecules taking into account periodicities of element properties. *MATCH Commun. Math. Comput. Chem.*, **21**, 115–122.
- Balaban, A.T. (1986b) Symmetry in chemical structures and reactions, in *Symmetry Unifying Human Understanding* (ed. I. Hargittai), Pergamon Press, New York, pp. 999–1020.
- Balaban, A.T. (1987) Numerical modelling of chemical structures: local graph invariants and topological indices, in *Graph Theory and Topology in Chemistry* (eds R.B. King and D.H. Rouvray), Elsevier, Amsterdam, The Netherlands, pp. 159–176.
- Balaban, A.T. (1988a) Chemical graphs. Part 49. Open problems in the area of condensed polycyclic benzenoids: topological stereoisomers of coronoids and congeners. *Rev. Roum. Chim.*, **33**, 699–707.
- Balaban, A.T. (1988b) Topological indices and their uses: a new approach for coding of alkanes. *J. Mol. Struct. (Theochem)*, **165**, 243–253.
- Balaban, A.T. (1991) Enumeration of Isomers, in *Chemical Graph Theory. Introduction and Fundamentals* (eds D. Bonchev and D.E. Rouvray), Abacus Press/Gordon and Breach Science Publishers, New York, pp. 177–234.
- Balaban, A.T. (1992) Using real numbers as vertex invariants for third-generation topological indexes. *J. Chem. Inf. Comput. Sci.*, **32**, 23–28.
- Balaban, A.T. (1993a) Benzenoid catafusenes: perfect matchings, isomerization, automerization. *Pure & Appl. Chem.*, **65**, 1–9.
- Balaban, A.T. (1993b) Confessions and reflections of a graph-theoretical chemist. *MATCH Commun. Math. Comput. Chem.*, **29**, 3–17.
- Balaban, A.T. (1993c) Lowering the intra- and intermolecular degeneracy of topological invariants. *Croat. Chem. Acta*, **66**, 447–458.
- Balaban, A.T. (1993d) Prediction of physical properties from chemical structures, in *Recent Advances in Chemical Information* (ed. H. Collier), Royal Society of Chemistry, Cambridge, UK, pp. 301–317.
- Balaban, A.T. (1993e) Solved and unsolved problems in chemical graph theory. *Ann. Disc. Math.*, **55**, 109–126.
- Balaban, A.T. (1994a) Local vs. global (i.e., atomic versus molecular) numerical modeling of molecular graphs. *J. Chem. Inf. Comput. Sci.*, **34**, 398–402.
- Balaban, A.T. (1994b) Reaction graphs, in *Graph Theoretical Approaches to Chemical Reactivity* (eds D. Bonchev and O. Mekenyan), Kluwer, Dordrecht, The Netherlands, pp. 137–180.
- Balaban, A.T. (1994c) Real-number local (atomic) invariants and global (molecular) topological indices. *Rev. Roum. Chim.*, **39**, 245–257.
- Balaban, A.T. (1995a) Chemical graphs: looking back and a glimpsing ahead. *J. Chem. Inf. Comput. Sci.*, **35**, 339–350.
- Balaban, A.T. (1995b) Local (atomic) and global (molecular) graph-theoretical descriptors. *SAR & QSAR Environ. Res.*, **3**, 81–95.
- Balaban, A.T. (1997a) From chemical graphs to 3D molecular modeling, in *From Chemical Topology to Three-Dimensional Geometry* (ed. A.T. Balaban), Plenum Press, New York, pp. 1–24.
- Balaban, A.T. (1997b) From chemical topology to 3D geometry. *J. Chem. Inf. Comput. Sci.*, **37**, 645–650.
- Balaban, A.T. (ed.) (1997c) *From Chemical Topology to Three-Dimensional Geometry*, Plenum Press, New York, 420.
- Balaban, A.T. (1998) Topological and stereochemical molecular descriptors for databases useful in QSAR. Similarity/dissimilarity and drug design. *SAR & QSAR Environ. Res.*, **8**, 1–21.
- Balaban, A.T. (2001) A personal view about topological indices for QSAR/QSPR, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 1–30.
- Balaban, A.T. and Artemi, C. (1987) Mathematical modeling of polymers. Part I. Enumeration of non redundant (irreducible) repeating sequences in stereoregular polymers, elastomers, or in binary copolymers. *MATCH Commun. Math. Comput. Chem.*, **22**, 3–32.

- Balaban, A.T. and Artemi, C. (1989) Chemical graphs. Part 51. Enumeration of nonbranched catafusenes according to the numbers of benzenoid rings in the catafusene and its longest linearly condensed portion. *Polycycl. Aromat. Comp.*, **1**, 171–189.
- Balaban, A.T. and Artemi, C. (1998) Mathematical modeling of polymers. 3. Enumeration and generation of repeating irreducible sequences in linear bi-, ter-, quater-, and quinquenary copolymers and in stereoregular homopolymers. *Macromol. Chem.*, **189**, 863–870.
- Balaban, A.T. and Balaban, T.-S. (1991) New vertex invariants and topological indices of chemical graphs based on information on distances. *J. Math. Chem.*, **8**, 383–397.
- Balaban, A.T. and Balaban, T.-S. (1992) Correlation using topological indices based on real graph invariants. *J. Chim. Phys.*, **89**, 1735–1745.
- Balaban, A.T., Basak, S.C., Beteringhe, A., Mills, D. and Supuran, C.T. (2004) QSAR study using topological indices for inhibition of carbonic anhydrase II by sulfanilamides and Schiff bases. *Mol. Div.*, **8**, 401–412.
- Balaban, A.T., Basak, S.C., Colburn, T. and Grunwald, G.D. (1994) Correlation between structure and normal boiling points of haloalkanes C_1 – C_4 using neural networks. *J. Chem. Inf. Comput. Sci.*, **34**, 1118–1121.
- Balaban, A.T., Bertelsen, S. and Basak, S.C. (1994) New centric topological indexes for acyclic molecules (trees) and substituents (rooted trees), and coding of rooted trees. *MATCH Commun. Math. Comput. Chem.*, **30**, 55–72.
- Balaban, A.T., Beteringhe, A., Constantinescu, T., Filip, P.A. and Ivanciuc, O. (2007) Four new topological indices based on the molecular path code. *J. Chem. Inf. Model.*, **47**, 716–731.
- Balaban, A.T., Biermann, D. and Schmidt, W. (1985) Dualist graph approach for correlating Diels–Alder reactivities of polycyclic aromatic hydrocarbons. *Nouv. J. Chim.*, **9**, 443–449.
- Balaban, A.T., Bonchev, D. and Seitz, W.A. (1993) Topological/chemical distances and graph centers in molecular graphs with multiple bonds. *J. Mol. Struct. (Theochem)*, **280**, 253–260.
- Balaban, A.T., Brunvoll, J., Cioslowski, J., Cyvin, B.N., Syvin, S.J., Gutman, I., Wenchen, H., Wenjie, H., von Knop, J., Kovacevic, M., Müller, W.R., Szymanski, K., Tosic, R. and Trinajstić, N. (1987) Enumeration of benzenoid and coronoid hydrocarbons. *Z. Naturforsch.*, **42**, 863–870.
- Balaban, A.T. and Catana, C. (1993) Search for nondegenerate real vertex invariants and derived topological indexes. *J. Comput. Chem.*, **14**, 155–160.
- Balaban, A.T. and Catana, C. (1994) New topological indices for substituents (molecular fragments). *SAR & QSAR Environ. Res.*, **2**, 1–16.
- Balaban, A.T., Catana, C., Dawson, M. and Niculescu-Duvaz, I. (1990) Applications of weighted topological index J for QSAR of carcinogenesis inhibitors (retinoic acid derivatives). *Rev. Roum. Chim.*, **35**, 997–1003.
- Balaban, A.T., Chiriac, A., Motoc, I. and Simon, Z. (1980) *Steric Fit in Quantitative Structure–Activity Relations*, Springer-Verlag, Berlin, Germany, p. 77.
- Balaban, A.T., Ciubotariu, D. and Ivanciuc, O. (1990) Design of topological indices. Part 2. Distance measure connectivity indices. *MATCH Commun. Math. Comput. Chem.*, **25**, 41–70.
- Balaban, A.T., Ciubotariu, D. and Medeleanu, M. (1991) Topological indices and real vertex invariants based on graph eigenvalues or eigenvectors. *J. Chem. Inf. Comput. Sci.*, **31**, 517–523.
- Balaban, A.T. and Diudea, M.V. (1993) Real number vertex invariants: regressive distance sums and related topological indices. *J. Chem. Inf. Comput. Sci.*, **33**, 421–428.
- Balaban, A.T. and Feroiu, V. (1990) Correlation between structure and critical data of vapor pressures of alkanes by means of topological indices. *Rep. Mol. Theory*, **1**, 133–139.
- Balaban, A.T. and Filip, P.A. (1984) Computer program for topological index J (average distance sum connectivity). *MATCH Commun. Math. Comput. Chem.*, **16**, 163–190.
- Balaban, A.T., Filip, P.A. and Balaban, T.-S. (1985) Computer program for finding all possible cycles in graphs. *J. Comput. Chem.*, **6**, 316–329.
- Balaban, A.T. and Harary, F. (1968) Chemical graphs. V. Enumeration and proposed nomenclature of benzenoid catacondensed polycyclic aromatic hydrocarbons. *Tetrahedron*, **24**, 2505–2516.
- Balaban, A.T. and Harary, F. (1971) The characteristic polynomial does not uniquely determine the topology of a molecule. *J. Chem. Doc.*, **11**, 258–259.
- Balaban, A.T. and Harary, F. (1976) Early history of the interplay between graph theory and chemistry, in *Chemical Applications of Graph Theory* (ed. A.T. Balaban), Academic Press, London, UK, pp. 1–4.
- Balaban, A.T., Ionescu-Pallas, N. and Balaban, T.-S. (1985) Asymptotic values of topological indices J and J' (average distance sum connectivities) for infinite acyclic and cyclic graphs. *MATCH Commun. Math. Comput. Chem.*, **17**, 121–146.
- Balaban, A.T. and Ivanciuc, O. (1989) FORTRAN-77 computer program for calculating topological index J for molecules containing heteroatoms, in

- MATH/CHEM/COMP 1988 (ed. A. Graovac), Elsevier, Amsterdam, The Netherlands, pp. 193–211.
- Balaban, A.T. and Ivanciuc, O. (1999) Historical development of topological indices, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 21–57.
- Balaban, A.T., Joshi, N., Kier, L.B. and Hall, L.H. (1992) Correlations between chemical structure and normal boiling points of halogenated alkanes C_1 – C_4 . *J. Chem. Inf. Comput. Sci.*, **32**, 233–237.
- Balaban, A.T., Kennedy, J.W. and Quintas, L.V. (1988) The number of alkanes having n carbons and a longest chain of length d . An application of a theorem of Polya. *J. Chem. Educ.*, **65**, 304–313.
- Balaban, A.T., Kier, L.B. and Joshi, N. (1992a) Correlations between chemical structure and normal boiling points of acyclic ethers, peroxides, acetals, and their sulfur analogues. *J. Chem. Inf. Comput. Sci.*, **32**, 237–244.
- Balaban, A.T., Kier, L.B. and Joshi, N. (1992b) Structure–property analysis of octane numbers for hydrocarbons (alkanes, cycloalkanes, alkenes). *MATCH Commun. Math. Comput. Chem.*, **28**, 13–27.
- Balaban, A.T., Liu, X., Cyvin, S.J. and Klein, D.J. (1993) Benzenoids with maximum Kekulé structure counts for given numbers of hexagons. *J. Chem. Inf. Comput. Sci.*, **33**, 429–436.
- Balaban, A.T., Liu, X., Klein, D.J., Babic, D., Schmalz, T.G., Seitz, W.A. and Randić, M. (1995) Graph invariants for fullerenes. *J. Chem. Inf. Comput. Sci.*, **35**, 396–404.
- Balaban, A.T., Mekenyan, O. and Bonchev, D. (1985a) Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). I. Algorithms for finding graph orbits and canonical numbering of atoms. *J. Comput. Chem.*, **6**, 538–551.
- Balaban, A.T., Mekenyan, O. and Bonchev, D. (1985b) Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). III. Topological, chemical, and stereochemical coding of molecular structure. *J. Comput. Chem.*, **6**, 552–569.
- Balaban, A.T., Mills, D. and Basak, S.C. (1999) Correlation between structure and normal boiling points of acyclic carbonyl compounds. *J. Chem. Inf. Comput. Sci.*, **39**, 758–764.
- Balaban, A.T., Mills, D. and Basak, S.C. (2002) Alkane ordering as a criterion for similarity between topological indices: index J as “sharpened Wiener index”. *MATCH Commun. Math. Comput. Chem.*, **45**, 5–26.
- Balaban, A.T., Mills, D., Ivanciuc, O. and Basak, S.C. (2000) Reverse Wiener indices. *Croat. Chem. Acta*, **73**, 923–941.
- Balaban, A.T., Mills, D., Kodali, V. and Basak, S.C. (2006) Complexity of chemical graphs in terms of size, branching, and cyclicity. *SAR & QSAR Environ. Res.*, **17**, 429–450.
- Balaban, A.T. and Motoc, I. (1979) Chemical graphs. XXXVI. Correlations between octane numbers and topological indices of alkanes. *MATCH Commun. Math. Comput. Chem.*, **5**, 197–218.
- Balaban, A.T., Motoc, I., Bonchev, D. and Mekenyan, O. (1983) Topological indices for structure–activity correlations, in *Steric Effects in Drug Design, Topics in Current Chemistry*, Vol. 114 (eds M. Charton and I. Motoc), Springer-Verlag, Berlin, Germany, pp. 21–55.
- Balaban, A.T., Niculescu-Duvaz, I. and Simon, Z. (1987) Topological aspects in QSAR for biologically active molecules. *Acta Pharm. Jugosl.*, **37**, 7–36.
- Balaban, A.T. and Quintas, L.V. (1983) The smallest graphs, trees and 4-trees with degenerate topological index J . *MATCH Commun. Math. Comput. Chem.*, **14**, 213–233.
- Balaban, A.T. and Rouvray, D.E. (1980) Graph-theoretical analysis of the bonding topology in polyhedral organic cations. *Tetrahedron*, **36**, 1851–1855.
- Balaban, A.T. and Tomescu, I. (1984) Chemical graphs. XL. Three relations between the Fibonacci sequence and the numbers of Kekulé structures for non-branched cata-condensed polycyclic aromatic hydrocarbons. *Croat. Chem. Acta*, **57**, 391–404.
- Balaban, A.T. and Tomescu, I. (1985) Chemical graphs. XLI. Numbers of conjugated circuits and Kekulé structures for zigzag catafusenes and (j, k) -hexes; generalized Fibonacci numbers. *MATCH Commun. Math. Comput. Chem.*, **17**, 91–120.
- Balaban, A.T. and Tomescu, I. (1988) Alternating 6-cycles in perfect matchings of graphs representing condensed benzenoid hydrocarbons, in *Application of Graphs in Chemistry and Physics* (eds J.W. Kennedy and L.V. Quintas), North-Holland, Amsterdam, The Netherlands, pp. 5–16.
- Balaban, T.-S., Balaban, A.T. and Bonchev, D. (2001) A topological approach to predicting properties of infinite polymers. Part VI. Rational formulas for the normalized Wiener index and a comparison with index J . *J. Mol. Struct. (Theochem)*, **535**, 81–92.
- Balaban, T.-S., Filip, P. and Ivanciuc, O. (1992) Computer generation of acyclic graphs based on local vertex invariants and topological indices.

- Derived canonical labelling and coding of trees and alkanes. *J. Math. Chem.*, **11**, 79–105.
- Balakin, K.V., Savchuk, N.P. and Tetko, I.V. (2006) *In silico* approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. *Curr. Med. Chem.*, **13**, 223–241.
- Balakin, K.V., Tkachenko, S.E., Lang, S.A., Okun, I., Ivashchenko, A.A. and Savchuk, N.P. (2002) Property-based design of GPCR-targeted library. *J. Chem. Inf. Comput. Sci.*, **42**, 1332–1342.
- Balasubramanian, K. (1982) Spectra of chemical trees. *Int. J. Quant. Chem.*, **21**, 581–590.
- Balasubramanian, K. (1984a) Computer generation of the characteristic polynomial of chemical graph. *J. Comput. Chem.*, **5**, 387–394.
- Balasubramanian, K. (1984b) The use of frame's method for the characteristic polynomials of chemical graphs. *Theor. Chim. Acta*, **65**, 49–58.
- Balasubramanian, K. (1985) Applications of combinatorics and graph theory to spectroscopy and quantum chemistry. *Chem. Rev.*, **85**, 599–618.
- Balasubramanian, K. (1990) Geometry-dependent characteristic polynomials of molecular structures. *Chem. Phys. Lett.*, **169**, 224–228.
- Balasubramanian, K. (1991) Graph theory and the PPP method. *J. Math. Chem.*, **7**, 353–362.
- Balasubramanian, K. (1994) Integration of graph theory and quantum chemistry for structure–activity relationships. *SAR & QSAR Environ. Res.*, **2**, 59–77.
- Balasubramanian, K. (1995a) Computer generation of nuclear equivalence classes based on the three-dimensional molecular structure. *J. Chem. Inf. Comput. Sci.*, **35**, 243–250.
- Balasubramanian, K. (1995b) Geometry-dependent connectivity indices for the characterization of molecular structures. *Chem. Phys. Lett.*, **235**, 580–586.
- Balasubramanian, K. and Basak, S.C. (1998) Characterization of isospectral graphs using graph invariants and derived orthogonal parameters. *J. Chem. Inf. Comput. Sci.*, **38**, 367–373.
- Balasubramanian, K., Kaufmann, J.J., Koski, W.S. and Balaban, A.T. (1980) Graph theoretical characterization and computer generation of certain carcinogenic benzenoid hydrocarbons and identification of bay regions. *J. Comput. Chem.*, **1**, 149–157.
- Balasubramanian, K. and Randić, M. (1982) The characteristic polynomials of structures with pending bonds. *Theor. Chim. Acta*, **61**, 307–323.
- Balawender, R., Komorowski, L., De Proft, F. and Geerlings, P. (1998) Derivatives of molecular valence as a measure of aromaticity. *J. Phys. Chem. A*, **102**, 9912–9917.
- Balàz, S., Wiese, M., Chi, H.-L. and Seydel, J.K. (1990) Subcellular pharmacokinetics and quantitative structure/time/activity relationships. *Anal. Chim. Acta*, **235**, 195–207.
- Baldi, P., Benz, R.W., Hirschberg, D.S. and Swamidass, S.J. (2007) Lossless compression of chemical fingerprints using integer entropy codes improves storage and retrieval. *J. Chem. Inf. Model.*, **47**, 2098–2109.
- Balinska, K.T., Gargano, M.L. and Quintas, L.V. (2001) Two models for random graphs with bounded degree. *Croat. Chem. Acta*, **74**, 207–223.
- Balogh, T. and Naray-Szabo, G. (1993) Application of the average molecular electrostatic field in quantitative structure–activity relationships. *Croat. Chem. Acta*, **66**, 129–140.
- Bangov, I.P. (1988) Use of the ^{13}C -NMR chemical shift/charge density linear relationship for recognition and ranking of chemical structures. *Anal. Chim. Acta*, **209**, 29.
- Bangov, I.P. (1990) Computer-assisted structure generation from a gross formula. 3. Alleviation of the combinatorial problem. *J. Chem. Inf. Comput. Sci.*, **30**, 277–289.
- Bangov, I.P. (1992) Toward the solution of the isomorphism problem in generation of chemical graphs: generation of benzenoid hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **32**, 167–173.
- Bangov, I.P. (2003) Topological structure generators, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 178–194.
- Barakat, N.A.M., Jiang, J.-H., Liang, Y.-Z. and Yu, R.-Q. (2004) Piece-wise quasi-linear modeling in QSAR and analytical calibration based on linear substructures detected by genetic algorithm. *Chemom. Intell. Lab. Syst.*, **72**, 73–82.
- Barbe, J. (1983) Convenient relations for the estimation of bond ionicity in A–B type compounds. *J. Chem. Educ.*, **60**, 640–642.
- Barbosa, F. and Horvath, D. (2004) Molecular similarity and property similarity. *Curr. Top. Med. Chem.*, **4**, 589–600.
- Barker, E.J., Gardiner, E.J., Gillet, V.J., Kitts, P. and Morris, J. (2003) Further development of reduced graphs for identifying bioactive compounds. *J. Chem. Inf. Comput. Sci.*, **43**, 346–356.
- Barlow, T.W. (1995) Self-organizing maps and molecular similarity. *J. Mol. Graph.*, **13**, 24–27.
- Barnard, J.M. (1993) Substructure searching methods: old and new. *J. Chem. Inf. Comput. Sci.*, **33**, 532–538.

- Barnard, J.M. (1994) Third international conference on chemical structures. *J. Chem. Inf. Comput. Sci.*, **34**, 1–2.
- Barnard, J.M. (2003) Representation of molecular structures – overview, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 27–50.
- Barnard, J.M. and Downs, G.M. (1997) Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.*, **37**, 141–142.
- Barnard, J.M., Lynch, M.F. and Welford, S.M. (1982) Computer storage and retrieval of generic structures in chemical patents. 4. An extended connection table representation for generic structures. *J. Chem. Inf. Comput. Sci.*, **22**, 160–164.
- Barone, P.M.V.B., Braga, R.S., Camilo, A., Jr and Galvão, D.S. (2000) Electronic indices from semi-empirical calculations to identify carcinogenic activity of polycyclic aromatic hydrocarbons. *J. Mol. Struct. (Theochem)*, **505**, 55–66.
- Barone, P.M.V.B., Camilo, A., Jr and Galvão, D.S. (1996) Theoretical approach to identify carcinogenic activity of polycyclic aromatic hydrocarbons. *Phys. Rev. Lett.*, **77**, 1186–1189.
- Barone, R. and Chanon, M. (2001) A new and simple approach to chemical complexity. Application to the synthesis of natural products. *J. Chem. Inf. Comput. Sci.*, **41**, 269–272.
- Baroni, M., Clementi, S., Cruciani, G., Costantino, G., Riganelli, D. and Oberrauch, E. (1992) Predictive ability of regression models. Part II. Selection of the best predictive PLS model. *J. Chemom.*, **6**, 347–356.
- Baroni, M., Clementi, S., Cruciani, G., Kettaneh-Wold, N. and Wold, S. (1993) D-optimal designs in QSAR. *Quant. Struct. - Act. Relat.*, **12**, 225–231.
- Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S. (1993a) Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct. - Act. Relat.*, **12**, 9–20.
- Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S. (1993b) GOLPE: an advanced chemometric tool for 3D QSAR problems, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 256–259.
- Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F. and Mason, J.S. (2007) A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *J. Chem. Inf. Model.*, **47**, 279–294.
- Barratt, M.D., Basketter, D.A. and Roberts, D.W. (1994) Skin sensitization structure–activity relationships for phenyl benzoates. *Toxicol. Vitro*, **8**, 823–826.
- Barroso, J.M. and Besalú, E. (2005) Design of experiments applied to QSAR: ranking a set of compounds and establishing a statistical significance test. *J. Mol. Struct. (Theochem)*, **727**, 89–96.
- Bartell, L.S. (1963) Resonance energies from Pauling bond orders. *J. Phys. Chem.*, **67**, 1865–1868.
- Bartell, L.S. (1964) Semiquantitative theory of resonance using Pauling bond orders. *Tetrahedron*, **20**, 139–153.
- Barysz, M., Bonchev, D. and Mekenyan, O. (1986) Graph-centre, self-returning walks, and critical pressure of alkanes. *MATCH Commun. Math. Comput. Chem.*, **20**, 125–140.
- Barysz, M., Jashari, G., Lall, R.S., Srivastava, A.K. and Trinajstić, N. (1983) On the distance matrix of molecules containing heteroatoms, in *Chemical Applications of Topology and Graph Theory* (ed. R.B. King), Elsevier, Amsterdam, The Netherlands, pp. 222–230.
- Barysz, M., Nikolić, S. and Trinajstić, N. (1986) A note on the characteristic polynomial. *MATCH Commun. Math. Comput. Chem.*, **19**, 117–126.
- Barysz, M., Plavšić, D. and Trinajstić, N. (1986) A note on topological indices. *MATCH Commun. Math. Comput. Chem.*, **19**, 89–116.
- Barysz, M. and Trinajstić, N. (1984) A novel approach to the characterization of chemical structures. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **18**, 661–673.
- Barysz, M., Trinajstić, N. and von Knop, J. (1983) On the similarity of chemical structures. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **17**, 441–451.
- Barysz, M., von Knop, J., Pajaković, S. and Trinajstić, N. (1985) Characterization of branching. *Pol. J. Chem.*, **59**, 405–432.
- Basak, S.C. (1987) Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.*, **15**, 605–609.
- Basak, S.C. (1988) Binding of barbiturates to cytochrome P₄₅₀: a QSAR study using log *P* and topological indices. *Med. Sci. Res.*, **16**, 281–282.
- Basak, S.C. (1990) A nonempirical approach to predicting molecular properties using graph-theoretic invariants, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 83–103.
- Basak, S.C. (1999) Information theoretic indices of neighborhood complexity and their applications, in

- Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 563–593.
- Basak, S.C., Balaban, A.T., Grunwald, G.D. and Gute, B.D. (2000) Topological indices: their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.*, **40**, 891–898.
- Basak, S.C., Bertelsen, S. and Grunwald, G.D. (1994) Application of graph theoretical parameters in quantifying molecular similarity and structure–activity relationships. *J. Chem. Inf. Comput. Sci.*, **34**, 270–276.
- Basak, S.C., Bertelsen, S. and Grunwald, G.D. (1995) Use of graph theoretic parameters in risk assessment of chemicals. *Toxicol. Lett.*, **79**, 239–250.
- Basak, S.C., Frane, C.M., Rosen, M.E. and Magnuson, V.R. (1987) Molecular topology and acute toxicity: a QSAR study of monoketones. *Med. Sci. Res.*, **15**, 887–888.
- Basak, S.C., Gieschen, D.P., Harriss, D.K. and Magnuson, V.R. (1983) Physico-chemical and topological correlates of the enzymatic acetyltransfer reaction. *J. Pharm. Sci.*, **72**, 934–937.
- Basak, S.C., Gieschen, D.P. and Magnuson, V.R. (1984) A quantitative correlation of the LC₅₀ values of esters in *Pimephales promelas* using physico-chemical and topological parameters. *Environ. Toxicol. Chem.*, **3**, 191–199.
- Basak, S.C., Gieschen, D.P., Magnuson, V.R. and Harriss, D.K. (1982) Structure–activity relationships and pharmacokinetics: a comparative study of hydrophobicity, van der Waals' volume and topological parameters. *Med. Sci. Res.*, **10**, 619–620.
- Basak, S.C. and Grunwald, G.D. (1993) Use of graph invariants, volume and total surface area in predicting boiling point of alkanes. *Math. Modelling and Sci. Computing*, **2**, 735–740.
- Basak, S.C. and Grunwald, G.D. (1994a) Molecular similarity and risk assessment: analog selection and property estimation using graph invariants. *SAR & QSAR Environ. Res.*, **2**, 289–307.
- Basak, S.C. and Grunwald, G.D. (1994b) Use of topological space and property space in selecting structural analogs. *Math. Modelling and Sci. Computing*, **4**, 464–469.
- Basak, S.C. and Grunwald, G.D. (1995a) Estimation of lipophilicity from structural similarity. *New J. Chem.*, **19**, 231–237.
- Basak, S.C. and Grunwald, G.D. (1995b) Molecular similarity and estimation of molecular properties. *J. Chem. Inf. Comput. Sci.*, **35**, 366–372.
- Basak, S.C. and Grunwald, G.D. (1995c) Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study. *Chemosphere*, **31**, 2529–2546.
- Basak, S.C. and Grunwald, G.D. (1995d) Tolerance space and molecular similarity. *SAR & QSAR Environ. Res.*, **3**, 265–277.
- Basak, S.C., Grunwald, G.D., Gute, B.D., Balasubramanian, K. and Opitz, D. (2000) Use of statistical and neural net approaches in predicting toxicity of chemicals. *J. Chem. Inf. Comput. Sci.*, **40**, 885–890.
- Basak, S.C., Grunwald, G.D., Host, G.E., Niemi, G.J. and Bradbury, S.P. (1998) A comparative study of molecular similarity, statistical, and neural methods for predicting toxic modes of action. *Environ. Toxicol. Chem.*, **6**, 1056–1064.
- Basak, S.C., Grunwald, G.D. and Niemi, G.J. (1997) Use of graph-theoretic and geometrical molecular descriptors in structure–activity relationships, in *From Chemical Topology to Three-Dimensional Geometry* (ed. A.T. Balaban), Plenum Press, New York, pp. 73–116.
- Basak, S.C. and Gute, B.D. (1997) Characterization of molecular structures using topological indices. *SAR & QSAR Environ. Res.*, **7**, 1–21.
- Basak, S.C., Gute, B.D. and Balaban, A.T. (2004) Interrelationship of major topological indices evidenced by clustering. *Croat. Chem. Acta*, **77**, 331–344.
- Basak, S.C., Gute, B.D. and Drewes, L.R. (1996a) Predicting blood–brain transport of drugs: a computational approach. *Pharm. Res.*, **13**, 775–778.
- Basak, S.C., Gute, B.D. and Ghatak, S. (1999a) Prediction of complement-inhibitory activity of benzamides using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, **39**, 255–260.
- Basak, S.C., Gute, B.D. and Grunwald, G.D. (1996b) A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.*, **36**, 1054–1060.
- Basak, S.C., Gute, B.D. and Grunwald, G.D. (1996c) Estimation of the normal boiling points of haloalkanes using molecular similarity. *Croat. Chem. Acta*, **69**, 1159–1173.
- Basak, S.C., Gute, B.D. and Grunwald, G.D. (1997) Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, **37**, 651–655.
- Basak, S.C., Gute, B.D. and Grunwald, G.D. (1998a) Characterization of the molecular similarity of chemicals using topological invariants, in *Advances in Molecular Similarity*, Vol. 2 (eds R. Carbó-Dorca

- and G. Mezey), JAI Press, Inc., Stanford, CO, pp. 171–185.
- Basak, S.C., Gute, B.D. and Grunwald, G.D. (1998b) Relative effectiveness of topological, geometrical and quantum chemical parameters in estimating mutagenicity of chemicals, in *Quantitative Structure–Activity Relationships in Environmental Sciences*, Vol. 7 (eds F. Chen and G. Schüürmann), SETAC Press, Pensacola, FL, pp. 245–261.
- Basak, S.C., Gute, B.D. and Grunwald, G.D. (1999b) A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 675–696.
- Basak, S.C., Gute, B.D. and Grunwald, G.D. (1999c) Assessment of the mutagenicity of aromatic amines from theoretical structural parameters: a hierarchical approach. *SAR & QSAR Environ. Res.*, **10**, 117–129.
- Basak, S.C., Gute, B.D. and Mills, D. (2006) Similarity methods in analog selection, property estimation and clustering of diverse chemicals. *ARKIVOC*, (ix), 157–210.
- Basak, S.C., Gute, B.D., Mills, D. and Hawkins, D.M. (2003) Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: a comparison of arbitrary versus tailored similarity spaces. *J. Mol. Struct. (Theochem)*, **622**, 127–145.
- Basak, S.C., Harriss, D.K. and Magnuson, V.R. (1984) Comparative study of lipophilicity versus topological molecular descriptors in biological correlations. *J. Pharm. Sci.*, **73**, 429–437.
- Basak, S.C. and Magnuson, V.R. (1983) Molecular topology and narcosis. A quantitative structure–activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim. Forsch. (German)*, **33**, 501–503.
- Basak, S.C., Magnuson, V.R., Niemi, G.J. and Regal, R.R. (1988) Determining structural similarity of chemicals using graph-theoretic indices. *Disc. Appl. Math.*, **19**, 17–44.
- Basak, S.C., Magnuson, V.R. and Veith, G.D. (1987) Topological indices: their nature, mutual relatedness, and applications, in *Mathematical Modelling in Science and Technology* (eds X.J.R. Avula, G. Leitmann, C.D. Mote, Jr and E.Y. Rodin), Pergamon Press, Oxford, UK, pp. 300–305.
- Basak, S.C. and Mills, D. (2001a) Prediction of mutagenicity utilizing a hierarchical QSAR approach. *SAR & QSAR Environ. Res.*, **12**, 481–496.
- Basak, S.C. and Mills, D. (2001b) Quantitative structure–property relationships (QSPRs) for the estimation of vapor pressure: a hierarchical approach using mathematical structural descriptors. *J. Chem. Inf. Comput. Sci.*, **41**, 692–701.
- Basak, S.C. and Mills, D. (2001c) Use of mathematical structural invariants in the development of QSAR models. *MATCH Commun. Math. Comput. Chem.*, **44**, 15–30.
- Basak, S.C. and Mills, D. (2005) Prediction of partitioning properties for environmental pollutants using mathematical structural descriptors. *ARKIVOC*, (ii), 60–76.
- Basak, S.C., Mills, D., El-Masri, H.A., Mumtaz, M.M. and Hawkins, D.M. (2004) Predicting blood:air partition coefficients using theoretical molecular descriptors. *Envir. Toxicol. Pharmacol.*, **16**, 45–55.
- Basak, S.C., Mills, D., Hawkins, D.M. and El-Masri, H.A. (2002) Prediction of tissue:air partition coefficients, a comparison of structure-based and property-based methods. *SAR & QSAR Environ. Res.*, **13**, 649–665.
- Basak, S.C., Mills, D., Hawkins, D.M. and El-Masri, H.A. (2003a) Prediction of human blood:air partition coefficient, a comparison of structure-based and property-based methods. *Risk Anal.*, **23**, 1173–1184.
- Basak, S.C., Mills, D., Mumtaz, M. and Balasubramanian, K. (2003b) Use of topological indices in predicting aryl hydrocarbon receptor binding potency of dibenzofurans: a hierarchical QSAR approach. *Indian J. Chem.*, **42**, 1385–1391.
- Basak, S.C., Mills, D.R., Balaban, A.T. and Gute, B.D. (2001) Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: a hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, **41**, 671–678.
- Basak, S.C., Monsrud, L.J., Rosen, M.E., Frane, C.M. and Magnuson, V.R. (1986) A comparative study of lipophilicity and topological indices in biological correlation. *Acta Pharm. Jugosl.*, **36**, 81–95.
- Basak, S.C., Niemi, G.J. and Veith, G.D. (1990a) A graph theoretic approach to predicting molecular properties. *Math. Comput. Modelling*, **14**, 511–516.
- Basak, S.C., Niemi, G.J. and Veith, G.D. (1990b) Optimal characterization of structure for prediction of properties. *J. Math. Chem.*, **4**, 185–205.
- Basak, S.C., Niemi, G.J. and Veith, G.D. (1990c) Recent developments in the characterization of chemical structure using graph-theoretic indices, in *Computational Chemical Graph Theory* (ed. D.H. Rouvray), Nova Science Publishers, New York, pp. 235–277.

- Basak, S.C., Niemi, G.J. and Veith, G.D. (1991) Predicting properties of molecules using graph invariants. *J. Math. Chem.*, **7**, 243–272.
- Basak, S.C., Nikolić, S., Trinajstić, N. and Amić, D. (2000) QSPR modeling: graph connectivity indices versus line graph connectivity indices. *J. Chem. Inf. Comput. Sci.*, **40**, 927–933.
- Basak, S.C., Rosen, M.E. and Magnuson, V.R. (1986) Molecular topology and mutagenicity: a QSAR study of nitrosoamines. *Med. Sci. Res.*, **14**, 848–849.
- Basak, S.C., Roy, A.B. and Ghosh, J.J. (1980) Study of the structure–function relationship of pharmacological and toxicological agents using information theory, in Proceedings of the Second International Conference on Mathematical Modelling (eds X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler), University of Missouri, Rolla, MS, pp. 851–856.
- Basilevsky, A. (1994) *Statistical Factor Analysis and Related Methods*, John Wiley & Sons, Inc., New York, p. 738.
- Baskin, I.I., Gordeeva, E.V., Devdariani, R.O., Zefirov, N.S., Palyulin, V.A. and Stankevitch, I.V. (1989) Solving the inverse problem of structure–property relations for the case of topological indexes. *Dokl. Akad. Nauk. SSSR*, **307**, 613–617.
- Baskin, I.I., Halberstam, N.M., Artemenko, N.V., Palyulin, V.A. and Zefirov, N.S. (2003) NASAWIN – a universal software for QSPR/QSAR studies, in *Designing Drugs and Crop Protectants: Processes, Problems and Solutions* (ed. M. Ford), Blackwell Publishing, Oxford, UK, pp. 260–263.
- Baskin, I.I., Palyulin, V.A. and Zefirov, N.S. (1997) A neural device for searching direct correlations between structures and properties of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **37**, 715–721.
- Baskin, I.I., Skvortsova, M.I., Stankevitch, I.V. and Zefirov, N.S. (1995) On the basis of invariants of labeled molecular graphs. *J. Chem. Inf. Comput. Sci.*, **35**, 527–531.
- Bassoli, A., Drew, M.G.B., Hattotuwegama, C.K., Merlini, L., Morini, G. and Wilden, G.R.H. (2001) Quantitative structure–activity relationships of sweet isovanillyl derivatives. *Quant. Struct. -Act. Relat.*, **20**, 3–16.
- Bate-Smith, E.C. and Westall, R.G. (1950) Chromatographic behaviour and chemical structure. I. Some naturally occurring phenolic substances. *Biochim. Biophys. Acta*, **4**, 427–440.
- Bath, P.A., Morris, C.A. and Willett, P. (1993) Effects of standardization on fragment-based measures of structural similarity. *J. Chemom.*, **7**, 543–550.
- Bath, P.A., Poirrette, A.R., Willett, P. and Allen, F.H. (1994) Similarity searching in files of three-dimensional chemical structures: comparison of fragment-based measures of shape similarity. *J. Chem. Inf. Comput. Sci.*, **34**, 141–147.
- Bath, P.A., Poirrette, A.R., Willett, P. and Allen, F.H. (1995) The extent of the relationship between the graph-theoretical and the geometrical shape coefficients of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **35**, 714–716.
- Batista, J. and Bajorath, J. (2007) Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *J. Chem. Inf. Model.*, **47**, 59–68.
- Battershell, C., Malhotra, D. and Hopfinger, A.J. (1981) Inhibition of dihydrofolate reductase: structure–activity correlations of quinazolines based upon molecular shape analysis. *J. Med. Chem.*, **24**, 812–818.
- Bauerschmidt, S. and Gasteiger, J. (1997) Overcoming the limitations of a connection table description: a universal representation of chemical species. *J. Chem. Inf. Comput. Sci.*, **37**, 705–714.
- Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J. and Gasteiger, J. (1996) Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.*, **36**, 1205–1213.
- Baum, E.J. (1997) *Chemical Property Estimation: Theory and Application*, Lewis Publishers, Boca Raton, FL, p. 448.
- Baumann, K. (1999) Uniform-length molecular descriptors for quantitative structure–property relationships (QSPR) and quantitative structure–activity relationships (QSAR): classification studies and similarity searching. *TRAC*, **18**, 36–46.
- Baumann, K. (2002a) An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features. *J. Chem. Inf. Comput. Sci.*, **42**, 26–35.
- Baumann, K. (2002b) Distance profiles (DiP): a translationally and rotationally invariant 3D structure descriptor capturing steric properties of molecules. *Quant. Struct. -Act. Relat.*, **21**, 507–519.
- Baumann, K. (2003) Cross-validation as the objective function for variable-selection techniques. *TRAC*, **22**, 395–406.
- Baumann, K., Affolter, C., Pretsch, E. and Clerc, J.T. (1997) Numerical structure representation and IR spectra prediction. *Mikrochim. Acta*, **14** (Suppl.), 275–276.
- Baumann, K., Albert, H. and von Korff, M. (2002) A systematic evaluation of the benefits and hazards of

- variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. *J. Chemom.*, **16**, 339–350.
- Baumann, K. and Clerc, J.T. (1997) Computer-assisted IR spectra prediction – linked similarity searches for structures and spectra. *Anal. Chim. Acta*, **348**, 327–343.
- Baumann, K. and Stiefl, N. (2004) Validation tools for variable subset regression. *J. Comput. Aid. Mol. Des.*, **18**, 549–562.
- Baumer, L., Sala, G. and Sello, G. (1989) Residual charges on atoms in organic structures: molecules containing charged and backdonating atoms. *Tetrahedron Comput. Methodol.*, **2**, 105–118.
- Baurin, N., Mozziconacci, J.C., Arnoult, E., Chavatte, P., Marot, C. and Morin-Allory, L. (2004) 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database. *J. Chem. Inf. Comput. Sci.*, **44**, 276–285.
- Baurin, N., Vangrevelinge, E., Morin-Allory, L., Méroux, J.-Y., Renard, P., Payard, M., Guillaumet, G. and Marot, C. (2000) 3D-QSAR CoMFA study on imidazolinergic I2 ligands: a significant model through a combined exploration of structural diversity and methodology. *J. Med. Chem.*, **43**, 1109–1122.
- Bawden, D. (1983) Computerized chemical structure–handling techniques in structure–activity studies and molecular property prediction. *J. Chem. Inf. Comput. Sci.*, **23**, 14–22.
- Bawden, D. (1990) Applications of two-dimensional chemical similarity measures to database analysis and querying, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 65–76.
- Baxter, S.J., Jenkins, H.D.B. and Samuel, C.J. (1996) A novel computational approach to the estimation of steric parameters. III. Extension to aliphatic amines and application to the adrenergic blocking activity of β -haloalkylamine. *Tetrahedron Lett.*, **37**, 4617–4620.
- Bayada, D.M., Hemersma, H. and van Geerestein, V. J. (1999) Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.*, **39**, 1–10.
- Bayley, M.J. and Willett, P. (1999) Binning schemes for partition-based compound selection. *J. Mol. Graph. Model.*, **17**, 10–18.
- Bayram, E., Santago, I.I.P., Harris, R., Xiao, Y.-D., Clausef, A.J. and Schmitt, J.D. (2004) Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems. *J. Comput. Aid. Mol. Des.*, **18**, 483–493.
- Bazylak, G. (1994) Differentiation of alkanolamines properties by multivariate analysis of database founded by their molecular parameters and chromatographic measurements results. *Chem. Anal.*, **39**, 295–308.
- Bazylak, G. and Nagels, L.J. (2002) Potentiometric detection of exogenic beta-adrenergic substances in liquid chromatography. *J. Chromat.*, **973**, 85–96.
- BCI fingerprints, Barnard Chemical Information Ltd, 46 Uppergate Road, Stannington, Sheffield S6 6BX, UK.
- Bearden, A.P. and Schultz, T.W. (1997) Structure–activity relationships for *Pimephales* and *Tetrahymena*: a mechanism of action approach. *Environ. Toxicol. Chem.*, **16**, 1311–1317.
- Beck, B., Breindl, A. and Clark, T. (2000) QM/NN QSPR models with error estimation: vapor pressure and log *P*. *J. Chem. Inf. Comput. Sci.*, **40**, 1046–1051.
- Beck, B., Glen, R.C. and Clark, T. (1996) The inhibition of α -chymotrypsin predicted using theoretically derived molecular properties. *J. Mol. Graph.*, **14**, 130–135.
- Beck, B., Horn, A., Carpenter, J.E. and Clark, T. (1998) Enhanced 3D-databases: a fully electrostatic database of AM1-optimized structures. *J. Chem. Inf. Comput. Sci.*, **38**, 1214–1217.
- Beckaus, H.-D. (1978) S_f parameters – a measure of the front strain of alkyl groups. *Angew. Chem. Int. Ed. Engl.*, **17**, 593–594.
- Becke, A.D. and Edgecombe, K.E. (1990) A simple measure of electron localization in atomic and molecular systems. *J. Chim. Phys.*, **92**, 5397–5403.
- Beger, R.D., Buzatu, D.A., Wilkes, J.G. and Lay, J.O., Jr (2001) ^{13}C NMR quantitative spectrometric data–activity relationship (QSDAR) models of steroids binding the aromatase enzyme. *J. Chem. Inf. Comput. Sci.*, **41**, 1360–1366.
- Beger, R.D., Buzatu, D.A., Wilkes, J.G. and Lay, J.O., Jr (2002) Comparative structural connectivity spectra analysis (CoSCoSA) models of steroid binding to the corticosteroid binding globulin. *J. Chem. Inf. Comput. Sci.*, **42**, 1123–1131.
- Beger, R.D. and Wilkes, J.G. (2001a) Developing ^{13}C NMR quantitative spectrometric data–activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. *J. Comput. Aid. Mol. Des.*, **15**, 659–669.
- Beger, R.D. and Wilkes, J.G. (2001b) Models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding affinity to the aryl

- hydrocarbon receptor developed using ^{13}C NMR data. *J. Chem. Inf. Comput. Sci.*, **41**, 1322–1329.
- Béliveau, M., Lipscomb, J., Tardif, R. and Krishnan, K. (2005) Quantitative structure–property relationships for interspecies extrapolation of the inhalation pharmacokinetics of organic chemicals. *Chem. Res. Toxicol.*, **18**, 475–485.
- Béliveau, M., Tardif, R. and Krishnan, K. (2003) Quantitative structure–property relationships for physiologically based pharmacokinetic modeling of volatile organic chemicals in rats. *Toxicol. Appl. Pharm.*, **189**, 221–232.
- Belvisi, L., Bravi, G., Catalano, G., Mabilia, M., Salimbeni, A. and Scolastico, C. (1996) A 3D QSAR CoMFA study of non-peptide angiotensin II receptor antagonists. *J. Comput. Aid. Mol. Des.*, **10**, 567–582.
- Belvisi, L., Bravi, G., Scolastico, C., Vulpetti, A., Salimbeni, A. and Todeschini, R. (1994) A 3D QSAR approach to the search for geometrical similarity in a series of nonpeptide angiotensin II receptor antagonists. *J. Comput. Aid. Mol. Des.*, **8**, 211–220.
- Belvisi, L., Brossa, S., Salimbeni, A., Scolastico, C. and Todeschini, R. (1991) Structure–activity relationship of Ca^{2+} channel blockers: a study using conformational analysis and chemometrics methods. *J. Comput. Aid. Mol. Des.*, **5**, 571–584.
- Belvisi, L., Salimbeni, A., Scolastico, C., Todeschini, R. and Vulpetti, A. (1991) Molecular modeling of non-peptide angiotensin II receptor antagonists. *Pharm. Pharmacol. Lett.*, **1**, 57–60.
- Bemis, G.W. and Kuntz, I.D. (1992) A fast and efficient method for 2D and 3D molecular shape description. *J. Comput. Aid. Mol. Des.*, **6**, 607–628.
- Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, **39**, 2887–2893.
- Bender, A. and Glen, R.C. (2005) A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.*, **45**, 1369–1375.
- Bender, A., Mussa, H.Y., Gill, G.S. and Glen, R.C. (2004a) Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.*, **47**, 6569–6583.
- Bender, A., Mussa, H.Y., Glen, R.C. and Reiling, S. (2004b) Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.*, **44**, 170–178.
- Bender, A., Mussa, H.Y., Glen, R.C. and Reiling, S. (2004c) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.*, **44**, 1708–1718.
- Benfenati, E., Gini, G., Piclin, N., Roncaglioni, A. and Vari, M.R. (2003) Predicting log *P* of pesticides using different software. *Chemosphere*, **53**, 1155–1164.
- Benicori, T., Consonni, V., Gramatica, P., Pilati, T., Rizzo, S., Sannicolò, F., Todeschini, R. and Zotti, G. (2001) Steric control of conductivity in highly conjugated polythiophenes. *Chem. Mater.*, **13**, 1665–1673.
- Benigni, R. (1991) QSAR prediction of rodent carcinogenicity for a set of chemicals currently bioassayed by the US National Toxicology Program. *Mutagenesis*, **6**, 423–425.
- Benigni, R. (1994) EVE, a distance based approach for discriminating nonlinearly separable groups. *Quant. Struct. -Act. Relat.*, **13**, 406–411.
- Benigni, R. (ed.) (2003) *Quantitative Structure–Activity Relationship (QSAR). Models of Mutagens and Carcinogens*, CRC Press, Boca Raton, FL, pp. 286.
- Benigni, R. (2005) Structure–activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem. Rev.*, **105**, 1767–1800.
- Benigni, R., Andreoli, C., Conti, L., Tafani, P., Cotta Ramusino, M., Carere, A. and Crebelli, R. (1993) Quantitative structure–activity relationship models correctly predict the toxic and aneuploidizing properties of 6-halogenated methanes in *Aspergillus nidulans*. *Mutagenesis*, **8**, 301–305.
- Benigni, R., Andreoli, C. and Giuliani, A. (1989) Interrelationships among carcinogenicity, mutagenicity, acute toxicity, and chemical structure in a genotoxicity data base. *J. Toxicol. Env. Health*, **27**, 1–20.
- Benigni, R., Andreoli, C. and Giuliani, A. (1994) QSAR models for both mutagenic potency and activity application to nitroarenes and aromatic amines. *Envir. Mol. Mutag.*, **24**, 208–219.
- Benigni, R. and Bosa, C. (2006) Structural alerts of mutagens and carcinogens. *Curr. Comput. -Aided Drug Des.*, **2**, 169–176.
- Benigni, R., Conti, L., Crebelli, R., Rodomonte, A. and Vari, M.R. (2005) Simple and α,β -unsaturated aldehydes: correct prediction of genotoxic activity through structure–activity relationship models. *Envir. Mol. Mutag.*, **46**, 268–280.
- Benigni, R., Cotta Ramusino, M., Giorgi, F. and Gallo, G. (1995) Molecular similarity matrices and quantitative structure–activity relationships: a case

- study with methodological implications. *J. Med. Chem.*, **38**, 629–635.
- Benigni, R., Gallo, G., Giorgi, F. and Giuliani, A. (1999) On the equivalence between different descriptions of molecules: value for computational approaches. *J. Chem. Inf. Comput. Sci.*, **39**, 575–578.
- Benigni, R. and Giuliani, A. (1991) What kind of statistics for QSAR research? *Quant. Struct. -Act. Relat.*, **10**, 99–100.
- Benigni, R. and Giuliani, A. (1993) Analysis of distance matrices for studying data structures and separating classes. *Quant. Struct. -Act. Relat.*, **12**, 397–401.
- Benigni, R. and Giuliani, A. (1994) Quantitative structure–activity relationship (QSAR) studies in genetic toxicology. Mathematical models and the biological activity term of the relationship. *Mut. Res.*, **306**, 181–186.
- Benigni, R., Giuliani, A. and Passerini, L. (2001) Infrared spectra as chemical descriptors for QSAR models. *J. Chem. Inf. Comput. Sci.*, **41**, 727–730.
- Benigni, R., Netzeva, T.I., Benfenati, E., Bossa, C., Franke, R., Helma, C., Hulzebos, E., Marchant, C., Richard, A., Woo, Y.-T. and Yang, C. (2007) The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. *J. Environ. Sci. Health*, **25**, 53–97.
- Benigni, R. and Passerini, L. (2002) Carcinogenicity of the aromatic amines: from structure–activity relationships to mechanisms of action and risk assessment. *Mut. Res.*, **511**, 191–206.
- Benigni, R., Passerini, L., Livingstone, D.J. and Johnson, M.A. (1999a) Infrared spectra information and their correlation with QSAR descriptors. *J. Chem. Inf. Comput. Sci.*, **39**, 558–562.
- Benigni, R., Passerini, L., Pino, A. and Giuliani, A. (1999b) The information content of the eigenvalues from modified adjacency matrices: large scale and small scale correlations. *Quant. Struct. -Act. Relat.*, **18**, 449–455.
- Benigni, R. and Zito, R. (2003) Designing safer drugs: (Q)SAR-based identification of mutagens and carcinogens. *Curr. Top. Med. Chem.*, **3**, 1289–1300.
- Bentley, T.W. and von Schleyer, P.R. (1977) Medium effects on the rates and mechanisms of solvolytic reactions. *Adv. Phys. Org. Chem.*, **14**, 1–67.
- Berger, B.M. and Wolfe, N.L. (1996) Hydrolysis and biodegradation of sulfonylurea herbicides in aqueous buffers and anaerobic water–sediment systems. Assessing fate pathways using molecular descriptors. *Environ. Toxicol. Chem.*, **15**, 1500–1507.
- Berglund, A., De Rosa, M.C. and Wold, S. (1997) Alignment of flexible molecules at their receptor site using 3D descriptors and Hi-PCA. *J. Comput. Aid. Mol. Des.*, **11**, 601–612.
- Bergmann, D. and Hinze, J. (1996) Electronegativity and molecular properties. *Angew. Chem. Int. Ed. Engl.*, **35**, 150–163.
- Bergström, C.A.S., Norinder, U., Luthman, K. and Artursson, P. (2003) Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.*, **43**, 1177–1185.
- Berinde, Z. and Berinde, M. (2004) On a matrix representation of molecular structures. *Carpathian J. Math.*, **20**, 205–209.
- Bermúdez-Saldaña, J.M., Escuder-Gilabert, L., Medina-Hernández, M.J., Villanueva-Camañas, R. M. and Sagrado, S. (2005) Modelling bioconcentration of pesticides in fish using biopartitioning micellar chromatography. *J. Chromat.*, **1063**, 153–160.
- Bernard, P., Golbraikh, A., Kireev, D.B., Chrétien, J.R. and Rozhkova, N. (1998) Comparison of chemical databases: analysis of molecular diversity with self organizing maps (SOM). *Analisis*, **26**, 333–341.
- Bernard, P., Kireev, D.B., Chrétien, J.R., Fortier, P.-L. and Coppet, L. (1999) Automated docking of 82 N-benzylpiperidine derivatives to mouse acetylcholinesterase and comparative molecular field analysis with ‘natural’ alignment. *J. Comput. Aid. Mol. Des.*, **13**, 355–371.
- Bernejo, J., Canga, J.S., Gayol, O.M. and Guillen, M. D. (1984) Utilization of physico-chemical properties and structural parameters for calculating retention indices of alkylbenzenes. *J. Chromatogr. Sci.*, **22**, 252.
- Bernstein, H.J. (1947) A relation between bond order and covalent bond distance. *J. Chim. Phys.*, **15**, 284–289.
- Bersohn, M. (1983) A fast algorithm for calculation of the distance matrix. *J. Comput. Chem.*, **4**, 110–113.
- Bersohn, M. (1987) A matrix method for partitioning the atoms of a molecule into equivalence classes. *Computers Chem.*, **11**, 67–72.
- Berson, J.A., Hamlet, Z. and Mueller, W.A. (1962) The correlation of solvent effects on the stereoselectivities of Diels–Alder reactions by means of linear free energy relationships. A new empirical measure of solvent polarity. *J. Am. Chem. Soc.*, **84**, 297–304.
- Bersuker, I.B. (2003) Pharmacophore identification and quantitative bioactivity prediction using the electron-conformational method. *Curr. Pharm. Design*, **9**, 1575–1606.
- Bersuker, I.B., Bahçeci, S. and Boggs, J.E. (2000a) Improved electron-conformational method of

- pharmacophore identification and bioactivity prediction. Application to angiotensin converting enzyme inhibitors. *J. Chem. Inf. Comput. Sci.*, **40**, 1363–1376.
- Bersuker, I.B., Bahçeci, S. and Boggs, J.E. (2000b) The electron-conformational method of identification of pharmacophore and anti-pharmacophore shielding, in *Pharmacophore Perception, Development, and Use in Drug Design* (ed. O.F. Guner), International University Line, La Jolla, CA, pp. 457–474.
- Bersuker, I.B., Bahçeci, S., Boggs, J.E. and Pearlman, R.S. (1999a) A novel electron-conformational approach to molecular modeling for QSAR by identification of pharmacophore and anti-pharmacophore shielding. *SAR & QSAR Environ. Res.*, **10**, 157–173.
- Bersuker, I.B., Bahçeci, S., Boggs, J.E. and Pearlman, R.S. (1999b) An electron-conformational method of identification of pharmacophore and anti-pharmacophore shielding: application to rice blast activity. *J. Comput. Aid. Mol. Des.*, **13**, 419–434.
- Bersuker, I.B. and Dimoglo, A.S. (1991) The electron-topological approach to the QSAR problem, in *Reviews in Computational Chemistry*, Vol. 2 (eds K.B. Lipkowitz and D. Boyd), VCH Publishers, New York, pp. 423–460.
- Bersuker, I.B., Dimoglo, A.S., Gorbachov, M.Yu., Vlad, P.F. and Pesaro, M. (1991) Origin of musk fragrance activity: the electron-topological approach. *New J. Chem.*, **15**, 307–320.
- Berthelot, M., Laurence, C., Lucon, M. and Rossignol, C. (1996) Partition coefficients and intramolecular hydrogen bonding. 2. The influence of partition solvents on the intramolecular hydrogen bond stability of salicylic acid derivatives. *J. Phys. Org. Chem.*, **9**, 626–630.
- Bertz, S.H. (1981) The first general index of molecular complexity. *J. Am. Chem. Soc.*, **103**, 3599–3601.
- Bertz, S.H. (1983a) A mathematical model of molecular complexity, in *Chemical Applications of Topology and Graph Theory* (ed. R.B. King), Elsevier, Amsterdam, The Netherlands, pp. 206–221.
- Bertz, S.H. (1983b) On the complexity of graphs and molecules. *Bull. Math. Biol.*, **45**, 849–855.
- Bertz, S.H. (1988) Branching in graphs and molecules. *Disc. Appl. Math.*, **19**, 65–83.
- Bertz, S.H. and Herndon, W.C. (1986) The similarity of graphs and molecules, in *Artificial Intelligence Applications in Chemistry* (eds T.H. Pierce and B.A. Hohne), ACS, Washington, DC, pp. 169–175.
- Bertz, S.H. and Rücker, C. (2004) In search of simplification: the use of topological complexity indices to guide retrosynthetic analysis. *Croat. Chem. Acta*, **77**, 221–235.
- Besalú, E., Carbó, R., Mestres, J. and Solà, M. (1995) Foundations and recent developments on molecular quantum similarity. *Top. Curr. Chem.*, **173**, 31–62.
- Besalú, E., Gallegos, A. and Carbó-Dorca, R. (2001) Topological quantum similarity indices and their use in QSAR: application to several families of antimalarial compounds. *MATCH Commun. Math. Comput. Chem.*, **44**, 41–64.
- Besalú, E., Ponc, R. and De Julián-Ortiz, V. (2003) Virtual generation of agents against *Mycobacterium tuberculosis*. A QSAR study. *Mol. Div.*, **6**, 107–120.
- Beteringhe, A. and Balaban, A.T. (2004) QSAR for toxicities of polychlorodibenzofurans, polychlorodibenzo-1,4-dioxins, and polychlorobiphenyls. *ARKIVOC*, (i), 163–182.
- Beteringhe, A., Filip, P. and Tarko, L. (2005) QSAR study for diarylguanidines, noncompetitive NMDA receptor antagonists. A new topological index AAd derived from local invariants of the chemical graphs of diarylguanidines. *ARKIVOC*, (x), 45–62.
- Beteringhe, A., Radutiu, A.C., Bem, M., Costantinescu, T. and Balaban, A.T. (2006) QSPR study for the hydrophobicity of 4-aryloxy-7-nitrobenzofurazan and 2-aryloxy-(α -acetyl)-phenoxathiin derivatives. *Internet Electron. J. Mol. Des.*, **5**, 237–246.
- Betso, J.E., Carreon, R.E. and Miner, V.M. (1991) The use of proton nuclear magnetic resonance spectrometry (^1H NMR) for monitoring the reaction of epoxides with butylamine and predictive capabilities of the relative alkylation index (RAI) for skin sensitization by epoxides. *Toxicol. Appl. Pharm.*, **108**, 483–488.
- Beyer, A., Mackey, D., Matthies, M., Wania, F. and Webster, E. (2000) Assessing long-range transport potential of persistent organic pollutants. *Environ. Sci. Technol.*, **34**, 699–703.
- Bhal, S.K., Kassam, K., Peirson, I.G. and Pearl, G.M. (2007) The rule of five revisited: applying log *D* in place of log *P* in drug-likeness filters. *Mol. Pharm.*, **4**, 556–560.
- Bhat, K.L., Hayik, S., Sztandera, L. and Bock, C.W. (2005) Mutagenicity of aromatic and heteroaromatic amines and related compounds: a QSAR investigation. *QSAR Comb. Sci.*, **24**, 831–843.
- Bhattacharjee, A.K. and Karle, J.M. (1999) Stereoelectronic properties of antimalarial artemisinin analogues in relation to neurotoxicity. *Chem. Res. Toxicol.*, **12**, 422–428.

- Bhattacharjee, A.K., Kyle, D.E., Vennerstrom, J.L. and Milhous, W.K. (2002) A 3D QSAR pharmacophore model and quantum chemical structure–activity analysis of chloroquine (CQ)-resistance reversal. *J. Chem. Inf. Comput. Sci.*, **42**, 1212–1220.
- Bhattacharjee, S. (1994) Haloethanes, geometric volume and atomic contribution method. *Computers Chem.*, **18**, 419–429.
- Bhattacharjee, S., Basak, S.C. and Dasgupta, P. (1992) Molecular property correlation in haloethanes with geometric volume. *Computers Chem.*, **16**, 223–228.
- Bhattacharjee, S. and Dasgupta, P. (1994) Molecular property correlation in alkanes with geometric volume. *Computers Chem.*, **18**, 61–71.
- Bhattacharjee, S., Rao, A.S. and Dasgupta, P. (1991) A new index for molecular property correlation in halomethanes. *Computers Chem.*, **15**, 319–322.
- Bienfait, B. (1994) Applications of high-resolution self-organizing maps to retrosynthetic and QSAR analysis. *J. Chem. Inf. Comput. Sci.*, **34**, 890–898.
- Biggs, A.I. and Robinson, R.A. (1961) The ionisation constants of some substituted anilines and phenols: a test of the Hammett relation. *J. Chem. Soc.*, 388–393.
- Bijloo, G.J. and Rekker, R.F. (1984a) Some critical remarks concerning the inductive parameter I. Part IV. Parametrization of the *ortho* effect in anilines and pyridines. *Quant. Struct. -Act. Relat.*, **3**, 111–115.
- Bijloo, G.J. and Rekker, R.F. (1984b) Some critical remarks concerning the inductive parameter σ_i . Part III. Parametrization of the *ortho* effect in benzoic acids and phenols. *Quant. Struct. -Act. Relat.*, **3**, 91–96.
- Bindal, M.C., Singh, P. and Gupta, S.P. (1980) Quantitative correlation of anesthetic potencies of halogenated hydrocarbons with boiling point and molecular connectivity. *Arzneim. Forsch. (German)*, **30**, 234–236.
- Binsch, G. and Heilbronner, E. (1968) Double bond fixation in non-alternant π -electron systems. *Tetrahedron*, **24**, 1215–1223.
- Bird, C.W. (1985) A new aromaticity index and its application to five-membered ring heterocycles. *Tetrahedron*, **41**, 1409–1414.
- Bird, C.W. (1986) The application of a new aromaticity index to six-membered ring heterocycles. *Tetrahedron*, **42**, 89–92.
- Bird, C.W. (1997) Heteroaromaticity. 10. The direct calculation of resonance energies of azines and azoles from molecular dimensions. *Tetrahedron*, **53**, 13111–13118.
- Birkner, M.D. and van der Laan, M.J. (2006) Application of a variable importance measure method. *Int. J. Biostat.*, **2**, 1–22.
- Bjorsvik, H.R., Hansen, U.M. and Carlson, R. (1997) Principal properties of monodentate phosphorus ligands. Predictive model for the carbonyl absorption frequencies in NiCo_3L complexes. *Acta Chem. Scand.*, **51**, 733–741.
- Bjorsvik, H.R. and Priebe, H. (1995) Multivariate data analysis of molecular descriptors estimated by use of semiempirical quantum chemistry methods. Principal properties for synthetic screening of 2-chloromethyloxirane and analogous bis-alkylating C-3 moieties. *Acta Chem. Scand.*, **49**, 446–456.
- Blackstock, W.P. and Weir, M.P. (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.*, **17**, 121–127.
- Blaney, F.E., Naylor, D. and Woods, J. (1993) Mambas: a real time graphics environment for QSAR. *J. Mol. Graph.*, **11**, 157–165.
- Blatova, O.A., Blatov, V.A. and Serezhkin, V.N. (2001) Study of rare-earth π -complexes by means of Voronoi–Dirichlet polyhedra. *Acta Cryst.*, **57**, 261–270.
- Blatova, O.A., Blatov, V.A. and Serezhkin, V.N. (2002) A new set of molecular descriptors. *Acta Cryst.*, **59**, 219–226.
- Blin, N., Federici, C., Koscielniak, T. and Strosberg, A. D. (1995) Predictive quantitative structure–activity relationships (QSAR) analysis of beta 3-adrenergic ligands. *Drug Design & Discovery*, **12**, 297–311.
- Blower, P.E., Fligner, M.A., Verducci, J.S. and Bjoraker, J. (2002) On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.*, **42**, 393–404.
- Blum, D.J., Suffet, I.H. and Duguet, J.P. (1994) Quantitative structure–activity relationship using molecular connectivity for the activated carbon adsorption of organic chemicals in water. *Water Res.*, **28**, 687–699.
- Blumenthal, L.M. (1970) *Theory and Applications of Distance Geometry*, American Mathematical Society, Providence, RI.
- Blurock, E.S. (1998) Use of atomic and bond parameters in a spectral representation of a molecule for physical property determination. *J. Chem. Inf. Comput. Sci.*, **38**, 1111–1118.
- Bocek, K., Kopecky, J., Krivucova, M. and Vlachová, D. (1964) Chemical structure and biological activity of *p*-disubstituted derivatives of benzene. *Experientia*, **20**, 667–668.
- Böcker, A., Schneider, G. and Teckentrup, A. (2004) Status of HTS data mining approaches. *QSAR Comb. Sci.*, **23**, 207–213.
- Bodor, N., Buchwald, P. and Huang, M.-J. (1998) Computer-assisted design of new drugs based on

- retrometabolic concepts. *SAR & QSAR Environ. Res.*, **8**, 41–92.
- Bodor, N., Gabanyi, Z. and Wong, C.-K. (1989) A new method for the estimation of partition coefficient. *J. Am. Chem. Soc.*, **111**, 3783–3786.
- Bodor, N., Harget, A. and Huang, M.-J. (1991) Neural network studies. 1. Estimation of the aqueous solubility of organic compounds. *J. Am. Chem. Soc.*, **113**, 9480–9483.
- Bodor, N. and Huang, M.-J. (1992a) A new method for the estimation of the aqueous solubility of organic compounds. *J. Pharm. Sci.*, **81**, 954–960.
- Bodor, N. and Huang, M.-J. (1992b) An extended version of a novel method for the estimation of partition coefficients. *J. Pharm. Sci.*, **81**, 272–281.
- Bodor, N., Huang, M.-J. and Harget, A. (1992) Neural network studies. 4. An extended study of the aqueous solubility of organic compounds. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **26**, 853–867.
- Bodor, N., Huang, M.-J. and Harget, A. (1994) Neural network studies. Part 3. Prediction of partition coefficients. *J. Mol. Struct. (Theochem)*, **309**, 259–266.
- Boecklen, W.J. and Niemi, G.J. (1994) Multivariate association of graph-theoretic variables and physico-chemical properties. *SAR & QSAR Environ. Res.*, **2**, 79–87.
- Boethling, R.S. and Sabljic, A. (1989) Screening-level model for aerobic biodegradability based on a survey of expert knowledge. *Environ. Sci. Technol.*, **23**, 672–679.
- Bogdanov, B., Nikolic, S. and Trinajstić, N. (1989) On the three-dimensional Wiener number. *J. Math. Chem.*, **3**, 299–309.
- Bogdanov, B., Nikolic, S. and Trinajstić, N. (1990) On the three-dimensional Wiener number. A comment. *J. Math. Chem.*, **5**, 305–306.
- Bögel, H., Dettmann, J. and Randić, M. (1997) Why is the topological approach in chemistry so successful? *Croat. Chem. Acta*, **70**, 827–840.
- Boggia, R., Forina, M., Fossa, P. and Mosti, L. (1997) Chemometric study and validation strategies in the structure–activity relationship of new cardiotonic agents. *Quant. Struct. -Act. Relat.*, **16**, 201–213.
- Bohac, M., Loeprecht, B., Damborsky, J. and Schüürmann, G. (2002) Impact of orthogonal signal correction (OSC) on the predictive ability of CoMFA models for the ciliate toxicity of nitrobenzenes. *Quant. Struct. -Act. Relat.*, **21**, 3–11.
- Bohl, M., Dumber, J., Gifford, E.M., Heritage, T.W., Wild, D.J., Willett, P. and Wilton, D.J. (2002) Scaffold searching: automated identification of similar ring systems for the design of combinatorial libraries. *Quant. Struct. -Act. Relat.*, **21**, 590–597.
- Böhm, H.-J. (1992a) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aid. Mol. Des.*, **6**, 593–606.
- Böhm, H.-J. (1992b) The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput. Aid. Mol. Des.*, **6**, 61–78.
- Böhm, H.-J. (1994a) On the use of LUDI to search the fine chemicals directory for ligands of proteins of known three-dimensional structure. *J. Comput. Aid. Mol. Des.*, **8**, 623–632.
- Böhm, H.-J. (1994b) The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J. Comput. Aid. Mol. Des.*, **8**, 243–256.
- Böhm, H.-J. (1996) Current computational tools for *de novo* ligand design. *Curr. Opin. Biotechnol.*, **7**, 433–436.
- Böhm, H.-J. (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from *de novo* design or 3D database search programs. *J. Comput. Aid. Mol. Des.*, **12**, 309–323.
- Böhm, H.-J. and Stahl, M. (2000) Structure-based library design: molecular modelling merges with combinatorial chemistry. *Curr. Opin. Chem. Biol.*, **4**, 283–286.
- Böhm, M. and Klebe, G. (2002) Development of new hydrogen-bond descriptors and their application to comparative molecular field analyses. *J. Med. Chem.*, **45**, 1585–1597.
- Bojarski, J. and Ekiert, L. (1982) Relationship between molecular connectivity indices of barbiturates and chromatographic parameters. *Chromatographia*, **15**, 172.
- Bojarski, J. and Ekiert, L. (1983) Evaluation of modified valence molecular connectivity index for correlation of chromatographic parameters. *J. Liquid Chromat.*, **6**, 73.
- Bolboacă, S. and Jäntschi, L. (2005a) Molecular descriptors family on structure–activity relationships. 2. Insecticidal activity of neonicotinoid compounds. *Leonardo Journal of Sciences*, **6**, 78–85.
- Bolboacă, S. and Jäntschi, L. (2005b) Molecular descriptors family on structure–activity relationships. 3. Antituberculous activity of some polyhydroxyxanthones. *Leonardo Journal of Sciences*, **6**, 58–64.
- Bonaccorsi, R., Scrocco, E. and Tomasi, J. (1970) Molecular SCF calculations for the ground state of some three-membered ring molecules: $(\text{CH}_2)_3$, $(\text{CH}_2)_2\text{NH}$, $(\text{CH}_2)_2\text{NH}_2^+$, $(\text{CH}_2)_2\text{O}$, $(\text{CH}_2)_2\text{S}$, $(\text{CH}_2)_2\text{CH}_2$, and N_2CH_2 . *J. Chim. Phys.*, **52**, 5270–5284.

- Bonacich, P. (1972) Factoring and weighing approaches to clique identification. *Journal of Mathematical Sociology*, **2**, 113–120.
- Bonacich, P. (2007) Some unique properties of eigenvector centrality. *Social Networks*, **29**, 555–564.
- Bonati, L., Fraschini, E., Lasagni, M., Palma Modoni, E. and Pitea, D. (1995) A hypothesis on the mechanism of PCDD biological activity based on molecular electrostatic potential modelling. Part 2. *J. Mol. Struct. (Theochem)*, **340**, 83–95.
- Bonchev, D. (1983) *Information Theoretic Indices for Characterization of Chemical Structures*, Research Studies Press, Chichester, UK, p. 249.
- Bonchev, D. (1989) The concept for the centre of a chemical structure and its applications. *J. Mol. Struct. (Theochem)*, **185**, 155–168.
- Bonchev, D. (1990) *The Problems of Computing Molecular Complexity*, in *Computational Chemical Graph Theory* (ed. D.H. Rouvray), Nova Science Publishers, New York, pp. 33–63.
- Bonchev, D. (1995) Topological order in molecules. 1. Molecular branching revisited. *J. Mol. Struct. (Theochem)*, **336**, 137–156.
- Bonchev, D. (1997) Novel indices for the topological complexity of molecules. *SAR & QSAR Environ. Res.*, **7**, 23–43.
- Bonchev, D. (1999) Overall connectivity and molecular complexity: a new tool for QSPR/QSAR, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 361–401.
- Bonchev, D. (2000) Overall connectivities/topological complexities: a new powerful tool for QSPR/QSAR. *J. Chem. Inf. Comput. Sci.*, **40**, 934–941.
- Bonchev, D. (2001a) Overall connectivity – a next generation molecular connectivity. *J. Mol. Graph. Model.*, **20**, 65–75.
- Bonchev, D. (2001b) The overall Wiener index – a new tool for characterization of molecular topology. *J. Chem. Inf. Comput. Sci.*, **41**, 582–592.
- Bonchev, D. (2003a) On the complexity of directed biological networks. *SAR & QSAR Environ. Res.*, **14**, 199–214.
- Bonchev, D. (2003b) Shannon's information and complexity, in *Complexity: Introduction and Fundamentals*, Vol. 7 (eds D. Bonchev and D.E. Rouvray), Taylor & Francis, London, UK, pp. 157–187.
- Bonchev, D. (2005) My life-long journey in mathematical chemistry. *Internet Electron. J. Mol. Des.*, **4**, 434–490.
- Bonchev, D. and Balaban, A.T. (1981) Topological centric coding and nomenclature of polycyclic hydrocarbons. I. Condensed benzenoid systems (polyhexes, fusenes). *J. Chem. Inf. Comput. Sci.*, **21**, 223–229.
- Bonchev, D. and Balaban, A.T. (1993) Central vertices versus central rings in polycyclic systems. *J. Math. Chem.*, **14**, 287–304.
- Bonchev, D., Balaban, A.T. Liu, X. and Klein, D.J. (1994) Molecular cyclicity and centrality of polycyclic graphs. I. Cyclicity based on resistance distances or reciprocal distances. *Int. J. Quant. Chem.*, **50**, 1–20.
- Bonchev, D., Balaban, A.T. and Mekenyan, O. (1980) Generalization of the graph center concept and derived topological centric indexes. *J. Chem. Inf. Comput. Sci.*, **20**, 106–113.
- Bonchev, D., Balaban, A.T. and Randić, M. (1981) The graph center concept for polycyclic graphs. *Int. J. Quant. Chem.*, **19**, 61–82.
- Bonchev, D. and Buck, G.A. (2005) Quantitative measures of network complexity, in *Complexity in Chemistry Biology and Ecology* (eds D. Bonchev and D.H. Rouvray), Springer, New York, pp. 191–235.
- Bonchev, D. and Buck, G.A. (2007) From molecular to biological structure and back. *J. Chem. Inf. Model.*, **47**, 909–917.
- Bonchev, D. and Gordeeva, E.V. (1995) Topological atomic charges, valencies, and bond orders. *J. Chem. Inf. Comput. Sci.*, **35**, 383–395.
- Bonchev, D., Gutman, I. and Polanski, J. (1987) Parity of the distance numbers and Wiener numbers of bipartite graphs. *MATCH Commun. Math. Comput. Chem.*, **22**, 209–214.
- Bonchev, D. and Kamenska, V. (1978) Information theory in describing the electronic structures of atoms. *Croat. Chem. Acta*, **51**, 19–27.
- Bonchev, D., Kamenska, V. and Tashkova, C. (1976) Equations for the elements in the periodic table, based on information theory. *MATCH Commun. Math. Comput. Chem.*, **2**, 117–122.
- Bonchev, D., Kamenski, D. and Kamenska, V. (1976) Symmetry and information content of chemical structures. *Bull. Math. Biol.*, **38**, 119–133.
- Bonchev, D. and Kier, L.B. (1992) Topological atomic indices and the electronic charges in alkanes. *J. Math. Chem.*, **9**, 75–85.
- Bonchev, D., Kier, L.B. and Mekenyan, O. (1993) Self-returning walks and fractional electronic charges of atoms in molecules. *Int. J. Quant. Chem.*, **46**, 635–649.
- Bonchev, D. and Klein, D.J. (2002) On the Wiener number of thorn trees, stars, rings, and rods. *Croat. Chem. Acta*, **75**, 613–620.
- Bonchev, D., Liu, X. and Klein, D.J. (1993) Weighted self-returning walks for structure–property correlations. *Croat. Chem. Acta*, **66**, 141–150.

- Bonchev, D., Markel, E.J. and Dekmezian, A.H. (2005) Long chain branch polymer chain dimensions: application of topology to the Zimm–Stockmayer model. *Polymer*, **43**, 203–222.
- Bonchev, D. and Mekenyan, O. (1980) Topological approach to the calculation of the π -electron energy and energy gap of infinite conjugated polymers. *Z. Naturforsch.*, **35**, 739–747.
- Bonchev, D. and Mekenyan, O. (1983) Comparability graphs and electronic spectra of condensed benzenoid hydrocarbons. *Chem. Phys. Lett.*, **98**, 134–138.
- Bonchev, D., Mekenyan, O. and Balaban, A.T. (1985) Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). IV. Recognition of graph isomorphism and graph symmetries. *MATCH Commun. Math. Comput. Chem.*, **18**, 83–89.
- Bonchev, D., Mekenyan, O. and Balaban, A.T. (1986) Algorithms for coding chemical compounds, in *Mathematics and Computational Concepts in Chemistry* (ed. N. Trinajstić), Ellis Horwood, Chichester, UK, pp. 34–47.
- Bonchev, D., Mekenyan, O. and Balaban, A.T. (1989) Iterative procedure for the generalized graph center in polycyclic graphs. *J. Chem. Inf. Comput. Sci.*, **29**, 91–97.
- Bonchev, D., Mekenyan, O. and Fritsche, H. (1980a) An approach to the topological modeling of crystal growth. *J. Cryst. Growth*, **49**, 90–96.
- Bonchev, D., Mekenyan, O. and Kamenska, V. (1992) A topological approach to the modeling of polymer properties (the TEMPO method). *J. Math. Chem.*, **11**, 107–132.
- Bonchev, D., Mekenyan, O. and Polansky, O.E. (1981a) Topological approach to the predicting of the electron energy characteristics of conjugated infinite polymers. II. PPP-calculations. *Z. Naturforsch.*, **36**, 643–646.
- Bonchev, D., Mekenyan, O. and Polansky, O.E. (1981b) Topological approach to the predicting of the electron energy characteristics of conjugated infinite polymers. III. The influence of some structural modifications. *Z. Naturforsch.*, **36a**, 647–650.
- Bonchev, D., Mekenyan, O. and Trinajstić, N. (1980b) Topological characterization of cyclic structures. *Int. J. Quant. Chem.*, **17**, 845–893.
- Bonchev, D., Mekenyan, O. and Trinajstić, N. (1981c) Isomer discrimination by topological information approach. *J. Comput. Chem.*, **2**, 127–148.
- Bonchev, D., Mekenyan, O., von Knop, J. and Trinajstić, N. (1979) On characterization of monocyclic structures. *Croat. Chem. Acta*, **52**, 361–367.
- Bonchev, D., Mountain, C.F., Seitz, W.A. and Balaban, A.T. (1993) Modeling the anticancer action of some retinoid compounds by making use of the OASIS method. *J. Med. Chem.*, **36**, 1562–1569.
- Bonchev, D. and Peev, T. (1973) A theoretic-information study of chemical elements. Mean information content of a chemical element. *God. Vissh khim. – Technol. Inst., Burgas, Bulg.*, **10**, 561–574.
- Bonchev, D., Peev, T. and Rousseva, B. (1976) Information study of atomic nuclei, information for proton–neutron composition. *MATCH Commun. Math. Comput. Chem.*, **2**, 123–137.
- Bonchev, D. and Polansky, O.E. (1987) On the topological complexity of chemical systems, in *Graph Theory and Topology in Chemistry* (eds R.B. King and D.H. Rouvray), Elsevier, Amsterdam, The Netherlands, pp. 126–158.
- Bonchev, D. and Rouvray, D.E. (eds) (1992) *Chemical Graph Theory: Reactivity and Kinetics*, Gordon and Breach Science Publishers, New York, p. 265.
- Bonchev, D. and Rouvray, D.H. (eds) (1991) *Chemical Graph Theory: Introduction and Fundamentals*, Gordon and Breach Science Publishers, Reading, UK, p. 266.
- Bonchev, D. and Rouvray, D.H. (eds) (1994) *Chemical Group Theory: Introduction and Fundamentals*, Gordon and Breach Science Publishers, New York, p. 262.
- Bonchev, D. and Rouvray, D.H. (eds) (1995) *Chemical Group Theory: Applications*, Gordon and Breach Science Publishers, Reading, UK, p. 243.
- Bonchev, D. and Rouvray, D.H. (eds) (1998) *Topology in Chemistry: Introduction and Fundamentals*, Gordon and Breach Science Publishers, Reading, UK, p. 324.
- Bonchev, D. and Rouvray, D.H. (eds) (2000) *Topology in Chemistry: Applications*, Gordon and Breach Science Publishers, Reading, UK, p. 351.
- Bonchev, D. and Rouvray, D.H. (eds) (2003) *Complexity in Chemistry: Introduction and Fundamentals*, Taylor & Francis, London, UK, p. 208.
- Bonchev, D. and Rouvray, D.H. (eds) (2007) *Complexity in Chemistry, Biology, and Ecology*, Springer, New York, p. 372.
- Bonchev, D. and Seitz, W.A. (1995) Relative atomic moments as squared principal eigenvector coefficients. *J. Chem. Inf. Comput. Sci.*, **35**, 237–242.
- Bonchev, D. and Seitz, W.A. (1996) The concept of complexity in chemistry, in *Concepts in Chemistry: Contemporary Challenge* (ed. D.H. Rouvray), Research Studies Press, Taunton, UK, pp. 353–381.

- Bonchev, D., Seitz, W.A., Mountain, C.F. and Balaban, A.T. (1994) Modeling the anticarcinogenic action of retinoids by making use of the OASIS method. III. Inhibition of the induction of ornithine decarboxylase by arotinoids. *J. Med. Chem.*, **37**, 2300–2307.
- Bonchev, D. and Trinajstić, N. (1977) Information theory, distance matrix, and molecular branching. *J. Chim. Phys.*, **67**, 4517–4533.
- Bonchev, D. and Trinajstić, N. (1978) On topological characterization of molecular branching. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **12**, 293–303.
- Bonchev, D. and Trinajstić, N. (1982) Chemical information theory: structural aspects. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **16**, 463–480.
- Bonchev, D. and Trinajstić, N. (2001) Overall molecular descriptors. 3. Overall Zagreb indices. *SAR & QSAR Environ. Res.*, **12**, 213–235.
- Bonchev, D., von Knop, J. and Trinajstić, N. (1979) Mathematical models of branching. *MATCH Commun. Math. Comput. Chem.*, **6**, 21–47.
- Bonckaert, P. and Egghe, L. (1991) Rational normalization of concentration measures. *Journal of the American Society for Information Science*, **42**, 715–722.
- Bondi, A. (1964) van der Waals volumes and radii. *J. Phys. Chem.*, **68**, 441–451.
- Bonelli, D., Cechetti, V., Clementi, S., Cruciani, G., Fravolini, A. and Savino, A.F. (1991) The antibacterial activity of quinolones against *Escherichia coli*: a chemometric study. *Quant. Struct. -Act. Relat.*, **10**, 333–343.
- Bonjean, M.-C. and Luu Duc, C. (1978) Connectivité Moléculaire: Relation dans une Série de Barbituriques. *Eur. J. Med. Chem.*, **13**, 73–76.
- Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M. and Wade, R.C. (1989) New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.*, **32**, 1083–1094.
- Boon, G., De Proft, F., Langenaeker, W. and Geerlings, P. (1998) The use of density functional theory-based reactivity descriptors in molecular similarity calculations. *Chem. Phys. Lett.*, **295**, 122–128.
- Boon, G., Langenaeker, W., De Proft, F., De Winter, H., Tollenaere, J.P. and Geerlings, P. (2001) Systematic study of the quality of various quantum similarity descriptors. Use of the autocorrelation function and principal component analysis. *J. Phys. Chem. A*, **105**, 8805–8814.
- Boon, G., Van Alsenoy, C., De Proft, F., Bultinck, P. and Geerlings, P. (2005) Molecular quantum similarity of enantiomers of amino acids: a case study. *J. Mol. Struct. (Theochem)*, **727**, 49–56.
- Booth, D.E., Isenhour, T.L., Mahaney, J.K., Jr, Suh, M. and Wright, C. (2003) Quality control and data analysis, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 410–422.
- Booth, T.D. and Wainer, I.W. (1996a) Investigation of the enantioselective separations of alpha-alkylarylcarboxylic acids on an amylose tris(3,5-dimethylphenylcarbamate) chiral stationary phase using quantitative structure–enantioselective retention relationships. Identification of a conformationally driven chiral recognition mechanism. *J. Chromat.*, **737**, 157–169.
- Booth, T.D. and Wainer, I.W. (1996b) Mechanistic investigation into the enantioselective separation of mexiletine and related compounds, chromatographed on an amylose tris(3,5-dimethylphenylcarbamate) chiral stationary phase. *J. Chromat.*, **741**, 205–211.
- Bordwell, F.G. and Cooper, G.D. (1952) Conjugative effects of methylsulfonyl and methylthio groupings. *J. Am. Chem. Soc.*, **74**, 1058.
- Borodina, Y., Filimonov, D. and Poroikov, V.V. (1998) Computer-aided estimation of synthetic compounds similarity with endogenous bioregulations. *Quant. Struct. -Act. Relat.*, **17**, 459–464.
- Borodina, Y., Filimonov, D. and Poroikov, V.V. (2002) Computer-aided prediction of receptor profile for drug-like compounds. *SAR & QSAR Environ. Res.*, **13**, 433–443.
- Borosy, A.P., Balogh, B. and Mátyus, P. (2005) Alignment-free descriptors for quantitative structure–rate constant relationships of [4 + 2] cycloadditions. *J. Mol. Struct. (Theochem)*, **729**, 169–176.
- Borosy, A.P., Morvay, M. and Mátyus, P. (1999) 3D QSAR analysis of novel 5-HT_{1A} receptor ligands. *Chemom. Intell. Lab. Syst.*, **47**, 239–252.
- Borowski, T., Krol, M., Broclawik, E., Baranowski, T., Strekowski, L. and Mokrosz, M.J. (2000) Application of similarity matrices and genetic neural networks in quantitative structure–activity relationships of 2- or 4-(4-methylpiperazino) pyrimidines: 5-HT_{2A} receptor antagonists. *J. Med. Chem.*, **43**, 1901–1909.
- Borth, D.M. (1996) Optimal experimental designs for (possibly) censored data. *Chemom. Intell. Lab. Syst.*, **32**, 25–35.
- Borth, D.M. and McKay, R.J. (1985) A difficulty information approach to substituent selection in QSAR studies. *Technometrics*, **27**, 25–35.
- Bosnjak, N., Aldler, N., Perić, M. and Trinajstić, N. (1987) On the structural origin of chromatographic

- retention data: alkanes and cycloalkanes, in *Modelling of Structure and Properties of Molecules* (ed. Z.B. Maksić), Ellis Horwood, Chichester, UK, pp. 103–122.
- Bosque, R. and Sales, J. (2003) A QSPR study of O–H bond dissociation energy in phenols. *J. Chem. Inf. Comput. Sci.*, **43**, 637–642.
- Boström, J., Böhm, M., Gundertofte, K. and Klebe, G. (2003) A 3D QSAR study on a set of dopamine D4 receptor antagonists. *J. Chem. Inf. Comput. Sci.*, **43**, 1020–1027.
- Böttcher, C.J.F., van Belle, O.C., Bordewijk, P. and Rip, A. (1973) *Theory of Electric Polarization*, Vol. 1, Elsevier, Amsterdam, The Netherlands, p. 378.
- Botzki, A., Salmen, S., Bernhardt, G., Buschauer, A. and Dove, S. (2005) Structure-based design of bacterial hyaluronan lyase inhibitors. *QSAR Comb. Sci.*, **24**, 458–469.
- Boudreau, R.J. and Efange, S.M. (1992) Computer-aided radiopharmaceutical design. *Invest. Radiol.*, **27**, 653–658.
- Boulu, L.G. and Crippen, G.M. (1989) Voronoi binding site models: calculation of binding modes and influence of drug binding data accuracy. *J. Comput. Chem.*, **10**, 673–682.
- Boulu, L.G., Crippen, G.M., Barton, H.A., Kwon, H. and Marletta, M.A. (1990) Voronoi binding site model of a polycyclic aromatic hydrocarbon binding protein. *J. Med. Chem.*, **33**, 771–775.
- Bowden, K. (1990) Electronic effects in drug design, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 205–238.
- Bowden, K. and Wooldridge, K.R.H. (1973) Structure–activity relations. 3. Bronchodilator activity of substituted 6-thioxanthines. *Biochem. Pharmacol.*, **22**, 1015–1021.
- Bowden, K. and Young, R.C. (1970) Structure–activity relations. I. A series of antagonists of acetylcholine and histamine at the postganglionic receptors. *J. Med. Chem.*, **13**, 225–230.
- Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978) *Statistics for Experimenters*, John Wiley & Sons, Inc., New York, p. 653.
- Boyd, J.C., Millership, J.S. and Woolfson, A.D. (1982) The relationship between molecular connectivity and partition coefficients. *J. Pharm. Pharmacol.*, **34**, 364–366.
- Boyd, R.J. and Edgecombe, K.E. (1988) Atomic and group electronegativities from the electron density distributions of molecules. *J. Am. Chem. Soc.*, **110**, 4182–4186.
- Boyd, R.J. and Markus, G.E. (1981) Electronegativities of the elements from a nonempirical electrostatic model. *J. Chim. Phys.*, **75**, 5385–5388.
- Bradbury, S.P. (1994) Predicting modes of toxic action from chemical structure: an overview. *SAR & QSAR Environ. Res.*, **2**, 89–104.
- Bradbury, S.P. (1995) Quantitative structure–activity relationships and ecological risk assessment. An overview of predictive aquatic toxicology research. *Toxicol. Lett.*, **79**, 229–237.
- Bradbury, S.P., Henry, T.R., Niemi, G.J., Carlson, R. W. and Snarski, V.M. (1989) Use of respiratory–cardiovascular responses of rainbow trout (*Salmo gairdneri*) in identifying acute toxicity syndromes in fish. Part 3. Polar narcotics, *Environ. Toxicol. Chem.*, **8**, 247–261.
- Bradbury, S.P. and Lipnick, R.L. (1990) Introduction: structural properties for determining mechanisms of toxic action. *Environ. Health Persp.*, **87**, 181–182.
- Bradbury, S.P., Mekenyan, O. and Ankley, G.T. (1996) Quantitative structure–activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity. An assessment of conformer flexibility. *Environ. Toxicol. Chem.*, **15**, 1945–1954.
- Bradfield, C. (2004) Genomics and proteomics. *Chem. Res. Toxicol.*, **17**, 2.
- Bradley, E.K., Beroza, P., Penzotti, J.E., Grootenhuis, P.D.J., Spellmeyer, D.C. and Miller, J.L. (2000) A rapid computational method for lead evolution: description and application to α_1 -adrenergic antagonists. *J. Med. Chem.*, **43**, 2770–2774.
- Bradley, M. and Waller, C.L. (2001) Polarizability fields for use in three-dimensional quantitative structure–activity relationship (3D-QSAR). *J. Chem. Inf. Comput. Sci.*, **41**, 1301–1307.
- Bradley, M.P. (2002) An overview of the diversity represented in commercially-available databases. *J. Comput. Aid. Mol. Des.*, **16**, 301–309.
- Bradley, M.P. and Crippen, G.M. (1993) Voronoi modeling: the binding of triazines and pyrimidines to *L. casei* dihydrofolate reductase. *J. Med. Chem.*, **36**, 3171–3177.
- Braga, R.S., Barone, P.M.V.B. and Galvão, D.S. (1999) Identifying carcinogenic activity of methylated polycyclic aromatic hydrocarbons (PAHs). *J. Mol. Struct. (Theochem)*, **464**, 257–266.
- Braga, R.S., Vendrame, R. and Galvão, D.S. (2000) Structure–activity relationship studies of substituted 17α -acetoxyprogesterone hormones. *J. Chem. Inf. Comput. Sci.*, **40**, 1377–1385.
- Braga, S.F. and Galvão, D.S. (2003) A structure–activity study of taxol, taxotere, and derivatives using the electronic indices methodology (EIM). *J. Chem. Inf. Comput. Sci.*, **43**, 699–706.
- Braga, S.F. and Galvão, D.S. (2004) Benzo[c]quinolizin-3-ones theoretical investigation: SAR

- analysis and application to nontested compounds. *J. Chem. Inf. Comput. Sci.*, **44**, 1987–1997.
- Branch, G.E. and Calvin, M. (1941) *The Theory of Organic Chemistry*, Prentice-Hall, Inc., New York.
- Bratsch, S.G. (1985) A group electronegativity method with Pauling units. *J. Chem. Educ.*, **62**, 101–103.
- Braun, J., Kerber, A., Meringer, M. and Rücker, C. (2005) Similarity of molecular descriptors: the equivalence of Zagreb indices and walk counts. *MATCH Commun. Math. Comput. Chem.*, **54**, 163–176.
- Brauner, N. and Shacham, M. (1998) Role of range and precision of the independent variable in regression of data. *AIChE*, **44**, 603–611.
- Brauner, N., Shacham, M., Cholakov, G.S. and Stateva, R.P. (2005) Property prediction by similarity of molecular structures – practical application and consistency analysis. *Chem. Eng. Sci.*, **60**, 5458–5471.
- Brauner, N., Stateva, R.P., Cholakov, G.S. and Shacham, M. (2006) Structurally “targeted” quantitative structure–property relationship method for property prediction. *Ind. Eng. Chem. Res.*, **45**, 8430–8437.
- Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R. and Zaliani, A. (1997) MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D-QSAR study in a series of steroids. *J. Comput. Aid. Mol. Des.*, **11**, 79–92.
- Bravi, G. and Wikel, J.H. (2000a) Application of MS-WHIM descriptors: (1) introduction of new molecular surface properties and (2) prediction of binding affinity data. *Quant. Struct. -Act. Relat.*, **19**, 29–38.
- Bravi, G. and Wikel, J.H. (2000b) Application of MS-WHIM descriptors. 3. Prediction of molecular properties. *Quant. Struct. -Act. Relat.*, **19**, 39–49.
- Bray, P.J. and Barnes, R.G. (1957) Estimates of Hammett’s sigma values from quadrupole resonance studies. *J. Chim. Phys.*, **27**, 551–560.
- Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **26**, 123–140.
- Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
- Brekke, T. (1989) Prediction of physical properties of hydrocarbons mixtures by partial-least-squares calibration of carbon-13 nuclear magnetic resonance data. *Anal. Chim. Acta*, **223**, 123–134.
- Brendlé, E. and Papirer, E. (1997) A new topological index for molecular probes used in inverse gas chromatography. 2. Application for the evaluation of the solid surface specific interaction potential. *J. Colloid Interf. Sci.*, **194**, 217–224.
- Breneman, C.M., Sundling, C.M., Sukumar, N., Shen, L. Katt, W.P. and Embrechts, M.J. (2003) New developments in PESTshape/property hybrid descriptors. *J. Comput. Aid. Mol. Des.*, **17**, 231–240.
- Breneman, C.M., Thompson, T.R., Rhem, M. and Dung, M. (1995) Electron density modeling of large systems using the transferable atom equivalent method. *Computers Chem.*, **19**, 161–179.
- Brereton, R.G. (1990) *Chemometrics*, Ellis Horwood, Chichester, UK, p. 308.
- Breyer, E.D., Strasters, J.K. and Khaledi, M.G. (1991) Quantitative retention–biological activity relationship study by micellar liquid chromatography. *Anal. Chem.*, **63**, 828–833.
- Brickmann, J. (1997) Linguistic variables in the molecular recognition problem, in *Fuzzy Logic in Chemistry* (ed. D.H. Rouvray), Academic Press, New York, pp. 225–247.
- Briem, H. and Kuntz, I.D. (1996) Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.*, **39**, 3401–3408.
- Briem, H. and Lessel, U.F. (2000) *In vitro* and *in silico* affinity fingerprints: finding similarities beyond structural classes. *Pers. Drug Disc. Des.*, **20**, 231–244.
- Briens, F., Bureau, R., Rault, S. and Robba, M. (1995) Applicability of CoMFA in ecotoxicology: a critical study on chlorophenols. *Ecotox. Environ. Safety*, **31**, 37–48.
- Briggs, G.G. (1981) Theoretical and experimental relationships between soil adsorption, octanol–water partition coefficients, water solubilities, bioconcentration factors, and the parachor. *J. Agr. Food Chem.*, **29**, 1050–1059.
- Briggs, J.M., Marrone, T.J. and McCammon, J.A. (1996) Computational science: new horizons and relevance to pharmaceutical design. *Trends Cardiovasc. Med.*, **6**, 198–204.
- Brillouin, L. (1962) *Science and Information Theory*, 2nd edn, Academic Press, New York.
- Brinck, T., Murray, J.S. and Politzer, P. (1993) Octanol/water partition coefficients expressed in terms of solute molecular surface areas and electrostatic potentials. *J. Org. Chem.*, **58**, 7070–7073.
- Bringmann, G. and Rummey, C. (2003) 3D QSAR investigations on antimalarial naphthylisoquinoline alkaloids by comparative molecular similarity indices analysis (CoMSIA), based on different alignment approaches. *J. Chem. Inf. Comput. Sci.*, **43**, 304–316.
- Brint, A.T. and Willett, P. (1987a) Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.*, **27**, 152–158.

- Brint, A.T. and Willett, P. (1987b) Identifying 3-D maximal common substructures using transputer networks. *J. Mol. Graph.*, **5**, 200–207.
- Brooker, L.G.S., Craig, A.C., Heseltine, D.W., Jenkins, P.W. and Lincoln, L.L. (1965) Color and constitution. XIII. Merocyanines as solvent property indicators. *J. Am. Chem. Soc.*, **87**, 2443–2450.
- Broto, P. and Devillers, J. (1990) Autocorrelation of properties distributed on molecular graphs, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 105–127.
- Broto, P., Moreau, G. and Vandycke, C. (1984a) Molecular structures: perception, autocorrelation descriptor and SAR studies. Autocorrelation descriptor. *Eur. J. Med. Chem.*, **19**, 66–70.
- Broto, P., Moreau, G. and Vandycke, C. (1984b) Molecular structures: perception, autocorrelation descriptor and SAR studies. System of atomic contributions for the calculation of the *n*-octane/water partition coefficients. *Eur. J. Med. Chem.*, **19**, 71–78.
- Broto, P., Moreau, G. and Vandycke, C. (1984c) Molecular structures: perception, autocorrelation descriptor and SAR studies. Use of the autocorrelation descriptors in the QSAR study of two non-narcotic analgesic series. *Eur. J. Med. Chem.*, **19**, 79–84.
- Broughton, H.B., Green, S.M. and Rzepa, H.S. (1992) Rank correlation of AM1 and PM3 derived molecular electrostatic potentials (RACEL) with Hammett σ_p -parameters. *J. Chem. Soc. Chem. Comm.*, 37–39.
- Brown, F.K. (1998) Chemoinformatics: what is it and how does it impact drug discovery. *Annu. Rep. Med. Chem.*, **33**, 375–384.
- Brown, H.C. and Okamoto, Y. (1958) Electrophilic substituent constants. *J. Am. Chem. Soc.*, **80**, 4979–4987.
- Brown, H.C., Okamoto, Y. and Inukai, T. (1958) Rates of solvolysis of the *m*- and *p*-phenyl-, *m*- and *p*-methylthio-, and *m*- and *p*-trimethylsilylphenyldimethylcarbinyl chlorides. Steric inhibition of resonance as a factor in electrophilic substituent constants. *J. Am. Chem. Soc.*, **80**, 4964–4968.
- Brown, N., McKay, B. and Gasteiger, J. (2004) The *de novo* design of median molecules within a property range of interest. *J. Comput. Aid. Mol. Des.*, **18**, 761–771.
- Brown, N., McKay, B. and Gasteiger, J. (2005) Fingal: a novel approach to geometric fingerprinting and a comparative study of its application to 3D-QSAR modelling. *QSAR Comb. Sci.*, **24**, 480–484.
- Brown, N., McKay, B. and Gasteiger, J. (2006) A novel workflow for the inverse QSPR problem using multiobjective optimization. *J. Comput. Aid. Mol. Des.*, **20**, 333–341.
- Brown, R.D. (1997) Descriptors for diversity analysis. *Persp. Drug Disc. Des.*, **7/8**, 31–49.
- Brown, R.D. and Martin, Y.C. (1996) Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, **36**, 572–584.
- Brown, R.D. and Martin, Y.C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *J. Chem. Inf. Comput. Sci.*, **37**, 1–9.
- Brown, R.D. and Martin, Y.C. (1998) An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR & QSAR Environ. Res.*, **8**, 23–39.
- Brown, R.E. and Simas, A.M. (1982) On the applicability of CNDO indices for the prediction of chemical reactivity. *Theor. Chim. Acta*, **62**, 1–16.
- Brownlee, R.T.C., Katritzky, A.R. and Topsom, R.D. (1965) Direct infrared determination of the resonance interaction in monosubstituted benzenes. *J. Am. Chem. Soc.*, **87**, 3260–3261.
- Brownlee, R.T.C., Katritzky, A.R. and Topsom, R.D. (1966) Distortions of the π -electron system in monosubstituted benzenes. *J. Am. Chem. Soc.*, **88**, 1413–1419.
- Brüggemann, R., Altschuh, J. and Matthies, M. (1990) QSAR for estimating physico-chemical data, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 197–212.
- Brüggemann, R. and Bartel, H.-G. (1999) A theoretical concept to rank environmentally significant chemicals. *J. Chem. Inf. Comput. Sci.*, **39**, 211–217.
- Brüggemann, R., Bücherl, C., Pudenz, S. and Steinberg, E.W. (1999) Application of the concept of partial order on comparative evaluation of environmental chemicals. *Acta Hydrochim. Hydrobiol.*, **27**, 170–178.
- Brüggemann, R., Pudenz, S., Carlsen, L., Sørensen, P. B., Thomsen, M. and Mishra, R.K. (2001) The use of Hasse diagrams as a potential approach for inverse QSAR. *SAR & QSAR Environ. Res.*, **11**, 473–487.
- Bruneau, P. (2001) Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.*, **41**, 1605–1616.

- Bruni, A.T. and Ferreira, M.M.C. (2002) Omeprazole and analogue compounds: a QSAR study of activity against *Helicobacter pylori* using theoretical descriptors. *J. Chemom.*, **16**, 510–520.
- Bruno-Blanch, L. and Estiú, G.L. (1995) Quantum chemistry in QSAR anticonvulsant activity of VPA derivatives. *Int. J. Quant. Chem.*, **56**, 39–49.
- Bruschi, M., Giuffreda, M.G. and Lüthi, H.P. (2002) *trans* versus *geminal* electron delocalization in tetra- and diethynylethenes: a new method of analysis. *Chem. Eur. J.*, **8**, 4216–4227.
- Brusseau, M.L. (1993) Using QSAR to evaluate phenomenological models for sorption of organic compounds by soil. *Environ. Toxicol. Chem.*, **12**, 1835–1846.
- Brüstle, M., Beck, B., Schindler, T., King, W., Mitchell, T. and Clark, T. (2002) Descriptors, physical properties, and drug-likeness. *J. Med. Chem.*, **45**, 3345–3355.
- Brzezinska, E., Koska, G. and Klimczak, A. (2003) Application of thin-layer chromatographic data in quantitative structure–activity relationship assay of thiazole and benzothiazole derivatives with H-antihistamine activity. II. *J. Chromat.*, **1007**, 157–164.
- Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Computers Chem.*, **20**, 3–23.
- Buchwald, P. (2000) Modeling liquid properties, solvation, and hydrophobicity: a molecular size-based perspective. *Persp. Drug Disc. Des.*, **19**, 19–45.
- Buchwald, P. and Bodor, N. (1998) Octanol–water partition: searching for predictive models. *Curr. Med. Chem.*, **5**, 353–360.
- Buckingham, A.D. (1967) Permanent and induced molecular moments and long-range intermolecular forces, in *Advances in Chemical Physics*, Vol. 12 (ed. J.O. Hirschfelder), Wiley–Interscience, New York, pp. 107.
- Buckingham, R.A. (1938) The classical equation of state of gaseous He, Ne and Ar. *Proc. Roy. Soc. London A*, **168**, 264–283.
- Buckley, F. and Harary, F. (1990) *Distance in Graphs*, Addison-Wesley, Reading, MA.
- Bucknum, M.J. and Castro, E.A. (2005a) Introduction to molecular Schläfli indices. *MATCH Commun. Math. Comput. Chem.*, **54**, 121–136.
- Bucknum, M.J. and Castro, E.A. (2005b) Towards an empirical relation between the elementary polygonal circuit area and the topological form index, *l*, in the polyhedra and 2- and 3-dimensional structures. *MATCH Commun. Math. Comput. Chem.*, **54**, 313–330.
- Buda, A.B., Auf der Heyde, T. and Mislow, K. (1992) On quantifying chirality. *Angew. Chem. Int. Ed. Engl.*, **31**, 989–1007.
- Buda, A.B. and Mislow, K. (1992) A Hausdorff chirality measure. *J. Am. Chem. Soc.*, **114**, 6006–6012.
- Bultinck, P. and Carbó-Dorca, R. (2003) Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comput. Sci.*, **43**, 170–177.
- Bultinck, P., Carbó-Dorca, R. and Van Alsenoy, C. (2003) Quality of approximate electron densities and internal consistency of molecular alignment algorithms in molecular quantum similarity. *J. Chem. Inf. Comput. Sci.*, **43**, 1208–1217.
- Bultinck, P., Kuppens, T., Gironés, X. and Carbó-Dorca, R. (2003) Quantum similarity superposition algorithm (QSSA): a consistent scheme for molecular alignment and molecular similarity based on quantum chemistry. *J. Chem. Inf. Comput. Sci.*, **43**, 1143–1150.
- Bultinck, P., Langenaeker, W., Carbó-Dorca, R. and Tollenaere, J.P. (2003) Fast calculation of quantum chemical molecular descriptors from the electronegativity equalization method. *J. Chem. Inf. Comput. Sci.*, **43**, 422–428.
- Bundy, J.G., Morriss, A.W.J., Durham, D.G., Campbell, C.D. and Paton, G.I. (2001) Development of QSARs to investigate the bacterial toxicity and biotransformation potential of aromatic heterocyclic compounds. *Chemosphere*, **42**, 885–892.
- Bünz, A.P., Braun, B. and Janowsky, R. (1998) Application of quantitative structure–performance relationship and neural network models for the prediction of physical properties from molecular structure. *Ind. Eng. Chem. Res.*, **37**, 3044–3051.
- Buolamwini, J.K. and Assefa, H. (2002) CoMFA and CoMSIA 3D QSAR and docking studies on conformationally-restrained cinnamoyl HIV-1 integrase inhibitors: exploration of a binding mode at the active site. *J. Med. Chem.*, **45**, 841–852.
- Buontempo, F.V., Wang, X.Z., Mwense, M., Horan, N., Young, A., and Osborn, D. (2005) Genetic programming for the induction of decision trees to model ecotoxicity data. *J. Chem. Inf. Model.*, **45**, 904–912.
- Burbidge, R., Trotter, M., Buxton, B. and Holden, S. (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers Chem.*, **26**, 5–14.
- Burden, F.R. (1989) Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.*, **29**, 225–227.

- Burden, F.R. (1996) Using artificial neural networks to predict biological activity of molecules from simple molecular structural considerations. *Quant. Struct. -Act. Relat.*, **15**, 7–11.
- Burden, F.R. (1997) A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Quant. Struct. -Act. Relat.*, **16**, 309–314.
- Burden, F.R. (1998) Holographic neural networks as nonlinear discriminants for chemical applications. *J. Chem. Inf. Comput. Sci.*, **38**, 47–53.
- Burden, F.R. (2001) Quantitative structure–activity relationship studies using Gaussian processes. *J. Chem. Inf. Comput. Sci.*, **41**, 830–835.
- Burden, F.R., Brereton, R.G. and Walsh, P.T. (1997) A comparison of cross-validation and non-cross-validation techniques: application to polycyclic aromatic hydrocarbons electronic absorption spectra. *The Analyst*, **122**, 1015–1022.
- Burden, F.R., Ford, M.G., Whitley, D.C. and Winkler, D.A. (2000) Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *J. Chem. Inf. Comput. Sci.*, **40**, 1423–1430.
- Burden, F.R., Rosewarne, B.S. and Winkler, D.A. (1997) Predicting maximum bioactivity by effective inversion of neural networks using genetic algorithms. *Chemom. Intell. Lab. Syst.*, **38**, 127–137.
- Burden, F.R. and Winkler, D.A. (1999a) New QSAR methods applied to structure–activity mapping and combinatorial chemistry. *J. Chem. Inf. Comput. Sci.*, **39**, 236–242.
- Burden, F.R. and Winkler, D.A. (1999b) Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.*, **42**, 3183–3187.
- Burden, F.R. and Winkler, D.A. (2000) A quantitative structure–activity relationships model for the acute toxicity of substituted benzenes to *Tetrahymena pyriformis* using Bayesian-regularized neural networks. *Chem. Res. Toxicol.*, **13**, 436–440.
- Burdett, J.K. (1995) Topological aspects of chemical bonding and structure explored through the method of moments. *J. Mol. Struct. (Theochem)*, **336**, 115–136.
- Bureau, R., Daveu, C., Baglin, I., Sopkova-De Oliveira Santos, J., Lancelot, J.-C. and Rault, S. (2001) Association of two 3D QSAR analyses. Application to the study of partial agonist serotonin-3 ligands. *J. Chem. Inf. Comput. Sci.*, **41**, 815–823.
- Bureau, R., Daveu, C., Lancelot, J.C. and Rault, S. (2002a) Molecular design based on 3D-pharmacophore. Application to 5-HT subtypes receptors. *J. Chem. Inf. Comput. Sci.*, **42**, 429–436.
- Bureau, R., Daveu, C., Lemaître, S., Dauphin, F., Landelle, H., Lancelot, J.C. and Rault, S. (2002b) Molecular design based on 3D-pharmacophore. Application to 5-HT4 receptor. *J. Chem. Inf. Comput. Sci.*, **42**, 962–967.
- Burger, A. (1991) Isosterism and bioisosterism in drug design. *Progress in Drug Research*, **37**, 287–371.
- Burke, B.J. and Hopfinger, A.J. (1993) Advances in molecular shape analysis, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 276–306.
- Bursi, R., Dao, T., van Wijk, T., de Gooyer, M., Kellenbach, E. and Verwer, P. (1999) Comparative spectra analysis (CoSA): spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.*, **39**, 861–867.
- Bursi, R., Verwer, P., Gazit, A., Beccari, A.R., Uccello Beretta, G., Balzano, F. and Guccione, S. (2001) From molecular spectra to biological activities: a comparative spectra analysis (CoSA) study on epidermal growth factor receptor protein tyrosine kinase inhibitors, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 211–213.
- Burt, C. and Richards, W.G. (1990) Molecular similarity: the introduction of flexible fitting. *J. Comput. Aid. Mol. Des.*, **4**, 231–238.
- Burt, C., Richards, W.G. and Huxley, P. (1990) The application of molecular similarity calculations. *J. Comput. Chem.*, **11**, 1139–1146.
- Burton, J., Ijjaali, I., Barberan, O., Petitot, F., Vercauteren, D.P. and Michel, A. (2006) Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset. *J. Med. Chem.*, **49**, 6231–6240.
- Bush, B.L. and Sheridan, R.P. (1993) PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.*, **33**, 756–762.
- Butina, D. (2004) Performance of Kier–Hall E-state descriptors in quantitative structure–activity relationship (QSAR) studies of multifunctional molecules. *Molecules*, **9**, 1004–1009.
- Butina, D. and Gola, M.R. (2003) Modeling aqueous solubility. *J. Chem. Inf. Comput. Sci.*, **43**, 837–841.
- Butina, D., Segall, M.D. and Frankcombe, K. (2002) Predicting ADME properties *in silico*: methods and models. *Drug Discov. Today*, **7** (Suppl.), S83–S88.
- Buydens, L., Coomans, D., Vanbelle, M., Massart, D. L. and Vanden Driessche, R. (1983) Comparative study of topological and linear free energy-related parameters for the prediction of gas

- chromatographic retention indices. *J. Pharm. Sci.*, **72**, 1327–1329.
- Buydens, L. and Massart, D.L. (1981) Prediction of gas chromatographic retention indices from linear free energy and topological parameters. *Anal. Chem.*, **53**, 1990–1993.
- Buydens, L., Massart, D.L. and Geerlings, P. (1983) Prediction of gas chromatographic retention indexes with topological, physico-chemical, and quantum chemical parameters. *Anal. Chem.*, **55**, 738–744.
- Buydens, L., Massart, D.L. and Geerlings, P. (1985) Relationship between gas chromatographic behaviour and topological, physico-chemical and quantum chemically calculated charge parameters for neuroleptics. *J. Chromatogr. Sci.*, **23**, 304–307.
- Bytautas, L. and Klein, D.J. (1998) Chemical combinatorics for alkane-isomer enumeration and more. *J. Chem. Inf. Comput. Sci.*, **38**, 1063–1078.
- Bytautas, L. and Klein, D.J. (1999) Alkane isomer combinatorics: stereostructure enumeration and graph-invariant and molecular-property distributions. *J. Chem. Inf. Comput. Sci.*, **39**, 803–818.
- Bytautas, L. and Klein, D.J. (2000) Mean Wiener numbers and other mean extensions for alkane trees. *J. Chem. Inf. Comput. Sci.*, **40**, 471–481.
- Byvatov, E., Fechner, U., Sadowski, J. and Schneider, G. (2003) Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.*, **43**, 1882–1889.
- Cabala, R., Svobodova, J., Feltl, L. and Tichy, M. (1992) Direct determination of partition coefficients of volatile liquids between oil and gas by gas chromatography and its use in QSAR analysis. *Chromatographia*, **34**, 601–606.
- Caballero, J. and Fernández, M. (2006) Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks. *J. Mol. Model.*, **12**, 168–181.
- Caetano, S., Aires-de-Sousa, J., Daszykowski, M. and Vander Heyden, Y. (2005) Prediction of enantioselectivity using chirality codes and classification and regression trees. *Anal. Chim. Acta*, **544**, 315–326.
- Caetano, S., Decaestecker, T., Put, R., Daszykowski, M., Van Bocklaer, J. and Vander Heyden, Y. (2005) Exploring and modelling the responses of electrospray and atmospheric pressure chemical ionization techniques based on molecular descriptors. *Anal. Chim. Acta*, **550**, 92–106.
- Caianiello, E.R. (1953) On the quantum field theory. I. Explicit solution of Dyson's equation in electrodynamics without use of Feynman graphs. *Nuovo Cimento*, **10**, 1634–1652.
- Caianiello, E.R. (1956) Proprietà di Pfaffiani e Hafniani. *Ricerca Napoli*, **7**, 25–31.
- Calamari, D. and Vighi, M. (1992) A proposal to define quality objectives for aquatic life for mixtures of chemical substances. *Chemosphere*, **25**, 531–542.
- Calgarotto, A.K., Miotto, S., Honório, K.M., da Silva, A.B.F., Marangoni, S., Silva, J.L., Comar, M., Jr, Oliveira, K.M.T. and da Silva, S.L. (2007) A multivariate study on flavonoid compounds scavenging the peroxynitrite free radical. *J. Mol. Struct. (Theochem)*, **808**, 25–33.
- Caliendo, G., Fattorusso, C., Greco, G., Novellino, E., Perissutti, E. and Santagada, V. (1995) Shape-dependent effects in a series of aromatic nitro compounds acting as mutagenic agents on *S. typhimurium* TA98. *SAR & QSAR Environ. Res.*, **4**, 21–27.
- Caliendo, G., Greco, G., Novellino, E., Perissutti, E. and Santagada, V. (1994) Combined use of factorial design and comparative molecular field analysis (CoMFA): a case study. *Quant. Struct. -Act. Relat.*, **13**, 249–261.
- Calixto, F. and Raso, A. (1982) Retention index, connectivity index and van der Waals volume of alkanes. *Chromatographia*, **15**, 521.
- Camacho-Zuñiga, C. and Ruiz-Treviño, F.A. (2003) A new group contribution scheme to estimate the glass transition temperature for polymers and diluents. *Ind. Eng. Chem. Res.*, **42**, 1530–1534.
- Camarda, K.V. and Maranas, C.D. (1999) Optimization in polymer design using connectivity indices. *Ind. Eng. Chem. Res.*, **38**, 1884–1892.
- Camarda, K.V. and Sunderesan, P. (2005) An optimization approach to the design of value-added soybean oil products. *Ind. Eng. Chem. Res.*, **44**, 4361–4367.
- Camargo, A.J., Mercadante, R., Honório, K.M., Alves, C.N. and da Silva, A.B.F. (2002) A structure–activity relationship (SAR) study of synthetic neolignans and related compounds with biological activity against *Escherichia coli*. *J. Mol. Struct. (Theochem)*, **583**, 105–116.
- Cambon, B. and Devillers, J. (1993) New trends in structure–biodegradability relationships. *Quant. Struct. -Act. Relat.*, **12**, 49–56.
- Camilleri, P., Livingstone, D.J., Murphy, J.A. and Manallack, D.T. (1993) Chiral chromatography and multivariate quantitative structure–property relationships of benzimidazole sulphoxides. *J. Comput. Aid. Mol. Des.*, **7**, 61–69.
- Camilleri, P., Watts, S.A. and Boraston, J.A. (1988) A surface area approach to determination of partition

- coefficients. *J. Chem. Soc. Perkin Trans. 2*, 1699–1707.
- Cammarata, A. (1972) Interrelationship of the regression models used for structure–activity analyses. *J. Med. Chem.*, **15**, 573–577.
- Cammarata, A. (1979) Molecular topology and aqueous solubility of aliphatic alcohols. *J. Pharm. Sci.*, **68**, 839–842.
- Cammarata, A. and Bustard, T.M. (1974) Reinvestigation of a “nonadditive” quantitative structure–activity relationship. *J. Med. Chem.*, **17**, 981–985.
- Cammarata, A. and Yau, S.J. (1970) Predictability of correlations between *in vitro* tetracycline potencies and substituent indices. *J. Med. Chem.*, **13**, 93–97.
- Campbell, J.L. and Johnson, K.E. (1993) Abductive networks generalization, pattern recognition, and prediction of chemical behavior. *Can. J. Chem.*, **71**, 1800–1804.
- Can, H., Dimoglo, A. and Kovalishyn, V.V. (2005) Application of artificial neural networks for the prediction of sulfur polycyclic aromatic compounds retention indices. *J. Mol. Struct. (Theochem)*, **723**, 183–188.
- Canfield, E.R., Robinson, R.W. and Rouvray, D.H. (1985) Determination of the Wiener molecular branching index for the general tree. *J. Comput. Chem.*, **6**, 598–609.
- Cannon, E.O., Amini, A., Bender, A., Sternberg, M.J. E., Muggleton, S.H., Glen, R.C. and Mitchell, J.B. O. (2007) Support vector inductive logic programming outperforms the naive Bayes classifier and inductive logic programming for the classification of bioactive chemical compounds. *J. Comput. Aid. Mol. Des.*, **21**, 269–280.
- Cao, C. (1996) Distance-edge topological index for research on structure–property relationships of alkanes. *Acta Chim. Sin.*, **54**, 533–538.
- Cao, C. and Gao, S. (2005) Estimating enthalpies of formation of monoalkenes by the bonding orbital-connecting matrix of molecular graphics and the steric effect of the *cis/trans* configuration. *J. Mol. Struct. (Theochem)*, **718**, 153–163.
- Cao, C., and Gao, S. (2007) Bond orbital-connection matrix method to predict refractive indices of alkanes. *Chinese J. Chem. Phys.*, **20**, 149–154.
- Cao, C. and Li, Z. (1998) Molecular polarizability. 1. Relationship to water solubility of alkanes and alcohols. *J. Chem. Inf. Comput. Sci.*, **38**, 1–7.
- Cao, C. and Liu, L. (2004) Topological steric effect index and its application. *J. Chem. Inf. Comput. Sci.*, **44**, 678–687.
- Cao, C., Liu, S. and Li, Z. (1999) On molecular polarizability. 2. Relationship to the boiling point of alkanes and alcohols. *J. Chem. Inf. Comput. Sci.*, **39**, 1105–1111.
- Cao, C. and Luo, J. (2007) Topological electronegativity index and its application. I. Ionization potentials of alkyl groups and alkyl halides. *QSAR Comb. Sci.*, **26**, 955–962.
- Cao, C. and Yuan, H. (2001) Topological indices based on vertex, distance, and ring: on the boiling points of paraffins and cycloalkanes. *J. Chem. Inf. Comput. Sci.*, **41**, 867–877.
- Cao, C. and Yuan, H. (2002) On molecular polarizability. 4. Evaluation of the ionization potential for alkanes and alkenes with polarizability. *J. Chem. Inf. Comput. Sci.*, **42**, 667–672.
- Cao, C. and Yuan, H. (2003) A new approach of evaluating bond dissociation energy from eigenvalue of bonding orbital-connection matrix for C–C and C–H bonds in alkanes. *J. Chem. Inf. Comput. Sci.*, **43**, 600–608.
- Cao, C., Yuan, H., Liu, S. and Zeng, R. (2000) On molecular polarizability. 3. Relationship to the ionization potential of haloalkanes, amines, alcohols, and ethers. *J. Chem. Inf. Comput. Sci.*, **40**, 1010–1014.
- Cao, Y.L. and Li, H.Y. (2003) Application of genetic programming in predicting infinite dilution activity coefficients of organic compounds in water. *Chinese Chem. Lett.*, **14**, 987–990.
- Capelli, A.M., Feriani, A., Tedesco, G. and Pozzan, A. (2006) Generation of a focused set of GSK compounds biased toward ligand-gated ion-channel ligands. *J. Chem. Inf. Model.*, **46**, 659–664.
- Caporossi, G., Gutman, I. and Hansen, P. (1999) Variable neighborhood search for extremal graphs. IV. Chemical trees with extremal connectivity index. *Computers Chem.*, **23**, 469–477.
- Caporossi, G., Gutman, I., Hansen, P. and Pavlović, L. (2003) Graphs with maximum connectivity index. *Comp. Biol. Chem.*, **27**, 85–90.
- Caporossi, G. and Hansen, P. (1998) Enumeration of polyhex hydrocarbons to $h = 21$. *J. Chem. Inf. Comput. Sci.*, **38**, 610–619.
- Caprioara, M. and Diudea, M.V. (2003) QSAR modeling of polychlorinated aromatic compounds. *Indian J. Chem.*, **42**, 1368–1378.
- Carabédian, M. and Dubois, J.-E. (1998) Large virtual enhancement of a ^{13}C NMR database. A structural crowing extrapolation method enabling spectral data transfer. *J. Chem. Inf. Comput. Sci.*, **38**, 100–107.
- Carbó, R., Besalú, E., Amat, L. and Fradera, X. (1995) Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical

- foundation of quantitative structure-properties relationships (QSPR). *J. Math. Chem.*, **18**, 237–246.
- Carbó, R., Besalú, E., Amat, L. and Fradera, X. (1996) On quantum molecular similarity measures (QMSM) and indices (QMSI). *J. Math. Chem.*, **19**, 47–56.
- Carbó, R. and Calabuig, B. (1990) Molecular similarity and quantum chemistry, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 147–171.
- Carbó, R. and Calabuig, B. (1992a) Molecular quantum similarity measures and N-dimensional representation of quantum objects. I. Theoretical foundations. *Int. J. Quant. Chem.*, **42**, 1681–1693.
- Carbó, R. and Calabuig, B. (1992b) Molecular quantum similarity measures and N-dimensional representation of quantum objects. II. Practical applications. *Int. J. Quant. Chem.*, **42**, 1695–1709.
- Carbó, R. and Calabuig, B. (1992c) Quantum similarity, in *Structure, Interactions and Reactivity* (ed. S. Fraga), Elsevier, Amsterdam, The Netherlands, pp. 300–324.
- Carbó, R. and Calabuig, B. (1992d) Quantum similarity measures, molecular cloud description, and structure-properties relationships. *J. Chem. Inf. Comput. Sci.*, **32**, 600–606.
- Carbó, R. and Domingo, L. (1987) LCAO–MO similarity measures and taxonomy. *Int. J. Quant. Chem.*, **32**, 517–545.
- Carbó, R., Leyda, L. and Arnau, M. (1980) How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quant. Chem.*, **17**, 1185–1189.
- Carbó-Dorca, R. (2001) Inward matrix products: extensions and applications to quantum mechanical foundations of QSAR. *J. Mol. Struct. (Theochem)*, **537**, 41–54.
- Carbó-Dorca, R. and Besalú, E. (1996) Extending molecular similarity to energy surfaces: Boltzmann similarity measures and indices. *J. Math. Chem.*, **20**, 247–261.
- Carbó-Dorca, R. and Besalú, E. (1998) A general survey of molecular quantum similarity. *J. Mol. Struct. (Theochem)*, **451**, 11–23.
- Carbó-Dorca, R. and Mezey, P.G. (eds) (1998) *Advances in Molecular Similarity*, JAI Press, London, UK, p. 297.
- Cardoso, D.R., Andrade-Sobrinho, L.G., Leite-Neto, A.F., Reche, R.V., Isique, W.D., Ferreira, M.M., Lima-Neto, B.S. and Franco, D.W. (2004) Comparison between cachaça and rum using pattern recognition methods. *J. Agr. Food Chem.*, **52**, 3429–3433.
- Cardozo, M.G., Iimura, Y., Sugimoto, H., Yamanishi, Y. and Hopfinger, A.J. (1992) QSAR analyses of the substituted indanone and benzylpiperidine rings of a series of indanone benzylpiperidine inhibitors of acetylcholinesterase. *J. Med. Chem.*, **35**, 584–589.
- Carhart, R.E. (1978) Erroneous claims concerning the perception of topological symmetry. *J. Chem. Inf. Comput. Sci.*, **18**, 108–110.
- Carhart, R.E., Smith, D.H. and Venkataraghavan, R. (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, **25**, 64–73.
- Carlsen, L. (2004) Giving molecules an identity. On the interplay between QSARs and partial order ranking. *Molecules*, **9**, 1010–1018.
- Carlsen, L. (2005) A QSAR approach to physico-chemical data for organophosphates with special focus on known and potential nerve agents. *Internet Electron. J. Mol. Des.*, **4**, 355–366.
- Carlsen, L., Lerche, D.B. and Sørensen, P.B. (2002) Improving the predicting power of partial order based QSARs through linear extensions. *J. Chem. Inf. Comput. Sci.*, **42**, 806–811.
- Carlsen, L., Sørensen, P.B. and Thomsen, M. (2001) Partial order ranking-based QSARs: estimation of solubilities and octanol–water partitioning. *Chemosphere*, **43**, 295–302.
- Carlsen, L., Sørensen, P.B., Thomsen, M. and Brüggemann, R. (2002) QSAR's based on partial order ranking. *SAR & QSAR Environ. Res.*, **13**, 153–165.
- Carlson, R. (1992) *Design and Optimization in Organic Synthesis*, Elsevier, Amsterdam, The Netherlands, p. 536.
- Carlton, T.S. (1998) Correlation of boiling points with molecular structure for chlorofluoroethanes. *J. Chem. Inf. Comput. Sci.*, **38**, 158–164.
- Carlucci, G., D'Archivio, A.A., Maggi, M.A., Mazzeo, P. and Ruggieri, F. (2007) Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure-retention relationships. *Anal. Chim. Acta*, **601**, 68–76.
- Caron, G., Carrupt, P.-A., Testa, B., Ermondi, G. and Gasco, A. (1996) Insight into the lipophilicity of the aromatic N-oxide moiety. *Pharm. Res.*, **13**, 1186–1190.
- Caron, G. and Ermondi, G. (2003) A comparison of calculated and experimental parameters as sources of structural information: the case of lipophilicity-related descriptors. *Mini Rev. Med. Chem.*, **3**, 821–830.
- Caron, G. and Ermondi, G. (2005) Calculating virtual log P in the alkane/water system ($\log P^{\text{Nalk}}$) and its

- derived parameters $\Delta \log P^{\text{Noct-alk}}$ and $\log D^{\text{Halk}}$. *J. Med. Chem.*, **48**, 3269–3279.
- Caron, G. and Ermondi, G. (2008) Lipophilicity: chemical nature and biological relevance, in *Molecular Drug Properties*, Vol. 37 (ed. R. Mannhold), Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 315–329.
- Caron, G., Gaillard, P., Carrupt, P.-A. and Testa, B. (1997) 34. Lipophilicity behavior of model and medicinal compounds containing a sulfide, sulfoxide, or sulfone moiety. *Helv. Chim. Acta*, **80**, 449–462.
- Carpenter, M.P. (1979) Similarity of Pratt's measure of class concentration to the Gini index. *Journal of the American Society for Information Science*, **30**, 108–110.
- Carrieri, A., Altomare, C., Barreca, M.L., Contento, A., Carotti, A. and Hansch, C. (1994) Papain catalyzed hydrolysis of aryl esters: a comparison of the Hansch, docking and CoMFA methods. *Il Farmaco*, **49**, 573–585.
- Carrieri, A., Carotti, A., Barreca, M.L. and Altomare, C. (2002) Binding models of reversible inhibitors to type-B monoamine oxidase. *J. Comput. Aid. Mol. Des.*, **16**, 769–778.
- Carrigan, S.W., Fox, P.C., Wall, M.E., Wani, M.C. and Bowen, J.P. (1997) Comparative molecular field analysis and molecular modeling studies of 20-(S)-camptothecin analogs as inhibitors of DNA topoisomerase I and anticancer/antitumor agents. *J. Comput. Aid. Mol. Des.*, **11**, 71–78.
- Carro, A.M., Campisi, B., Camelio, P. and Phan-Tan-Luu, R. (2002) Improving an EVM QSPR model for glass transition temperature prediction using optimal design. *Chemom. Intell. Lab. Syst.*, **62**, 79–88.
- Carroll, F.I., Mascarella, S.W., Kuzemko, M.A., Gao, Y.G., Abraham, P., Lewin, A.H., Boja, J.W. and Kuhar, M.J. (1994) Synthesis, ligand binding, and QSAR (CoMFA and classical) study of 3beta-(3'-substituted phenyl), 3beta-(4'-substituted phenyl), and 3beta-(3',4'-disubstituted phenyl) tropane-2beta-carboxylic acid methyl esters. *J. Med. Chem.*, **37**, 2865–2873.
- Carrupt, P.-A., Testa, B. and Gaillard, P. (1997) Computational approaches to lipophilicity: methods and applications, in *Reviews in Computational Chemistry*, Vol. 11 (eds K.B. Lipkowitz and D. Boyd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 241–315.
- Carter, P.G. (1949) An empirical equation for the resonance energy of polycyclic aromatic hydrocarbons. *Trans. Faraday Soc.*, **45**, 597–602.
- Carter, S., Trinajstić, N. and Nikolić, S. (1987) A note on the use of ID numbers in QSAR studies. *Acta Pharm. Jugosl.*, **37**, 37–42.
- Cartier, A. and Rivail, J.-L. (1987) Electronic descriptors in quantitative structure–activity relationships. *Chemom. Intell. Lab. Syst.*, **1**, 335–347.
- Caruso, L., Musumarra, G. and Katritzky, A.R. (1993) “Classical” and “magnetic” aromaticities as new descriptors for heteroaromatics in QSAR. Part 3 [1]. Principal properties for heteroaromatics. *Quant. Struct. -Act. Relat.*, **12**, 146–151.
- Casaban-Ros, E., Antón-Fos, G.M., Gálvez, J., Duarte, M.J. and García-Domenech, R. (1999) Search for new antihistaminic compounds by molecular connectivity. *Quant. Struct. -Act. Relat.*, **18**, 35–42.
- Casalegno, M., Benfenati, E. and Sello, G. (2005) An automated group contribution method in predicting aquatic toxicity: the diatomic fragment approach. *Chem. Res. Toxicol.*, **18**, 740–746.
- Casalegno, M., Benfenati, E. and Sello, G. (2006) Application of a fragment-based model to the prediction of the genotoxicity of aromatic amines. *Internet Electron. J. Mol. Des.*, **5**, 431–446.
- Casalegno, M. and Sello, G. (2005) Quantitative aquatic toxicity prediction: using group contribution and classification methods on polar and non-polar narcotics. *J. Mol. Struct. (Theochem)*, **727**, 71–80.
- Casañola-Martín, G.M., Khan, M.T.H., Marrero-Ponce, Y., Ather, A., Sultankhodzhayev, M.N. and Torrens, F. (2006) New tyrosinase inhibitors selected by atomic linear indices-based classification models. *Bioorg. Med. Chem. Lett.*, **16**, 324–330.
- Casañola-Martín, G.M., Marrero-Ponce, Y., Khan, M.T.H., Ather, A., Khan, K.M., Torrens, F. and Rotondo, R. (2007a) Dragon method for finding novel tyrosinase inhibitors: biosilico identification and experimental *in vitro* assays. *Eur. J. Med. Chem.*, **42**, 1370–1381.
- Casañola-Martín, G.M., Marrero-Ponce, Y., Khan, M.T.H., Ather, A., Sultan, S., Torrens, F. and Rotondo, R. (2007b) TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: evaluation of different classification model combinations using bond-based linear indices. *Bioorg. Med. Chem.*, **15**, 1483–1503.
- Cash, G.G. (1995a) A fast computer algorithm for finding the permanent of adjacency matrices. *J. Math. Chem.*, **18**, 115–119.
- Cash, G.G. (1995b) Correlation of physico-chemical properties of alkylphenols with their graph-theoretical ϵ parameter. *Chemosphere*, **31**, 4307–4315.

- Cash, G.G. (1995c) Heats of formation of polyhex polycyclic aromatic hydrocarbons from their adjacency matrices. *J. Chem. Inf. Comput. Sci.*, **35**, 815–818.
- Cash, G.G. (1995d) Prediction of inhibitory potencies of arenesulfonamides toward carbonic anhydrase using easily calculated molecular connectivity indexes. *Struct. Chem.*, **6**, 157–160.
- Cash, G.G. (1998) A simple means of computing the Kekulé structure count for toroidal polyhex fullerenes. *J. Chem. Inf. Comput. Sci.*, **38**, 58–61.
- Cash, G.G. (1999) A simple program for computing characteristic polynomials with mathematica. *J. Chem. Inf. Comput. Sci.*, **39**, 833–834.
- Cash, G.G. (2000a) Permanental polynomials of the smaller fullerenes. *J. Chem. Inf. Comput. Sci.*, **40**, 1207–1209.
- Cash, G.G. (2000b) The permanental polynomial. *J. Chem. Inf. Comput. Sci.*, **40**, 1203–1206.
- Cash, G.G. (2002a) A differential-operator approach to the permanental polynomial. *J. Chem. Inf. Comput. Sci.*, **45**, 1132–1135.
- Cash, G.G. (2002b) Relationship between the Hosoya polynomial and the hyper-Wiener index. *Appl. Math. Lett.*, **15**, 893–895.
- Cash, G.G. (2003) Immanants and immanantal polynomials of chemical graphs. *J. Chem. Inf. Comput. Sci.*, **43**, 1942–1946.
- Cash, G.G. and Gutman, I. (2004) The Laplacian permanental polynomial: formulas and algorithms. *MATCH Commun. Math. Comput. Chem.*, **51**, 129–136.
- Cash, G.G., Klavžar, S. and Petrovšek, M. (2002) Three methods for calculation of the hyper-Wiener index of molecular graphs. *J. Chem. Inf. Comput. Sci.*, **42**, 571–576.
- Castillo-Garit, J.A., Marrero-Ponce, Y. and Torrens, F. (2006) Atom-based 3D-chiral quadratic indices. Part 2. Prediction of the corticosteroid-binding globulin binding affinity of the 31 benchmark steroids data set. *Bioorg. Med. Chem.*, **14**, 2398–2408.
- Castillo-Garit, J.A., Marrero-Ponce, Y., Torrens, F. and Rotondo, R. (2007) Atom-based stochastic and non-stochastic three-dimensional-chiral bilinear indices and their applications to central chirality codification. *J. Mol. Graph. Model.*, **26**, 32–47.
- Castro, E.A., Gutman, I., Marino, D. and Peruzzo, P. (2002) Upgrading the Wiener index. *J. Serb. Chem. Soc.*, **67**, 647–651.
- Castro, E.A., Toropov, A.A., Netserova, A.I. and Nazarov, A.U. (2003) QSAR study of the toxic action of aliphatic compounds to the bacteria *Vibrio fischeri* based on correlation weighting of local graph invariants. *J. Mol. Struct. (Theochem)*, **639**, 129–135.
- Castro, E.A., Tueros, M. and Toropov, A.A. (2000) Maximum topological distances based indices as molecular descriptors for QSPR 2 – application to aromatic hydrocarbons. *Computers Chem.*, **24**, 571–576.
- Cauchy, A.L. (1813) Recherche sur les polyèdres – premier mémoire. *Journal de l'Ecole Polytechnique*, **9**, 66–86.
- Cavalli, A., Carloni, P. and Recanatini, M. (2007) Target-related applications of first principles quantum chemical methods in drug design. *Chem. Rev.*, **106**, 3497–3519.
- Cayley, A. (1857) On the theory of the analytical forms called trees. *Philosophical Magazine*, **13**, 172–176.
- Cayley, A. (1859) On the analytical forms called trees. Part 2. *Philosophical Magazine*, **18**, 374–378.
- Cayley, A. (1874) On the mathematical theory of isomers. *Philosophical Magazine*, **47**, 444–447.
- Cedeño, W. and Agrafiotis, D.K. (2003) Using particle swarms for the development of QSAR models based on *K*-nearest neighbor and kernel regression. *J. Comput. Aid. Mol. Des.*, **17**, 255–263.
- Centner, V., Massart, D.L., de Noord, O.E., De Jong, S., Vandeginste, B.G.M. and Sterna, C. (1996) Elimination of uninformative variables for multivariate calibration. *Anal. Chem.*, **68**, 3851–3858.
- Cercos-del-Pozo, R.A., Perez-Gimenez, F., Salabert-Salvador, M.T. and García-March, F.J. (2000) Discrimination and molecular design of new theoretical hypolipaeic agents using the molecular connectivity functions. *J. Chem. Inf. Comput. Sci.*, **40**, 178–184.
- Cerruti, L. (2005) The inherent complexity of chemistry, in *Complexity in the Living: A Problem Oriented Approach* (eds R. Benigni, A. Colosimo, A. Giuliani, P. Sirabella and J.P. Zbilut), ISS, Rome, Italy, pp. 3–20.
- Chakraborty, K. and Devakumar, C. (2005) Quantitative structure–activity relationship analysis as a tool to evaluate the mode of action of chemical hybridizing agents for wheat (*Triticum aestivum* L.). *J. Agr. Food Chem.*, **53**, 3468–3475.
- Chalk, A.J., Beck, B. and Clark, T. (2001) A temperature-dependent quantum mechanical/neural net model for vapor pressure. *J. Chem. Inf. Comput. Sci.*, **41**, 1053–1059.
- Chan, E.C.Y., Tan, W.L., Ho, P.C. and Fang, L.J. (2005) Modeling Caco-2 permeability of drugs using immobilized artificial membrane chromatography and physico-chemical descriptors. *J. Chromat.*, **1072**, 159–168.

- Chan, O., Gutman, I., Lam, T.-K. and Merris, R. (1998) Algebraic connections between topological indices. *J. Chem. Inf. Comput. Sci.*, **38**, 62–65.
- Chan, O., Lam, T.-K. and Merris, R. (1997) Wiener number as an immanent of the Laplacian of molecular graphs. *J. Chem. Inf. Comput. Sci.*, **37**, 762–765.
- Chandrakumar, K.R.S., Ghanty, T.K. and Ghosh, S.K. (2004) Relationship between ionization potential, polarizability, and softness: a case study of lithium and sodium metal clusters. *J. Phys. Chem. A*, **108**, 6661–6666.
- Chang, C.M., Jalbout, A.F. and Lin, C. (2003) Novel descriptors based on density functional theory for predicting divalent metal ions adsorbed onto silica–disiloxane cluster model study. *J. Mol. Struct. (Theochem)*, **664–665**, 27–35.
- Chapman, D. (1996) The measurement of molecular diversity: a three-dimensional approach. *J. Comput. Aid. Mol. Des.*, **10**, 501–512.
- Chapman, N.B. and Shorter, J. (eds) (1978) *Correlation Analysis in Chemistry*, Plenum Press, New York, p. 546.
- Charifson, P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P. (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.*, **42**, 5100–5109.
- Charton, M. (1963) The estimation of Hammett substituent constants. *J. Org. Chem.*, **28**, 3121–3124.
- Charton, M. (1964) Definition of “inductive” substituent constants. *J. Org. Chem.*, **29**, 1222–1227.
- Charton, M. (1969) The nature of the *ortho* effect. II. Composition of the Taft steric parameters. *J. Am. Chem. Soc.*, **91**, 615–620.
- Charton, M. (1971) The quantitative treatment of the *ortho* effect. *Prog. Phys. Org. Chem.*, **8**, 235–317.
- Charton, M. (1975) Steric effects. I. Esterification and acid-catalyzed hydrolysis of esters. *J. Am. Chem. Soc.*, **97**, 1552–1556.
- Charton, M. (1976) Steric effects. 7. Additional ν constants. *J. Org. Chem.*, **41**, 2217–2220.
- Charton, M. (1978a) Applications of linear free energy relationships to polycyclic arenes and to heterocyclic compounds, in *Correlation Analysis in Chemistry* (eds N.B. Chapman and J. Shorter), Plenum Press, New York, pp. 175–268.
- Charton, M. (1978b) Steric effects. 13. Composition of the steric parameter as a function of alkyl branching. *J. Org. Chem.*, **43**, 3995–4001.
- Charton, M. (1981) Electrical effect substituent constants for correlation analysis. *Prog. Phys. Org. Chem.*, **13**, 119–251.
- Charton, M. (1983) The upsilon steric parameter. X. Definition and determination, in *Steric Effects in Drug Design, Topics in Current Chemistry*, Vol. 114 (eds M. Charton and I. Motoc), Springer-Verlag, Berlin, Germany, pp. 57–91.
- Charton, M. (1984) The validity of the revised F and R electrical effect substituent parameters. *J. Org. Chem.*, **49**, 1997–2001.
- Charton, M. (1987) A general treatment of electrical effects. *Prog. Phys. Org. Chem.*, **16**, 287–315.
- Charton, M. (1990) The quantitative description of amino acid, peptide, and protein properties and bioactivities. *Prog. Phys. Org. Chem.*, **18**, 163–284.
- Charton, M. (ed.) (1996) *Advances in Quantitative Structure–Property Relationships*, JAI Press, Greenwich, UK, p. 229.
- Charton, M. (2003) The nature of topological parameters. I. Are topological parameters ‘fundamental properties’? *J. Comput. Aid. Mol. Des.*, **17**, 197–209.
- Charton, M. and Charton, B.I. (1978) Steric effects. Park 12. Substituents at phosphorus. *J. Org. Chem.*, **43**, 2383–2386.
- Charton, M. and Charton, B.I. (1982) The structural dependence of amino acid hydrophobicity parameter. *J. Theor. Biol.*, **99**, 629–644.
- Charton, M. and Charton, B.I. (eds) (1999) *Advances in Quantitative Structure–Property Relationships*, JAI Press, Stamford, CT, p. 257.
- Charton, M. and Charton, B.I. (eds) (2002) *Advances in Quantitative Structure–Property Relationships*, JAI Press, Amsterdam, The Netherlands, p. 228.
- Chastrette, M., Rajzmann, M., Chanon, M. and Purcell, K.F. (1985) Approach to a general classification of solvents using a multivariate statistical treatment of quantitative solvent parameters. *J. Am. Chem. Soc.*, **107**, 1–11.
- Chastrette, M., Zakarya, D. and El Mouaffek, A. (1986) Relations Structure–Odeur dans une Famille des Muscs Benzénique Nitrés. *Eur. J. Med. Chem.*, **21**, 505–510.
- Chastrette, M., Zakarya, D. and Peyraud, J.F. (1994) Structure–musk odour relationships for indan and tetralins using neural network. On the contribution of descriptors to classification. *Eur. J. Med. Chem.*, **29**, 343–348.
- Chattaraj, P.K., Chamorro, E. and Fuentealba, P. (1999) Chemical bonding and reactivity: a local thermodynamic viewpoint. *Chem. Phys. Lett.*, **314**, 114–121.
- Chattaraj, P.K. and Roy, D.R. (2007) Update 1: of electrophilicity index. *Chem. Rev.*, **107**, PR46–PR74.
- Chattaraj, P.K., Roy, D.R., Elango, M. and Subramanian, V. (2005) Stability and reactivity of

- all-metal aromatic and antiaromatic systems in light of the principles of maximum hardness and minimum polarizability. *J. Phys. Chem. A*, **109**, 9590–9597.
- Chatterjee, A., Balaji, T., Matsunaga, H. and Mizukami, F. (2006) A reactivity index study to monitor the role of solvation on the interaction of the chromophores with amino-functional silanol surface for colorimetric sensors. *J. Mol. Graph. Model.*, **25**, 208–218.
- Chaudry, U.A. and Popelier, P.L.A. (2004) Estimation of pK_a using quantum topological molecular similarity descriptors: application to carboxylic acids, anilines and phenols. *J. Org. Chem.*, **69**, 233–241.
- Chaumat, E., Chamel, A., Taillandier, G. and Tissut, M. (1992) Quantitative relationships between structure and penetration of phenylurea herbicides through isolated plant cuticles. *Chemosphere*, **24**, 189–200.
- Chavatte, P., Yous, S., Beaurain, N., Mésangeau, C., Ferry, G. and Lesieur, D. (2002) Three-dimensional quantitative structure–activity relationship of arylalkylamine *N*-acetyltransferase (AANAT) inhibitors: a comparative molecular field analysis. *Quant. Struct. -Act. Relat.*, **20**, 414–421.
- Cheeseman, M.A., Machuga, E.J. and Bailey, A.B. (1999) A tiered approach to threshold of regulation. *Food Chem. Toxicol.*, **37**, 387–412.
- Chem-X Software, Oxford Molecular Ltd, The Magdalen Centre, Oxford Science Park, Sandford-on-Thames, Oxford, UK.
- ChemDiverse/Chem-X, Chemical Design Ltd, Roundway House, Cromwell Park, Chipping Norton, Oxfordshire, UK.
- Chen, D., Yin, C.-S., Wang, X. and Wang, L.-S. (2004) Holographic QSAR of selected esters. *Chemosphere*, **57**, 1739–1745.
- Chen, G., Zheng, S., Luo, X., Shen, J., Zhu, W., Liu, H., Gui, C., Zhang, J., Zheng, M., Puah, C.M., Chen, K. and Jiang, H. (2005) Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score. *J. Comb. Chem.*, **7**, 398–406.
- Chen, H., Zhou, J. and Xie, G. (1998) PARM: a genetic evolved algorithm to predict bioactivity. *J. Chem. Inf. Comput. Sci.*, **38**, 243–250.
- Chen, H.F., Yao, X.-J., Petitjean, M., Xia, H., Yao, J.H., Panaye, A., Doucet, J.P. and Fan, B.T. (2004) Insight into the bioactivity and metabolism of human glucagon receptor antagonists from 3D-QSAR analyses. *QSAR Comb. Sci.*, **23**, 603–620.
- Chen, J. and Wang, X.Z. (2001) A new approach to near-infrared spectral data analysis using independent component analysis. *J. Chem. Inf. Comput. Sci.*, **41**, 992–1001.
- Chen, J., Feng, L., Liao, Y., Han, S., Wang, L.-S. and Hu, H. (1996) Using AM1 Hamiltonian in quantitative structure–properties relationship studies of alkyl(1-phenylsulfonyl) cycloalkane-carboxylates. *Chemosphere*, **33**, 537–546.
- Chen, J., Harner, T., Schramm, K.-W., Quan, X., Xue, X. and Kettrup, A. (2003) Quantitative relationships between molecular structures, environmental temperatures and octanol–air partition coefficients of polychlorinated biphenyls. *Comp. Biol. Chem.*, **27**, 405–421.
- Chen, J., Peijnenburg, W.J.G.M., Quan, X., Zhao, Y., Xue, D. and Yang, F. (1998a) The application of quantum chemical and statistical technique in developing quantitative structure–property relationships for the photohydrolysis quantum yields of substituted aromatic halides. *Chemosphere*, **37**, 1169–1186.
- Chen, J., Peijnenburg, W.J.G.M. and Wang, L.-S. (1998b) Using PM3 Hamiltonian, factor analysis and regression analysis in developing quantitative structure–property relationships for the photohydrolysis quantum yields of substituted aromatic halides. *Chemosphere*, **36**, 2833–2853.
- Chen, J., Quan, X., Peijnenburg, W.J.G.M. and Yang, F. (2001a) Quantitative structure–property relationships (QSPRs) on direct photolysis quantum yields of PCDDs. *Chemosphere*, **43**, 235–241.
- Chen, J., Quan, X., Schramm, K.-W., Kettrup, A. and Yang, F. (2001b) Quantitative structure–property relationships (QSPRs) on direct photolysis of PCDDs. *Chemosphere*, **45**, 151–159.
- Chen, J., Quan, X., Yan, Y., Yang, F. and Peijnenburg, W.J.G.M. (2001c) Quantitative structure–property relationship studies on direct photolysis of selected polycyclic aromatic hydrocarbons in atmospheric aerosol. *Chemosphere*, **42**, 263–270.
- Chen, J., Quan, X., Zhao, Y., Yan, Y. and Yang, F. (2001d) Quantitative structure–property relationship studies on *n*-octanol/water partitioning coefficients of PCDD/Fs. *Chemosphere*, **44**, 1369–1374.
- Chen, J. and Wang, L.-S. (1997) Using MTLSE model and AM1 Hamiltonian in quantitative structure–activity relationship studies of alkyl(1-phenylsulfonyl)cycloalkane-carboxylates. *Chemosphere*, **35**, 623–631.
- Chen, J., Xue, X., Schramm, K.-W., Quan, X., Yang, F. and Kettrup, A. (2002) Quantitative structure–property relationships for octanol–air partition coefficients of polychlorinated biphenyls. *Chemosphere*, **48**, 535–544.

- Chen, J., Xue, X., Schramm, K.-W., Quan, X., Yang, F. and Kettrup, A. (2003) Quantitative structure–property relationships for octanol–air partition coefficients of polychlorinated naphthalenes, chlorobenzenes and *p,p'*-DDT. *Comp. Biol. Chem.*, **27**, 165–171.
- Chen, Q., Wu, C., Maxwell, D., Krudy, G.A., Dixon, R. A.F. and You, T.J. (1999) A 3D QSAR analysis of *in vitro* binding affinity and selectivity of 3-isoxazolylylsulfonylaminothiophenes as endothelin receptor antagonists. *Quant. Struct. - Act. Relat.*, **18**, 124–133.
- Chen, S.-W., Li, Z. and Li, X. (2005) Prediction of antifungal activity by support vector machine approach. *J. Mol. Struct. (Theochem)*, **731**, 73–81.
- Chen, X. and Reynolds, C.H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **42**, 1407–1414.
- Chen, X., Rusinko, A. III and Young, S.S. (1998) Recursive partitioning analysis of a large structure–activity data set using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci.*, **38**, 1054–1062.
- Chen, Y., Chen, D., He, C. and Hu, S. (1999) Quantitative structure–activity relationships study of herbicides using neural networks and different statistical methods. *Chemom. Intell. Lab. Syst.*, **45**, 267–276.
- Cheng, C., Maggiora, G.M., Lajiness, M.S. and Johnson, M.A. (1996) Four association constants for relating molecular similarity measures. *J. Chem. Inf. Comput. Sci.*, **36**, 909–915.
- Cheng, H., Kontogeorgis, G.M. and Stenby, E.H. (2005) Correlation and prediction of environmental properties of alcohol ethoxylate surfactants using the UNIFAC method. *Ind. Eng. Chem. Res.*, **44**, 7255–7261.
- Cheng, Y. and Chen, M. (2003) An approach to comparative analysis of chromatographic fingerprints for assuring the quality of botanical drugs. *J. Chem. Inf. Comput. Sci.*, **43**, 1068–1076.
- Cheng, Y. and Yuan, H. (2006) Quantitative study of electrostatic and steric effects on physico-chemical property and biological activity. *J. Mol. Graph. Model.*, **24**, 219–226.
- Chepoi, V. (1996) On distances in benzenoid systems. *J. Chem. Inf. Comput. Sci.*, **36**, 1169–1172.
- Chepoi, V. and Klavžar, S. (1997) The Wiener index and the Szeged index of benzenoid systems in linear time. *J. Chem. Inf. Comput. Sci.*, **37**, 752–755.
- Cherkasov, A.R. (2003) Inductive electronegativity scale. Iterative calculation of inductive partial charges. *J. Chem. Inf. Comput. Sci.*, **43**, 2039–2047.
- Cherkasov, A.R. (2005) ‘Inductive’ descriptors. 10. Successful years in QSAR. *Curr. Comput. -Aided Drug Des.*, **1**, 21–42.
- Cherkasov, A.R., Galkin, V.I. and Cherkasov, R. (2000) “Inductive” electronegativity scale. 2. ‘Inductive’ analog of chemical hardness. *J. Mol. Struct. (Theochem)*, **497**, 115–121.
- Cherkasov, A.R., Galkin, V.I. and Cherkasov, R.A. (1998) A new approach to the theoretical estimation of inductive constants. *J. Phys. Org. Chem.*, **11**, 437–447.
- Cherkasov, A.R., Galkin, V.I. and Cherkasov, R.A. (1999) “Inductive” electronegativity scale. *J. Mol. Struct. (Theochem)*, **489**, 43–46.
- Cherkasov, A.R. and Jankovic, B. (2004) Application of ‘inductive’ QSAR descriptors for quantification of antibacterial activity of cationic polypeptides. *Molecules*, **9**, 1034–1052.
- Cherkasov, A.R. and Jonsson, M. (1998) Substituent effects on thermochemical properties of free radicals. New substituent scales for C-centered radicals. *J. Chem. Inf. Comput. Sci.*, **38**, 1151–1156.
- Cherkasov, A.R. and Jonsson, M. (1999) Substituent effects on thermochemical properties of C-, N-, O-, and S-centered radicals. Physical interpretation of substituent effects. *J. Chem. Inf. Comput. Sci.*, **39**, 1057–1063.
- Cherkasov, A.R., Jonsson, M. and Galkin, V.I. (1999) A novel approach to the analysis of substituent effects: quantitative description of ionization energies and gas basicity of amines. *J. Mol. Graph. Model.*, **17**, 28–42.
- Cherkasov, A.R., Sprou, D.G. and Chen, R. (2003) Three-dimensional correlation analysis – a novel approach to the quantification of substituent effects. *J. Phys. Chem. A*, **107**, 9695–9704.
- Cherqaoui, D., Villemin, D., Mesbah, A., Cense, J.-M. and Kvasnička, V. (1994) Use of a neural network to determine the normal boiling points of acyclic ethers, peroxides, acetals and their sulfur analogues. *J. Chem. Soc. Faraday Trans.*, **90**, 2015–2019.
- Chester, N.A., Famini, G.R., Haley, M.V., Kurnas, C.W., Sterling, P.A. and Wilson, L.Y. (1996) Aquatic toxicity of chemical agent simulants as determined by quantitative structure–activity relationships and acute bioassays. *ACS Symp. Ser.*, **643**, 191–204.
- Chickos, J.S., Nichols, G. and Ruelle, P. (2002) The estimation of melting points and fusion enthalpies using experimental solubilities, estimated total phase change entropies, and mobile order and disorder theory. *J. Chem. Inf. Comput. Sci.*, **42**, 368–374.
- Chicu, S.A. (2000) An approach to calculate the toxicity of simple organic molecules on the basis of

- QSAR analysis in *Hydractinia echinata* (Hydrozoa, Cnidaria). *Quant. Struct. -Act. Relat.*, **19**, 227–236.
- Chicu, S.A. and Berking, S. (1997) Interference with metamorphosis induction in the marine cnidaria *Hydractinia echinata* (Hydrozoa): a structure–activity relationship analysis of lower alcohols, aliphatic and aromatic hydrocarbons, thiophenes, tributyl tin and crude oil. *Chemosphere*, **34**, 1851–1866.
- Chiorboli, C., Piazza, R., Carassiti, V., Passerini, L., Pino, A., Tosato, M.L. and Riganelli, D. (1996) A model for the tropospheric persistency of hydrohalo alkanes. *Gazz. Chim. It.*, **126**, 685–694.
- Chiorboli, C., Piazza, R., Carassiti, V., Passerini, L. and Tosato, M.L. (1993a) Application of chemometrics to the screening of hazardous substances. Part II. Advances in the multivariate characterization and reactivity modelling of haloalkanes. *Chemom. Intell. Lab. Syst.*, **19**, 331–336.
- Chiorboli, C., Piazza, R., Carassiti, V., Passerini, L. and Tosato, M.L. (1993b) Modelling of ionization potential of halogenated aliphatics. *Quant. Struct. -Act. Relat.*, **12**, 38–43.
- Chiorboli, C., Piazza, R., Tosato, M.L. and Carassiti, V. (1993c) Atmospheric chemistry: rate constants of the gas-phase reactions between haloalkanes of environmental interest and hydroxyl radicals. *Coordin. Chem. Rev.*, **125**, 241–250.
- Chiu, T.-L. and So, S.-S. (2003) Genetic neural networks for functional approximation. *QSAR Comb. Sci.*, **22**, 519–526.
- Chiu, T.-L. and So, S.-S. (2004) Development of neural network QSPR models for Hansch substituent constants. 1. Method and validation. *J. Chem. Inf. Comput. Sci.*, **44**, 147–153.
- Cho, S.J., Garsia, M.L.S., Bier, J. and Tropsha, A. (1996) Structure-based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors. *J. Med. Chem.*, **39**, 5064–5071.
- Cho, S.J. and Hermsmeier, M.A. (2002) Genetic algorithm guided selection: variable selection and subset selection. *J. Chem. Inf. Comput. Sci.*, **42**, 927–936.
- Cho, S.J., Shen, C.F. and Hermsmeier, M.A. (2000) Binary formal inference-based recursive modeling using multiple atom and physico-chemical property class pair and torsion descriptors as decision criteria. *J. Chem. Inf. Comput. Sci.*, **40**, 668–680.
- Cho, S.J. and Tropsha, A. (1995) Cross-validated R^2 -guided region selection for comparative molecular field analysis (CoMFA): a simple method to achieve consistent results. *J. Med. Chem.*, **38**, 1060–1066.
- Cho, S.J., Zheng, W. and Tropsha, A. (1998) Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J. Chem. Inf. Comput. Sci.*, **38**, 259–268.
- Choho, K., Langenaeker, W., Van De Woude, G. and Geerlings, P. (1995) Reactivity of fullerenes. Quantum-chemical descriptors versus curvature. *J. Mol. Struct. (Theochem)*, **338**, 293–301.
- Choho, K., Langenaeker, W., Van De Woude, G. and Geerlings, P. (1996) Local softness and hardness as reactivity indices in the fullerenes C_{24} – C_{76} . *J. Mol. Struct. (Theochem)*, **362**, 305–315.
- Cholakov, G.S., Wakeham, W.A. and Stateva, R.P. (1999) Estimation of normal boiling points of hydrocarbons from descriptors of molecular structure. *Fluid Phase Equil.*, **163**, 21–42.
- Chong, I.-G. and Jun, C.-H. (2005) Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.*, **78**, 103–112.
- Chou, J.T. and Jurs, P.C. (1979) Computer-assisted computation of partial coefficients from molecular structures using fragment constants. *J. Chem. Inf. Comput. Sci.*, **19**, 172–178.
- Chroust, K., Pavlová, M., Prokop, Z., Mendel, J., Božková, K., Kubát, Z., Zajčková, V. and Damborsky, J. (2007) Quantitative structure–activity relationships for toxicity and genotoxicity of halogenated aliphatic compounds: wing spot test of *Drosophila melanogaster*. *Chemosphere*, **67**, 152–159.
- Chumakov, Y., Terletskaya, A., Dimoglo, A. and Andronati, S.A. (2000) The electron-conformational approach to QSAR study in series of benzodiazepine derivatives. *Quant. Struct. -Act. Relat.*, **19**, 443–447.
- Chuman, H., Goto, S., Karasawa, M., Sasaki, M., Nagashima, U., Nishimura, K. and Fujita, T. (2000a) Three-dimensional structure–activity relationships of synthetic pyrethroids. 1. Similarity in bioactive conformations and their structure–activity pattern. *Quant. Struct. -Act. Relat.*, **19**, 10–21.
- Chuman, H., Goto, S., Karasawa, M., Satake, M., Nagashima, U., Nishimura, K. and Fujita, T. (2000b) Three-dimensional structure–activity relationships of synthetic pyrethroids. 2. Three-dimensional and classical QSAR studies. *Quant. Struct. -Act. Relat.*, **19**, 455–466.
- Chuman, H., Karasawa, M. and Fujita, T. (1998) A novel 3-dimensional QSAR procedure – Voronoi field analysis. *Quant. Struct. -Act. Relat.*, **17**, 313–326.
- Chung, F.R.K. (1988) The average distance and the independence number. *J. Serb. Chem. Soc.*, **12**, 229–235.

- Chung, F.R.K. (1997) *Spectral Graph Theory*, AMM, Providence, RI.
- Churchwell, C.J., Rintoul, M.D., Martin, S., Visco, D., Kotu, A., Larson, R.S., Sillerud, L.O., Brown, D.C. and Faulon, J.-L. (2004) The signature molecular descriptor. 3. Inverse quantitative structure–activity relationship of ICAM-1 inhibitory peptides. *J. Mol. Graph. Model.*, **22**, 263–273.
- Cianchetta, G., Li, Y., Singleton, R., Zhang, M., Wildgoose, M., Rampe, D., Kang, J. and Vaz, R.J. (2006) Molecular interaction fields in ADME and safety, in *Molecular Interaction Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 197–218.
- Cianchetta, G., Mannhold, R., Cruciani, G., Baroni, M. and Cecchetti, V. (2004) Chemometric studies on the bactericidal activity of quinolones via an extended VolSurf approach. *J. Med. Chem.*, **47**, 3193–3201.
- Cinone, N., Hölte, H.-D. and Carotti, A. (2000) Development of a unique 3D interaction model of endogenous and synthetic peripheral benzodiazepine receptor ligands. *J. Comput. Aid. Mol. Des.*, **14**, 753–768.
- Ciofini, I., Bedioui, F., Zagal, J.H. and Adamo, C. (2003) Environment effects on the oxidation of thiols: cobalt phthalocyanine as a test case. *Chem. Phys. Lett.*, **376**, 690–697.
- Cioslowski, J. and Fleischmann, E.D. (1991) Assessing molecular similarity from results of *ab initio* electronic structure calculations. *J. Am. Chem. Soc.*, **113**, 64–67.
- Ciubotariu, D., Deretey, E., Oprea, T.I., Sulea, T., Simon, Z., Kurunczi, L. and Chiriac, A. (1993) Multiconformational minimal steric difference. Structure–acetylcholinesterase hydrolysis rates relations for acetic acid esters. *Quant. Struct. -Act. Relat.*, **12**, 367–372.
- Ciubotariu, D., Gogonea, V. and Medeleanu, M. (2001) van der Waals molecular descriptors. Minimal steric difference, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 281–361.
- Ciubotariu, D., Grozav, A., Gogonea, V., Ciubotariu, C., Medeleanu, M., Dragos, D., Pasere, M. and Simon, Z. (2001) The effects of retinoids upon epithelial differentiation of hamster trachea, induction of ornithine decarboxylase and promotion of tumors in mouse epidermis. A QSAR study by MTD method. *MATCH Commun. Math. Comput. Chem.*, **44**, 65–92.
- Ciubotariu, D., Medeleanu, M., Vlaia, V., Olariu, T., Ciubotariu, C., Dragos, D. and Corina, S. (2004) Molecular van der Waals space and topological indices from the distance matrix. *Molecules*, **9**, 1053–1078.
- Clare, B.W. (1994) Frontier orbital energies in quantitative structure–activity relationships: a comparison of quantum chemical methods. *Theor. Chim. Acta*, **87**, 415–430.
- Clare, B.W. (1995a) Charge transfer complexes and frontier orbital energies in QSAR. A congeneric series of electron acceptors. *J. Mol. Struct. (Theochem)*, **337**, 139–150.
- Clare, B.W. (1995b) Structure–activity correlations for psychotomimetics. 3. Tryptamines. *Aust. J. Chem.*, **48**, 1385–1400.
- Clare, B.W. (1995c) The relationship of charge transfer complexes to frontier orbital energies in QSAR. *J. Mol. Struct. (Theochem)*, **331**, 63–78.
- Clare, B.W. (2002) QSAR of benzene derivatives: comparison of classical descriptors, quantum theoretic parameters and flip regression, exemplified by phenylalkylamine hallucinogens. *J. Comput. Aid. Mol. Des.*, **16**, 611–633.
- Clare, B.W. (2004) A novel quantum theoretic QSAR for hallucinogenic tryptamines: a major factor is the orientation of π orbital nodes. *J. Mol. Struct. (Theochem)*, **712**, 143–148.
- Clare, B.W. and Supuran, C.T. (1994) Carbonic anhydrase activators. 3. Structure–activity correlations of a series of isozyme II activators. *J. Pharm. Sci.*, **83**, 768–773.
- Clare, B.W. and Supuran, C.T. (1998) Semi-empirical atomic charges and dipole moments in hypervalent sulfonamide molecules: descriptors in QSAR studies. *J. Mol. Struct. (Theochem)*, **428**, 109–121.
- Clark, A.M., Labute, P. and Sanravy, M. (2006) 2D structure depiction. *J. Chem. Inf. Model.*, **46**, 1107–1123.
- Clark, D.E. (1999a) Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.*, **88**, 807–814.
- Clark, D.E. (1999b) Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–barrier penetration. *J. Pharm. Sci.*, **88**, 815–821.
- Clark, D.E. and Murray, C.W. (1995) PRO_LIGAND: an approach to *de novo* molecular design. 5. Tools for the analysis of generated structures. *J. Chem. Inf. Comput. Sci.*, **35**, 914–923.
- Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of ‘drug-likeness’. *Drug Discov. Today*, **5**, 49–58.
- Clark, D.E., Willett, P. and Kenny, P. (1992) Pharmacophoric pattern matching in files of three-

- dimensional chemical structures: use of smoothed-bounded distance matrices for the representation and searching of conformationally-flexible molecules. *J. Mol. Graph.*, **10**, 194–204.
- Clark, D.E., Willett, P. and Kenny, P. (1993) Pharmacophoric pattern matching in files of three-dimensional chemical structures: implementation of flexible searching. *J. Mol. Graph.*, **11**, 146–156.
- Clark, M. and Cramer, R.D. III (1993) The probability of chance correlation using partial least squares (PLS). *Quant. Struct. -Act. Relat.*, **12**, 137–145.
- Clark, M., Cramer, R.D. III and Van Opdenbosch, N. (1989) Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.*, **10**, 982–1012.
- Clark, R.D. (2003a) Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics. *J. Comput. Aid. Mol. Des.*, **17**, 265–275.
- Clark, R.D. and Fox, P.C. (2004) Statistical variation in progressive scrambling. *J. Comput. Aid. Mol. Des.*, **18**, 563–576.
- Clark, R.D. and Langton, W.J. (1998) Balancing representativeness against diversity using optimizable K-dissimilarity and hierarchical clustering. *J. Chem. Inf. Comput. Sci.*, **38**, 1079–1086.
- Clark, R.D., Parlow, J.J., Brannigan, L.H., Schnur, D. M. and Duewer, D.L. (1995) Applications of scaled rank sum statistics in herbicide QSAR. *ACS Symp. Ser.*, **606**, 264–281.
- Clark, R.D., Sprou, D.G. and Leonard, J.M. (2001) Validating models based on large data sets, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 475–485.
- Clark, R.D., Strizhev, A., Leonard, J.M., Blake, J.F. and Matthew, J.B. (2002) Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.*, **20**, 281–295.
- Clark, T. (2001) Quantum chemoinformatics: an oxymoron? Part II, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 29–40.
- Clark, T. (2003b) Quantum mechanics, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 947–976.
- Clark, T. (2004) QSAR and QSPR based solely on surface properties? *J. Mol. Graph. Model.*, **22**, 519–525.
- Clementi, S., Cruciani, G., Baroni, M. and Costantino, G. (1993) Series design, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 567–582.
- Clementi, S., Cruciani, G., Fifi, P., Riganelli, D., Valigi, R. and Musumarra, G. (1996) A new set of principal properties for heteroaromatics obtained by GRID. *Quant. Struct. -Act. Relat.*, **15**, 108–120.
- Clementi, S., Cruciani, G., Riganelli, D., Valigi, R., Costantino, G., Baroni, M. and Wold, S. (1993b) Autocorrelation as a tool for a congruent description of molecules in 3D QSAR studies. *Pharm. Pharmacol. Lett.*, **3**, 5–8.
- Clerc, J.T. and Terkovich, A.L. (1990) Versatile topological structure descriptor for quantitative structure/property studies. *Anal. Chim. Acta*, **235**, 93–102.
- Clifford, A.F. (1959) The electronegativity of groups. *J. Phys. Chem.*, **63**, 1227–1231.
- Cnubben, N.H., Peelen, S., Borst, J.W., Vervoort, J., Veeger, C. and Rietjens, I.M.C.M. (1994) Molecular orbital based quantitative structure–activity relationship for the cytochrome P450 catalyzed 4 hydroxylation of halogenated anilines. *Chem. Res. Toxicol.*, **7**, 590–598.
- Coats, E.A. (1998) The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ ESCOM, Dordrecht, The Netherlands, pp. 199–213.
- Cocchi, M., Corbellini, M., Foca, G., Lucisano, M., Pagani, M.A., Tassi, L. and Ulrici, A. (2005) Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra. *Anal. Chim. Acta*, **544**, 100–107.
- Cocchi, M., De Benedetti, P.G., Seeber, R., Tassi, L. and Ulrici, A. (1999) Development of quantitative structure–property relationships using calculated descriptors for the prediction of the physico-chemical properties (n_D , ρ , bp, ϵ , η) of a series of organic solvents. *J. Chem. Inf. Comput. Sci.*, **39**, 1190–1203.
- Cocchi, M., Fanelli, F., Menziani, M.C. and De Benedetti, P.G. (1997) Conformational analysis and theoretical quantitative size and shape–affinity relationships of N_4 -protonated N_1 -arylpiperazine 5-HT_{1A} serotoninergic ligands. *J. Mol. Struct. (Theochem)*, **397**, 129–145.
- Cocchi, M. and Johansson, E. (1993) Amino acids characterization by GRID and multivariate data analysis. *Quant. Struct. -Act. Relat.*, **12**, 1–8.
- Cocchi, M., Menziani, M.C. and De Benedetti, P.G. (1992) Theoretical versus empirical molecular descriptors in monosubstituted benzenes. *Chemom. Intell. Lab. Syst.*, **14**, 209–224.
- Cocchi, M., Menziani, M.C., Fanelli, F. and De Benedetti, P.G. (1995) Theoretical quantitative

- structure–activity relationship analysis of congeneric and non-congeneric α_1 -adrenoceptor antagonists: a chemometric study. *J. Mol. Struct. (Theochem)*, **331**, 79–93.
- Cocchi, M., Seeber, M. and Ulrici, A. (2001) WPTER: wavelet packet transform for efficient pattern recognition of signals. *Chemom. Intell. Lab. Syst.*, **57**, 97–119.
- Cocchi, M., Seeber, R. and Ulrici, A. (2003) Multivariate calibration of analytical signals by WILMA (wavelet interface to linear modeling analysis). *J. Chemom.*, **17**, 512–527.
- Coccini, T., Giannoni, L., Karcher, W., Manzo, L. and Roi, R. (eds) (1992) *Quantitative Structure/Activity Relationships (QSAR) in Toxicology*, Joint Research Centre – EEC, Brussels, Belgium, pp. 90.
- CODESSA – Reference Manual, Ver. 2.0, Katritzky, A. R., Lobanov, V.S. and Karelson, M., Gainesville, FL.
- Cohen, J.L., Lee, W. and Lien, E.J. (1974) Dependence of toxicity on molecular structure: group theory analysis. *J. Pharm. Sci.*, **63**, 1068–1072.
- Cohen, L.A. and Jones, W.M. (1963) A study of free energy relationships in hindered phenols. Linear dependence for solvation effects in ionization. *J. Am. Chem. Soc.*, **85**, 3397–3402.
- Cohen, N. and Benson, S.W. (1987) Empirical correlations for rate coefficients of reactions of OH with haloalkanes. *J. Phys. Chem.*, **91**, 171–175.
- Collantes, E.R. and Dunn, W.J. III (1995) Amino acid side chain descriptors for quantitative structure–activity relationship studies of peptide analogues. *J. Med. Chem.*, **38**, 2705–2713.
- Collantes, E.R., Tong, W. and Welsh, W.J. (1996) Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. *Anal. Chem.*, **68**, 2038–2043.
- Coluci, V.R., Vendrame, R., Braga, R.S. and Gamper, A.M. (2002) Identifying relevant molecular descriptors related to carcinogenic activity of polycyclic aromatic hydrocarbons (PAHs) using pattern recognition methods. *J. Chem. Inf. Comput. Sci.*, **42**, 1479–1489.
- Combes, R.D. and Judson, P.N. (1995) The use of artificial intelligence systems for predicting toxicity. *Pestic. Sci.*, **45**, 179–194.
- Compadre, R.L.L., Byrd, C. and Compadre, C.M. (1998) Comparative QSAR and 3-D-QSAR analysis of the mutagenicity of nitroaromatic compounds, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 111–136.
- Cone, M.M., Venkataraghavan, R. and McLafferty, F. W. (1977) Molecular structure comparison program for the identification of maximal common substructures. *J. Am. Chem. Soc.*, **99**, 7668–7671.
- Congreve, M., Carr, R., Murray, C. and Jhoti, H. (2003) A ‘rule of three’ for fragment-based lead discovery? *Drug Discov. Today*, **8**, 876–877.
- Connolly, M.L. (1983a) Analytical molecular surface calculation. *J. Appl. Cryst.*, **16**, 548–558.
- Connolly, M.L. (1983b) Solvent-accessible surfaces of proteins and nucleic acids. *Science*, **221**, 709–713.
- Connolly, M.L. (1985) Computation of molecular volume. *J. Am. Chem. Soc.*, **107**, 1118–1124.
- Connolly, M.L. (1994) Adjoint join volumes. *J. Math. Chem.*, **15**, 339–352.
- Consonni, V. and Todeschini, R. (2001) Getaway descriptors: new molecular descriptors combining geometrical, topological and chemical information for physico-chemical properties modelling and drug design, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 235–240.
- Consonni, V. and Todeschini, R. (2008) New spectral indices for molecule description. *MATCH Commun. Math. Comput. Chem.*, **60**, 3–14.
- Consonni, V., Todeschini, R. and Pavan, M. (2002a) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. Part 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.*, **42**, 682–692.
- Consonni, V., Todeschini, R., Pavan, M. and Gramatica, P. (2002b) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. Part 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.*, **42**, 693–705.
- Constans, P. and Carbó, R. (1995) Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.*, **35**, 1046–1053.
- Contrera, J.F., MacLaughlin, P., Hall, L.H. and Kier, L. B. (2005) QSAR modeling of carcinogenic risk using discriminant analysis and topological molecular descriptors. *Curr. Drug Discov. Technol.*, **2**, 55–67.
- Convrad, T., Dubost, J.P., Le Solleu, H. and Kummer, E. (1994) SmilogP: a program for a fast evaluation of theoretical log P from the smiles code of a molecule. *Quant. Struct. -Act. Relat.*, **13**, 34–37.
- Corbella, R., Rodriguez, M.A., Sánchez, M.J. and Montelongo, F.G. (1995) Correlations between gas chromatographic retention data of polycyclic aromatic hydrocarbons and several molecular descriptors. *Chromatographia*, **40**, 532–538.
- Cosentino, U., Moro, G., Bonalumi, D., Bonati, L., Lasagni, M., Todeschini, R. and Pitea, D. (2000) A combined use of global and local approaches in 3D-QSAR. *Chemom. Intell. Lab. Syst.*, **52**, 183–194.

- Cosentino, U., Moro, G., Pitea, D., Scolastico, S., Todeschini, R. and Scolastico, C. (1992) Pharmacophore identification by molecular modeling and chemometrics: the case of HMG-CoA reductase inhibitors. *J. Comput. Aid. Mol. Des.*, **6**, 47–60.
- Costa, M.C.A., Gaudio, A.C. and Takahata, Y. (2003) Core electron binding energy (CEBE) shifts as descriptors in structure–activity relationship (SAR) analysis of cytotoxicities of a series of simple phenols. *J. Mol. Struct. (Theochem)*, **664–665**, 171–174.
- Costa, M.C.A. and Takahata, Y. (2003) Core electron binding energy (CEBE) shifts as descriptors in structure–activity relationship (SAR) analysis of neolignans tested against *Leishmania donovani*. *J. Mol. Struct. (Theochem)*, **638**, 21–25.
- Costescu, A. and Diudea, M.V. (2006) QSTR study on aquatic toxicity against *Poecilia reticulata* and *Tetrahymena pyriformis* using topological indices. *Internet Electron. J. Mol. Des.*, **5**, 116–134.
- Costescu, A., Moldovan, C. and Diudea, M.V. (2006) QSAR modeling of steroid hormones. *MATCH Commun. Math. Comput. Chem.*, **55**, 315–329.
- Cotta Ramusino, M., Benigni, R., Passerini, L. and Giuliani, A. (2003) Looking for an unambiguous geometrical definition of organic series from 3-D molecular similarity indices. *J. Chem. Inf. Comput. Sci.*, **43**, 248–254.
- Coulson, C.A. (1939) The electronic structure of some polyenes and aromatic molecules. VII. Bonds of fractional order by the molecular orbital method. *Proc. Roy. Soc. London A*, **169**, 413–428.
- Coulson, C.A. (1946) Bond fixation in compounds containing the carbonyl group. *Trans. Faraday Soc.*, **42**, 106–112.
- Coulson, C.A. (1960) Present state of molecular structure calculations. *Rev. Mod. Phys.*, **32**, 170.
- Craig, P.N. (1972) Structure–activity correlations of antimalarial compounds. 1. Free–Wilson analysis of 2-phenylquinoline-4-carbinols. *J. Med. Chem.*, **15**, 144–149.
- Craig, P.N. (1984) QSAR – origins and present status: a historical perspective. *Drug Inf. J.*, **18**, 123–130.
- Craig, N. (1990) Substructural analysis and compound selection, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 645–666.
- Craik, D.J. and Brownlee, R.T.C. (1983) Substituent effects on chemical shifts in the sidechains of aromatic systems. *Prog. Phys. Org. Chem.*, **14**, 1–73.
- Cramer, R.D. III (1980a) BC(DEF) parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Am. Chem. Soc.*, **102**, 1837–1849.
- Cramer, R.D. III (1980b) BC(DEF) parameters. 2. An empirical structure-based scheme for the prediction of some physical properties. *J. Am. Chem. Soc.*, **102**, 1849–1859.
- Cramer, R.D. III (1983a) BC(DEF) coordinates. 3. Their acquisition from physical property data. *Quant. Struct. -Act. Relat.*, **2**, 7–12.
- Cramer, R.D. III (1983b) BC(DEF) coordinates. 4. Correlations with general anesthesia, nerve blockade, and erythrocyte stabilization. *Quant. Struct. -Act. Relat.*, **2**, 13–19.
- Cramer, R.D. III (1993) Partial least squares (PLS): its strengths and limitations. *Persp. Drug Disc. Des.*, **1**, 269–278.
- Cramer, R.D., III, Bunce, J.D., Patterson, D.E. and Frank, I.E. (1988) Crossvalidation, bootstrapping and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct. -Act. Relat.*, **7**, 18–25.
- Cramer, R.D., III, Clark, R.D., Patterson, D.E. and Ferguson, A.M. (1996) Bioisosterism as a molecular diversity descriptor: steric fields of single “topomeric” conformers. *J. Med. Chem.*, **39**, 3060–3069.
- Cramer, R.D., III, DePriest, S.A., Patterson, D.E. and Hecht, P. (1993) The developing practice of comparative molecular field analysis, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 443–485.
- Cramer, R.D., III, Patterson, D.E. and Bunce, J.D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, **110**, 5959–5967.
- Cramer, R.D., III, Patterson, D.E., Clark, R.D., Soltanshahi, F. and Lawless, M.S. (1998) Visual compound libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.*, **38**, 1010–1023.
- Cramer, R.D., III, Redl, G. and Berkoff, C.E. (1974) Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.*, **17**, 533–535.
- Cramer, C.J. (1995) Continuum solvation models: classical and quantum mechanical implementations, in *Reviews in Computational Chemistry*, Vol. 6 (eds K.B. Lipkowitz and D.B. Boyd), VCH Publishers, New York, pp. 1–72.
- Cramer, C.J., Famini, G.R. and Lowrey, A.H. (1993) Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure–activity relationships. *Acc. Chem. Res.*, **26**, 599–605.

- Crebelli, R., Andreoli, C., Carere, A., Conti, G., Conti, L., Ramusino, M.C. and Benigni, R. (1992) The induction of mitotic chromosome malsegregation in *Aspergillus nidulans*. quantitative structure–activity relationship (QSAR) analysis with chlorinated aliphatic hydrocarbons. *Mut. Res.*, **266**, 117–134.
- Crebelli, R., Andreoli, C., Carere, A., Conti, L., Crochi, B., Cotta Ramusino, M. and Benigni, R. (1995) Toxicology of halogenated aliphatic hydrocarbons structural and molecular determinants for the disturbance of chromosome segregation and the induction of lipid peroxidation. *Chem. -Biol. Inter.*, **98**, 113–129.
- Cringean, J.K. and Lynch, M.F. (1989) Subgraphs of reduced chemical graphs as screen for substructure searching of specific chemical structures. *J. Inform. Sci.*, **15**, 211–222.
- Crippen, G.M. (1977) A novel approach to calculation of conformation. *J. Comput. Phys.*, **24**, 96–107.
- Crippen, G.M. (1978) Rapid calculation of coordinates from distance matrices. *J. Comput. Phys.*, **26**, 449.
- Crippen, G.M. (1979) Distance geometry approach to rationalizing binding data. *J. Med. Chem.*, **22**, 988–997.
- Crippen, G.M. (1980) Quantitative structure–activity relationships by distance geometry: systematic analysis of dihydrofolate reductase inhibitors. *J. Med. Chem.*, **23**, 599–606.
- Crippen, G.M. (1981) *Distance Geometry and Conformational Calculations*, Research Studies Press, Letchworth, UK, p. 58.
- Crippen, G.M. (1987) Voronoi binding site models. *J. Comput. Chem.*, **8**, 943–955.
- Crippen, G.M. (1991) Chemical distance geometry: current realization and future projection. *J. Math. Chem.*, **6**, 307–324.
- Cristante, M., Selves, J.L., Grassy, G. and Colin, J.P. (1993) Structure–activity relationship study on paraffin inhibitors for crude oils (INIPAR model II). *Anal. Chim. Acta*, **274**, 303–316.
- Crivori, P., Cruciani, G., Carrupt, P.-A. and Testa, B. (2000) Predicting blood–brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.*, **43**, 2204–2216.
- Crivori, P., Zamora, I., Speed, B., Orrenius, C. and Poggesi, I. (2004) Model based on GRID-derived descriptors for estimating CYP3A4 enzyme stability of potential drug candidates. *J. Comput. Aid. Mol. Des.*, **18**, 155–166.
- Croizet, F., Langlois, M.H., Dubost, J.P., Braquet, P., Audry, E., Dallet, Ph. and Colleter, J.C. (1990) Lipophilicity force field profile: an expressive visualization of the lipophilicity molecular potential gradient. *J. Mol. Graph.*, **8**, 153–155.
- Cronin, M.T.D. (1992) Molecular descriptors of QSAR, in *Quantitative Structure/Activity Relationships (QSAR) in Toxicology* (eds T. Coccini, L. Giannoni, W. Karcher, L. Manzo and R. Roi), Joint Research Centre – EEC, Brussels, Belgium, pp. 43–54.
- Cronin, M.T.D. (1996) Quantitative structure–activity relationship (QSAR): analysis of the acute sublethal neurotoxicity of solvents. *Toxicol. Vitro*, **10**, 103–110.
- Cronin, M.T.D., Aptula, A.O., Dearden, J.C., Duffy, J. C., Netzeva, T.I., Patel, H., Rowe, P.H., Schultz, T. W., Worth, A.P., Voutzoulidis, K. and Schüürmann, G. (2002a) Structure-based classification of antibacterial activity. *J. Chem. Inf. Comput. Sci.*, **42**, 869–878.
- Cronin, M.T.D., Aptula, A.O., Duffy, J.C., Netzeva, T. I., Rowe, P.H., Valkova, I.V. and Schultz, T.W. (2002b) Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. *Chemosphere*, **49**, 1201–1221.
- Cronin, M.T.D. and Dearden, J.C. (1995a) QSAR in toxicology. 1. Prediction of aquatic toxicity. *Quant. Struct. -Act. Relat.*, **14**, 1–7.
- Cronin, M.T.D. and Dearden, J.C. (1995b) QSAR in toxicology. 2. Prediction of acute mammalian toxicity and interspecies correlations. *Quant. Struct. -Act. Relat.*, **14**, 117–120.
- Cronin, M.T.D. and Dearden, J.C. (1995c) QSAR in toxicology. 3. Prediction of chronic toxicity. *Quant. Struct. -Act. Relat.*, **14**, 329–334.
- Cronin, M.T.D. and Dearden, J.C. (1995d) QSAR in toxicology. 4. Prediction of non-lethal mammalian toxicological end points, and expert systems for toxicity prediction. *Quant. Struct. -Act. Relat.*, **14**, 518–523.
- Cronin, M.T.D., Dearden, J.C. and Dobbs, A.J. (1991) QSAR studies of comparative toxicity in aquatic organisms. *Sci. Total Environ.*, **109/110**, 431–439.
- Cronin, M.T.D., Gregory, B.W. and Schultz, T.W. (1998) Quantitative structure–activity analyses of nitrobenzene toxicity to *Tetrahymena pyriformis*. *Chem. Res. Toxicol.*, **11**, 902–908.
- Cronin, M.T.D., Netzeva, T.I., Dearden, J.C., Edwards, R. and Worgan, A.D.P. (2004) Assessment and modeling of the toxicity of organic chemicals to *Chlorella vulgaris*: development of novel database. *Chem. Res. Toxicol.*, **17**, 545–554.
- Cronin, M.T.D. and Schultz, T.W. (1996) Structure–toxicity relationships for phenols to *Tetrahymena pyriformis*. *Chemosphere*, **32**, 1453–1468.

- Cronin, M.T.D. and Schultz, T.W. (2001) Development of quantitative structure–activity relationships for the toxicity of aromatic compounds to *Tetrahymena pyriformis*: comparative assessment of the methodologies. *Chem. Res. Toxicol.*, **14**, 1284–1295.
- Cronin, M.T.D. and Schultz, T.W. (2003) Pitfalls in QSAR. *J. Mol. Struct. (Theochem)*, **622**, 39–51.
- Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environ. Health Persp.*, **111**, 1376–1390.
- Cronin, M.T.D. and Worth, A.P. (2008) (Q)SARs for predicting effects relating to reproductive toxicity. *QSAR Comb. Sci.*, **27**, 91–100.
- Crowe, J.E., Lynch, M.F. and Town, W.G. (1970) Analysis of structural characteristics of chemical compounds in a large computer-based file. Part 1. Non-cyclic fragments. *J. Chem. Soc., C*, **23**, 990–997.
- Cruciani, G., Baroni, M., Carosati, E., Clementi, M., Valigi, R. and Clementi, S. (2004) Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *J. Chemom.*, **18**, 146–155.
- Cruciani, G., Baroni, M., Clementi, S., Costantino, G., Riganelli, D. and Skagerberg, B. (1992) Predictive ability of regression models. Part I. Standard deviation of prediction errors (SDEP). *J. Chemom.*, **6**, 335–346.
- Cruciani, G., Benedetti, P., Caltabiano, G., Condorelli, D.F., Fortuna, G.C. and Musumarra, G. (2004) Structure-based rationalization of antitumor drugs mechanism of action by a MIF approach. *Eur. J. Med. Chem.*, **39**, 281–289.
- Cruciani, G. and Clementi, S. (1994) GOLPE: philosophy and applications in 3D QSAR, in *Advanced Computer-Assisted Techniques in Drug Discovery* (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 61–88.
- Cruciani, G., Clementi, S. and Baroni, M. (1993) Variable selection in PLS analysis, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 551–564.
- Cruciani, G., Clementi, S. and Pastor, M. (1998) GOLPE-guided region selection, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 71–86.
- Cruciani, G., Crivori, P., Carrupt, P.-A. and Testa, B. (2000) Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *J. Mol. Struct. (Theochem)*, **503**, 17–30.
- Cruciani, G., Meniconi, M., Carosati, E., Zamora, I. and Mannhold, R. (2003) VolSurf: a tool for drug ADME-properties prediction, in *Drug Bioavailability*, Vol. 18, Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 406–419.
- Cruciani, G., Pastor, M., Benedetti, P. and Clementi, S. (2001a) From molecular interactions fields to a widely applicable set of descriptors, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 159–170.
- Cruciani, G., Pastor, M., Clementi, M. and Clementi, S. (2001b) GRIND (grid independent descriptors) in 3D structure–metabolism relationships, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 251–260.
- Cruciani, G., Pastor, M. and Clementi, S. (1997) Region selection in 3D-QSAR, in *Computer-Assisted Lead Finding and Optimization* (eds H. van de Waterbeemd, B. Testa and G. Folkers), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 381–395.
- Cruciani, G., Pastor, M. and Guba, W. (2000) VolSurf: a new tool for the pharmaceutical optimization of lead compounds. *Eur. J. Pharm. Sci.*, **11** (Suppl.), S29–S39.
- Cruciani, G., Pastor, M. and Mannhold, R. (2002) Suitability of molecular descriptors for database mining. A comparative analysis. *J. Med. Chem.*, **45**, 2685–2694.
- Cruciani, G. and Watson, K.A. (1994) Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.*, **37**, 2589–2601.
- Crum-Brown, A. (1864) On the theory of isomeric compounds. *Trans. Roy. Soc. Edinburgh*, **23**, 707–719.
- Crum-Brown, A. (1867) On an application of mathematics to chemistry. *Proceedings of the Royal Society, Edinburgh*, Vol. VI (No. 73), pp. 89–90.
- Crum-Brown, A. and Fraser, T.R. (1868) On the connection between chemical constitution and physiological action. Part 1. On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia. *Trans. Roy. Soc. Edinburgh*, **25**, 151–203.
- Cruz-Monteagudo, M., Borges, F., Pérez González, M. and Soeiro Cordeiro, N.D. (2007) Computational modeling tools for the design of potent antimalarial bisbenzamidines: overcoming

- the antimalarial potential of pentamidine. *Bioorg. Med. Chem.*, **15**, 5322–5339.
- Cruz-Monteagudo, M., González Díaz, H., Borges, F., Domínguez, E.R. and Cordeiro, M.N.D.S. (2008) 3D-MEDNEs: an alternative “*in silico*” technique for chemical research in toxicology. 2. Quantitative proteome–toxicity relationships (QPTR) based on mass spectrum spiral entropy. *Chem. Res. Toxicol.*, **21**, 619–632.
- Cserhádi, T., Forgács, E., Deyl, Z., Miksik, I. and Eckhardt, A. (2002) Modification of nonlinear mapping technique for quantitative structure–retention relationship studies. *Croat. Chem. Acta*, **75**, 13–24.
- Csorvássy, I. and Tötsér, L. (1991) Functions and metrics in molecular transform and their application, in *QSAR: Rational Approaches to the Design of Bioactive Compounds* (eds C. Silipo and A. Vittoria), Elsevier, Amsterdam, The Netherlands, pp. 193–196.
- Cuadras, C.M. (1989) Distancias Estadísticas. *Estadística Española*, **30**, 295–378.
- Cuissart, B., Touffet, F., Crémilleux, B., Bureau, R. and Rault, S. (2002) The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *J. Chem. Inf. Comput. Sci.*, **42**, 1043–1052.
- Cummins, D.J. and Andrews, C.W. (1995) Iteratively reweighted partial least squares: a performance analysis by Monte Carlo simulation. *J. Chemom.*, **9**, 489–507.
- Cummins, D.J., Andrews, C.W., Bentley, J.A. and Cory, M. (1996) Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.*, **36**, 750–763.
- Cvetković, D.M., Doob, M. and Sachs, H. (1995) *Spectra of Graphs. Theory and Applications*, Johann Ambrosius Barth Verlag, Heidelberg, Germany, p. 447.
- Cvetković, D.M. and Fowler, P.W. (1999) A group-theoretical bound for the number of main eigenvalues of a graph. *J. Chem. Inf. Comput. Sci.*, **39**, 638–641.
- Cvetković, D.M. and Gutman, I. (1977) Note on branching. *Croat. Chem. Acta*, **49**, 115–121.
- Cvetković, D.M. and Gutman, I. (1985) The computer system GRAPH: a useful tool in chemical graph theory. *J. Comput. Chem.*, **7**, 640–644.
- Cyranski, M. and Krygowski, T.M. (1996) Separation of the energetic and geometric contributions to aromaticity. 3. Analysis of the aromatic character of benzene rings in their various topological and chemical environments in the substituted benzene derivatives. *J. Chem. Inf. Comput. Sci.*, **36**, 1142–1145.
- Cyvin, B.N., Brunvoll, J., Cyvin, S.J. and Gutman, I. (1988) All-benzenoid systems: enumeration and classification of benzenoid hydrocarbons. *MATCH Commun. Math. Comput. Chem.*, **23**, 163–174.
- Czerminski, R., Yasri, A. and Hartsough, D. (2001) Use of support vector machine in pattern classification: application to QSAR studies. *Quant. Struct. -Act. Relat.*, **20**, 227–240.
- D'Archivio, A.A., Ruggieri, F., Mazzeo, P. and Tettamanti, E. (2007) Modelling of retention of pesticides in reversed-phase high-performance liquid chromatography: quantitative structure–retention relationships based on solute quantum-chemical descriptors and experimental (solvatochromic and spin-probe) mobile phase descriptors. *Anal. Chim. Acta*, **593**, 140–151.
- da Silva Junkes, B., Amboni, D., de, M.C., Heinzen, V. E.F. and Yunes, R.A. (2002) Quantitative structure–retention relationships (QSRR), using the optimum semi-empirical topological index, for methyl-branched alkanes produced by insects. *Chromatographia*, **55**, 707–713.
- da Silva Junkes, B., Amboni, D., de, M.C., Heinzen, V. E.F. and Yunes, R.A. (2007) Use of a semi-empirical topological method to predict the chromatographic retention of branched alkenes. *Chromatographia*, **55**, 75–80.
- da Silva Junkes, B., Amboni, D., de, M.C., Yunes, R.A. and Heinzen, V.E.F. (2003a) Prediction of the chromatographic retention of saturated alcohols on stationary phases of different polarity applying the novel semi-empirical topological index. *Anal. Chim. Acta*, **477**, 29–39.
- da Silva Junkes, B., Amboni, D., de, M.C., Yunes, R.A. and Heinzen, V.E.F. (2003b) Semiempirical topological index: a novel molecular descriptor for quantitative-structure–retention relationship studies. *Internet Electron. J. Mol. Des.*, **2**, 33–49.
- da Silva Junkes, B., Amboni, D., de, M.C., Yunes, R.A. and Heinzen, V.E.F. (2004) Application of the semi-empirical topological index in quantitative structure–chromatographic retention relationship (QSRR) studies of aliphatic ketones and aldehydes on stationary phases of different polarity. *J. Braz. Chem. Soc.*, **15**, 183–189.
- da Silva Junkes, B., Silva Arruda, A.C., Yunes, R.A., Cucco Porto, L. and Heinzen, V.E.F. (2005) Semi-empirical topological index: a tool for QSPR/QSAR studies. *J. Mol. Model.*, **11**, 128–134.
- Da, Y.-Z., Ito, K. and Fujiwara, H. (1992) Energy aspects of oil/water partition leading to the novel

- hydrophobic parameters for the analysis of quantitative structure–activity relationships. *J. Med. Chem.*, **35**, 3382–3387.
- Da, Y.-Z., Yanagi, J., Tanaka, K. and Fujiwara, H. (1993) Thermochemical aspects of partition quantitative structure–activity relationships of benzyldimethylalkylammonium chlorides. *Chem. Pharm. Bull.*, **41**, 227–230.
- Dai, J., Jin, L., Yao, S. and Wang, L.-S. (2001) Prediction of partition coefficient and toxicity for benzaldehyde compounds by their capacity factors and various molecular descriptors. *Chemosphere*, **42**, 899–907.
- Dai, Q., Liu, X. and Wang, T. (2006) A novel 2D graphical representation of DNA sequences and its application. *J. Mol. Graph. Model.*, **25**, 340–344.
- Damborsky, J. and Schultz, T.W. (1997) Comparison of the QSAR models for toxicity and biodegradability of anilines and phenols. *Chemosphere*, **34**, 429–446.
- Danauskas, S.M. and Jurs, P.C. (2001) Prediction of C60 solubilities from solvent molecular structures. *J. Chem. Inf. Comput. Sci.*, **41**, 419–424.
- Dancoff, S.M. and Quastler, H. (eds) (1953) *Essays on the Use of Information Theory in Biology*, University of Illinois, Urbana, IL.
- Dang, P. and Madan, A.K. (1994) Structure–activity study on anticonvulsant (thio)hydantoins using molecular connectivity indices. *J. Chem. Inf. Comput. Sci.*, **34**, 1162–1166.
- Dannenfelser, R.-M., Surendran, N. and Yalkowsky, S.H. (1993) Molecular symmetry and related properties. *SAR & QSAR Environ. Res.*, **1**, 273–292.
- Dannenfelser, R.-M. and Yalkowsky, S.H. (1996) Estimation of entropy of melting from molecular structure: a non-group contribution method. *Ind. Eng. Chem. Res.*, **35**, 1483–1486.
- Daren, Z. (2001) QSPR studies of PCBs by the combination of genetic algorithms and PLS analysis. *Computers Chem.*, **25**, 197–204.
- Decision Analysis by Ranking Techniques (DART), Ver. 2.0, Milano Chemometrics & QSAR Research Group, University of Milano-Bicocca, P.za. della Scienza 1, Milano, Italy.
- Das, A., Dömötör, G., Gutman, I., Joshi, S., Karmarkar, S., Khaddar, D., Khaddar, T., Khadikar, P.V., Popovic, L., Sapre, N.S., Sapre, N. and Shirhatti, A. (1997) A comparative study of the Wiener, Schultz and Szeged indices of cycloalkanes. *J. Serb. Chem. Soc.*, **62**, 261–239.
- Das, K.C. and Gutman, I. (2004) Some properties of the second Zagreb index. *MATCH Commun. Math. Comput. Chem.*, **52**, 103–112.
- Dash, S.C. and Behera, G.B. (1980) A new steric parameter to explain *ortho*-substituent effect. *Indian J. Chem.*, **19**, 541–543.
- Dauben, H.J., Jr, Wilson, J.D. and Laity, J.L. (1968) Diamagnetic susceptibility exaltation as a criterion of aromaticity. *J. Am. Chem. Soc.*, **90**, 811–813.
- Daudel, P. and Daudel, R. (1966) *Chemical Carcinogenesis and Molecular Biology*, Wiley–Interscience, New York.
- David, V. and Mihailciuc, C. (2006) Characterization of multi-signals analytical outcome by means of the information entropy and energy. *Rev. Roum. Chim.*, **51**, 317–322.
- Davis, A.M., Gensmantel, N.P., Johansson, E. and Marriott, D.P. (1994) The use of the grid program in the 3-D QSAR analysis of a series of calcium channel agonists. *J. Med. Chem.*, **37**, 963–972.
- Davis, L. (ed.) (1991) *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York.
- Daylight Theory Manual, Ver. 4.9, Daylight Chemical Information Systems, Inc., 18500 Von Karman, 450, Irvine, CA.
- De Benedetti, P.G. (1992) Electrostatics in quantitative structure–activity relationship analysis. *J. Mol. Struct. (Theochem)*, **256**, 231–248.
- De Benedetti, P.G., Menziani, M.C., Cocchi, M. and Fanelli, F. (1995) Prototropic molecular forms and theoretical descriptors in QSAR analysis. *J. Mol. Struct. (Theochem)*, **333**, 1–17.
- de Bruijn, J. and Hermens, J.L.M. (1990) Relationships between octanol/water partition coefficients and total molecular surface area and total molecular volume of hydrophobic organic chemicals. *Quant. Struct. -Act. Relat.*, **9**, 11–21.
- de Bruijn, J. and Hermens, J.L.M. (1993) Inhibition of acetylcholinesterase and acute toxicity of organophosphorous compounds to fish: a preliminary structure–activity analysis. *Aquat. Toxicol.*, **24**, 257–274.
- De Castro, L.F.P. and Reissmann, S. (1995) QSAR in bradykinin antagonists. Inhibition of the bradykinin induced contraction of the isolated rat uterus and guinea pig ileum. *Quant. Struct. -Act. Relat.*, **14**, 249–257.
- de Cerqueira Lima, P., Golbraikh, A., Oloff, S. Xiao, Y.-D. and Tropsha, A. (2006) Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.*, **46**, 1245–1254.
- de Gregorio, C., Kier, L.B. and Hall, L.H. (1998) QSAR modeling with the electrotopological state indices: corticosteroids. *J. Comput. Aid. Mol. Des.*, **12**, 557–561.
- De Julián-Ortiz, V., de Gregorio Alapont, C., Rios-Santamarina, I., García-Domenech, R. and Gálvez,

- J. (1998) Prediction of properties of chiral compounds by molecular topology. *J. Mol. Graph. Model.*, **16**, 14–18.
- De Julián-Ortiz, V., García-Domenech, R., Gálvez, J., Soler-Roca, R., García-March, F.J. and Antón-Fos, G.M. (1996) Use of topological descriptors in chromatographic chiral separations. *J. Chromat.*, **719**, 37–44.
- De La Guardia, M., Carrión, J.L. and Galdú, M.V. (1988) The use of topological models in analytical chemistry. *J. Chemom.*, **3**, 193–207.
- de Lima Ribeiro, F.A. and Castro Ferreira, M.M. (2005) QSAR model of the phototoxicity of polycyclic aromatic hydrocarbons. *J. Mol. Struct. (Theochem)*, **719**, 191–200.
- de Lima Ribeiro, F.A. and Ferreira, M.M.C. (2003) QSPR models of boiling point, octanol–water partition coefficient and retention time index of polycyclic aromatic hydrocarbons. *J. Mol. Struct. (Theochem)*, **663**, 109–126.
- De Maria, P., Fini, A. and Hall, F.M. (1973) Thermodynamic acid dissociation constants of aromatic thiols. *J. Chem. Soc. Perkin Trans. 2*, 1969–1971.
- De Rienzo, F., Grant, G.H. and Menziani, M.C. (2002) Theoretical descriptors for the quantitative rationalisation of plastocyanin mutant functional properties. *J. Comput. Aid. Mol. Des.*, **16**, 501–509.
- Dean, P.M. (1987) *Molecular Foundations of Drug–Receptor Interaction*, Cambridge University Press, Cambridge, UK, p. 381.
- Dean, P.M. (1990) Molecular recognition: the measurement and search for molecular similarity in ligand–receptor interaction, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 211–238.
- Dean, P. M. Molecular similarity, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi) (1993) ESCOM, Leiden, The Netherlands, pp. 150–172.
- Dean, P.M. (ed.) (1995) *Molecular Similarity in Drug Design*, Chapman & Hall, London, UK.
- Dean, P.M. and Callow, P. (1987) Molecular recognition: identification of local minima for matching in rotational 3-space by cluster analysis. *J. Mol. Graph.*, **5**, 159–164.
- Dean, P.M., Callow, P. and Chau, P.-L. (1988) Molecular recognition: blind-searching for regions of strong structural match on the surfaces of two dissimilar molecules. *J. Mol. Graph.*, **6**, 28–34.
- Dean, P.M. and Chau, P.-L. (1987) Molecular recognition: optimized searching through rotational 3-space for pattern matches on molecular surfaces. *J. Mol. Graph.*, **5**, 152–158.
- Dean, P.M. and Perkins, T.D.J. (1993) Searching for molecular similarity between flexible molecules, in *Trends in QSAR and Molecular Modelling 92* (ed. C. G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 207–215.
- Deanda, F. and Pearlman, R.S. (2002) A novel approach for identifying the surface atoms of macromolecules. *J. Mol. Graph. Model.*, **20**, 415–425.
- Dearden, J.C. (ed.) (1983) *Quantitative Approaches to Drug Design*, Elsevier, Amsterdam, The Netherlands.
- Dearden, J.C. (1985) Partitioning and lipophilicity in quantitative structure–activity relationships. *Environ. Health Persp.*, **61**, 203–228.
- Dearden, J.C. (1990) Physico-chemical descriptors, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 25–59.
- Dearden, J.C. (1991) The QSAR prediction of melting point, a property of environmental relevance. *Sci. Total Environ.*, **109/110**, 59–68.
- Dearden, J.C. (2002) Prediction of environmental toxicity and fate using quantitative structure–activity relationships (QSARs). *J. Braz. Chem. Soc.*, **13**, 754–762.
- Dearden, J.C. (2003a) *In silico* prediction of drug toxicity. *J. Comput. Aid. Mol. Des.*, **17**, 119–127.
- Dearden, J.C. (2003b) Quantitative structure–property relationships for predicting of boiling point, vapor pressure, and melting point. *Environ. Toxicol. Chem.*, **22**, 1696–1709.
- Dearden, J.C., Bradburne, S.J.A. and Abraham, M.H. (1991) The nature of molar refractivity, in *QSAR: Rational Approaches to the Design of Bioactive Compounds* (eds C. Silipo and A. Vittoria), Elsevier, Amsterdam, The Netherlands, pp. 143–150.
- Dearden, J.C., Cronin, M.T.D. and Dobbs, A.J. (1995a) Quantitative structure–activity relationships as a tool to assess the comparative toxicity of organic chemicals. *Chemosphere*, **31**, 2521–2528.
- Dearden, J.C., Cronin, M.T.D., Schultz, T.W. and Lin, D.T. (1995b) QSAR study of the toxicity of nitrobenzenes to *Tetrahymena pyriformis*. *Quant. Struct. -Act. Relat.*, **14**, 427–432.
- Dearden, J.C., Cronin, M.T.D. and Wee, D. (1997) Prediction of hydrogen bond donor ability using new quantum chemical parameters. *J. Pharm. Pharmacol.*, **49** (Suppl 4), 110.
- Dearden, J.C., Cronin, M.T.D., Zhao, Y.H. and Raevsky, O.A. (2000) QSAR studies of compounds acting by polar and non-polar narcosis: an examination of the

- role of polarisability and hydrogen bonding. *Quant. Struct. -Act. Relat.*, **19**, 3–9.
- Dearden, J.C. and Ghafourian, T. (1995) Investigation of calculated hydrogen bonding parameters for QSAR, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, Barcelona, Spain, pp. 117–119.
- Dearden, J.C. and Ghafourian, T. (1999) Hydrogen bonding parameters for QSAR: comparison of indicator variables, hydrogen bond counts, molecular orbital and other parameters. *J. Chem. Inf. Comput. Sci.*, **39**, 231–235.
- Dearden, J.C. and Schüürmann, G. (2003) Quantitative structure–property relationships for predicting Henry's law constant from molecular structure. *Environ. Toxicol. Chem.*, **22**, 1755–1770.
- Debnath, A.K. (1998) Comparative molecular field analysis (CoMFA) of a series of symmetrical bis-benzamide cyclic urea derivatives as HIV-1 protease inhibitors. *J. Chem. Inf. Comput. Sci.*, **38**, 761–767.
- Debnath, A.K., Compadre, R.L.L., Debnath, G., Shusterman, A.J. and Hansch, C. (1991) Structure–activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J. Med. Chem.*, **34**, 787–797.
- Debnath, A.K., Compadre, R.L.L. and Hansch, C. (1992a) Mutagenicity of quinolines in *Salmonella typhimurium* TA100. A QSAR study based on hydrophobicity and molecular orbital determinants. *Mut. Res.*, **280**, 55–65.
- Debnath, A.K., Compadre, R.L.L., Shusterman, A.J. and Hansch, C. (1992b) Quantitative structure–activity relationship investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test. 2. mutagenicity of aromatic and heteroaromatic nitro-compounds in *Salmonella typhimurium* TA100. *Envir. Mol. Mutag.*, **19**, 53–70.
- Debnath, A.K., Debnath, G., Shusterman, A.J. and Hansch, C. (1992) A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test. 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Envir. Mol. Mutag.*, **19**, 37–52.
- Debnath, A.K. and Hansch, C. (1992) Structure–activity relationship of genotoxic polycyclic aromatic nitro-compounds. Further evidence for the importance of hydrophobicity and molecular orbital energies in genetic toxicity. *Envir. Mol. Mutag.*, **20**, 140–144.
- Debnath, A.K. and Hansch, C. (1993) The importance of hydrophobicity in the mutagenicity of methanesulfonic acid esters with *Salmonella typhimurium* TA100. *Chem. Res. Toxicol.*, **6**, 310–312.
- Debnath, A.K., Hansch, C., Kim, K.H. and Martin, Y. C. (1993) Mechanistic interpretation of the genotoxicity of nitrofurans (antibacterial agents) using quantitative structure–activity relationships and comparative molecular field analysis. *J. Med. Chem.*, **36**, 1007–1016.
- Debnath, A.K., Jiang, S., Strick, N., Lin, K., Haberfield, P. and Neurath, A.R. (1994) Three dimensional structure–activity analysis of a series of porphyrin derivatives with anti HIV-1 activity targeted to the V3 loop of the gp120 envelope glycoprotein of the human immunodeficiency virus type 1. *J. Med. Chem.*, **37**, 1099–1108.
- Debnath, A.K., Shusterman, A.J., Compadre, R.L.L. and Hansch, C. (1994) The importance of the hydrophobic interaction in the mutagenicity of organic compounds. *Mut. Res.*, **305**, 63–72.
- Deconinck, E., Ates, H., Callebaut, N., van Gysegheem, E. and Vander Heyden, Y. (2007) Evaluation of chromatographic descriptors for the prediction of gastro-intestinal absorption of drugs. *J. Chromat.*, **1138**, 190–202.
- Deconinck, E., Hancock, T., Coomans, D., Massart, D.L. and Vander Heyden, Y. (2005) Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. *J. Pharm. Biomed. Anal.*, **39**, 91–103.
- Dehmer, M. (2008a) A novel method for measuring the structural information content of networks. *Cybernetics and Systems*, in press.
- Dehmer, M. (2008b) Information processing in complex networks: graph entropy and information functionals. *Appl. Math. Computat.*, **201**, 82–94.
- Dehmer, M. and Emmert-Streib, F. (2008) Structural information content of networks: graph entropy based on local vertex functionals. *Comp. Biol. Chem.*, **32**, 131–138.
- Deka, R.Ch., Roy, R.K. and Hirao, K. (2004) Local reactivity descriptors to predict the strength of Lewis acid sites in alkali cation-exchanged zeolites. *Chem. Phys. Lett.*, **389**, 186–190.
- Del, Re.G. (1958) A simple MO-LCAO method for the calculation of charge distribution in saturated organic molecules. *J. Chem. Soc.*, **40**, 4031–4040.
- Delaney, J.S., Mullaley, A., Mullier, G.W., Sexton, G.J., Taylor, R. and Viner, R.C. (1993) Rapid construction of data tables for quantitative structure–activity relationship studies. *J. Chem. Inf. Comput. Sci.*, **33**, 174–178.
- Delgado, E.J. and Alderete, J. (2002) On the calculation of Henry's law constants of chlorinated

- benzenes in water from semiempirical quantum chemical methods. *J. Chem. Inf. Comput. Sci.*, **42**, 559–563.
- Demeter, D.A., Weintraub, H.J.R. and Knittel, J.J. (1998) The local minima method (LMM) of pharmacophore determination: a protocol for predicting the bioactive conformation of small, conformationally flexible molecules. *J. Chem. Inf. Comput. Sci.*, **38**, 1125–1136.
- Demirev, P.A., Dylgerov, A.S. and Bangov, I.P. (1991) CTI: a novel charge-related topological index with low degeneracy. *J. Math. Chem.*, **8**, 367–382.
- Demuth, W., Karlovits, M. and Varmuza, K. (2004) Spectral similarity versus structural similarity: mass spectrometry. *Anal. Chim. Acta*, **516**, 75–85.
- Demyttenaere-Kovatcheva, A., Cronin, M.T.D., Benfenati, E., Roncaglioni, A. and LoLiparo, E. (2005) Identification of the structural requirements of the receptor-binding affinity of diphenolic azoles to estrogen receptors α and β by three-dimensional quantitative structure–activity relationship and structure–activity relationship analysis. *J. Med. Chem.*, **48**, 7628–7636.
- Denny, W.A., Cain, B.F., Atwell, C., Hansch, C., Panthanackal, A. and Leo, A. (1982) Potential antitumor agents. 36. Quantitative relationships between antitumor activity, toxicity and structure for the general class of 9-anilinoacridine antitumor agents. *J. Med. Chem.*, **25**, 276–315.
- Depczynski, U., Frost, V.J. and Molt, K. (2000) Genetic algorithms applied to the selection of factors in principal component regression. *Anal. Chim. Acta*, **420**, 217–227.
- Depczynski, U., Jetter, K., Molt, K. and Niemöller, A. (1997) The fast wavelet transform on compact intervals as a tool in chemometrics. I. Mathematical background. *Chemom. Intell. Lab. Syst.*, **39**, 19–27.
- Depczynski, U., Jetter, K., Molt, K. and Niemöller, A. (1999a) Quantitative analysis of near infrared spectra by wavelet coefficient regression using a genetic algorithm. *Chemom. Intell. Lab. Syst.*, **47**, 179–187.
- Depczynski, U., Jetter, K., Molt, K. and Niemöller, A. (1999b) The fast wavelet transform on compact intervals as a tool in chemometrics. II. Boundary effects, denoising and compression. *Chemom. Intell. Lab. Syst.*, **49**, 151–161.
- DePriest, S.A., Mayer, D., Naylor, C.B. and Marshall, G.R. (1993) 3D-QSAR of angiotensin converting enzyme and thermolysin inhibitors. A comparison of CoMFA models based on deduced and experimentally determined active site geometries. *J. Am. Chem. Soc.*, **115**, 5372–5384.
- Deretey, E., Feher, M. and Schmidt, J.M. (2002) Rapid prediction of human intestinal absorption. *Quant. Struct. -Act. Relat.*, **21**, 493–506.
- Desiraju, G.R., Gopalakrishnan, B., Jetty, R.K.R., Raveendra, D., Sarma, J.A.R.P. and Subramanya, H.S. (2000) Three-dimensional quantitative structural activity relationship (3D-QSAR) studies of some 1,5-diarylpyrazoles: analogue based design of selective cyclooxygenase-2 inhibitors. *Molecules*, **5**, 945–955.
- DeTar, D.F. and Delahunty, C. (1983) Ester aminolysis: new reaction series for the quantitative measurement of steric effects. *J. Am. Chem. Soc.*, **105**, 2734–2739.
- DeTar, D.F. and Tenpas, C.J. (1976) Theoretical calculation of steric effects in ester hydrolysis. *J. Am. Chem. Soc.*, **98**, 7903–7908.
- Devillers, J. (1996a) Genetic algorithms in computer-aided molecular design, in *Genetic Algorithms in Molecular Modeling. Principles of QSAR and Drug Design*, Vol. 1 (ed. J. Devillers), Academic Press, London, UK, pp. 131–157.
- Devillers, J. (ed.) (1996b) *Genetic Algorithms in Molecular Modeling. Principles of QSAR and Drug Design*, Vol. 1 Academic Press, London, UK, p. 327.
- Devillers, J. (ed.) (1998) *Comparative QSAR*, Taylor & Francis, Washington, DC, p. 371.
- Devillers, J. (1999a) Autocorrelation descriptors for modeling (eco)toxicological endpoints, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 595–612.
- Devillers, J. (1999b) No-free lunch molecular descriptors in QSAR and QSPR, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 1–20.
- Devillers, J. (2000) EVA/PLS versus autocorrelation/neural network estimation of partition coefficients. *Persp. Drug Disc. Des.*, **19**, 117–131.
- Devillers, J. and Balaban A.T. (eds) (1999) *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science Publishers, Amsterdam, The Netherlands, p. 824.
- Devillers, J., Chambon, P., Zakarya, D. and Chastrette, M. (1986) Quantitative structure–activity relations of the lethal effects of 38 halogenated compounds against *Lepomis macrochirus*. *Comp. Rend. Acad. Sci. (Paris, French)*, **303**, 613–616.
- Devillers, J., Domine, D., Bintein, S. and Karcher, W. (1998) Comparison of fish bioconcentration

- models, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 1–50.
- Devillers, J. and Karcher, W. (eds) (1991) *Applied Multivariate Analysis in SAR and Environmental Studies*, Kluwer Academic Publishers for the European Communities, Dordrecht, The Netherlands, p. 530.
- Devillers, J. and Lipnick, R.L. (1990) Practical applications of regression analysis in environmental QSAR studies, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 129–144.
- Dewar, M.J.S. (1969) *The Molecular Orbital Theory of Organic Chemistry*, McGraw-Hill, New York.
- Dewar, M.J.S. and de Llano, C. (1969) Ground states of conjugated molecules. XI. Improved treatment of hydrocarbons. *J. Am. Chem. Soc.*, **91**, 789–795.
- Dewar, M.J.S. and Gleicher, G.J. (1965) Ground states of conjugated molecules. III. Classical polyenes. *J. Am. Chem. Soc.*, **87**, 692–696.
- Dewar, M.J.S., Golden, R. and Harris, J.M. (1971) Substituent effects. X. An improved treatment (FMMF) of substituent effects. *J. Am. Chem. Soc.*, **93**, 4187–4195.
- Dewar, M.J.S. and Grisdale, P.J. (1962a) Substituent effects. I. Introduction. *J. Am. Chem. Soc.*, **84**, 3539–3541.
- Dewar, M.J.S. and Grisdale, P.J. (1962b) Substituent effects. IV. A quantitative theory. *J. Am. Chem. Soc.*, **84**, 3548–3553.
- Dewar, M.J.S. and Harget, A.J. (1970a) Ground states of conjugated molecules. XVI. Treatment of hydrocarbons by l.c.a.o. s.c.f. m.o. *Proc. Roy. Soc. London A*, **315**, 443–455.
- Dewar, M.J.S. and Harget, A.J. (1970b) Ground states of conjugated molecules. XVII. The l.c.a.o. s.c.f. m.o. treatment of compounds containing nitrogen and oxygen. *Proc. Roy. Soc. London A*, **315**, 457–464.
- Dewar, M.J.S., Harget, A.J. and Trinajstić, N. (1969) Ground states of conjugated molecules. XV. Bond localization and resonance energies in compounds containing nitrogen or oxygen. *J. Am. Chem. Soc.*, **91**, 6321–6325.
- Dewar, M.J.S., Haselbach, E. and Worley, S.D. (1970) Calculated and observed ionization potentials of unsaturated polycyclic hydrocarbons; calculated heats of formation by several semiempirical s.c.f. m.o. methods. *Proc. Roy. Soc. London A*, **315**, 431–442.
- Dewar, M.J.S., Kohn, M.C. and Trinajstić, N. (1971) Cyclobutadiene and diphenylcyclobutadiene. *J. Am. Chem. Soc.*, **93**, 3437–3440.
- Dewar, M.J.S. and Longuet-Higgins, H.C. (1952) The correspondence between the resonance and molecular orbital theories. *Proc. Roy. Soc. London A*, **214**, 482–493.
- Dewar, M.J.S. and Trinajstić, N. (1970) Resonance energies of some compounds containing nitrogen or oxygen. *Theor. Chim. Acta*, **17**, 235–238.
- Di Marzio, W., Galassi, S., Todeschini, R. and Consolaro, F. (2001) Traditional versus WHIM molecular descriptors in QSAR approaches applied to fish toxicity studies. *Chemosphere*, **44**, 401–406.
- Di Paolo, T. (1978a) Molecular connectivity in quantitative structure–activity relationship study of anesthetic and toxic activity of aliphatic hydrocarbons, ethers, and ketones. *J. Pharm. Sci.*, **67**, 566–568.
- Di Paolo, T. (1978b) Structure–activity relationships of anesthetic ethers using molecular connectivity. *J. Pharm. Sci.*, **67**, 564–565.
- Di Paolo, T., Kier, L.B. and Hall, L.H. (1977) Molecular connectivity and structure–activity relationship of general anesthetics. *Mol. Pharm.*, **13**, 31–37.
- Di Paolo, T., Kier, L.B. and Hall, L.H. (1979) Molecular connectivity study of halocarbon anesthetics. *J. Pharm. Sci.*, **68**, 39–42.
- Diaconis, P. and Efron, B. (1983) Computer intensive methods in statistics. *Sci. Am.*, **248**, 96–108.
- Dias, J.C., Rebelo, M.M. and Alves, C.N. (2004) A semi-empirical study of biflavonoid compounds with biological activity against tuberculosis. *J. Mol. Struct. (Theochem)*, **676**, 83–87.
- Dias, J.R. (1987a) A periodic table for polycyclic aromatic hydrocarbons. Part X. On the characteristic polynomial and other structural invariants. *J. Mol. Struct. (Theochem)*, **149**, 213–241.
- Dias, J.R. (1987b) Facile calculations of select eigenvalues and the characteristic polynomial of small molecular graphs containing heteroatoms. *Can. J. Chem.*, **65**, 734–739.
- Dias, J.R. (1992) Algebraic structure count. *J. Math. Chem.*, **9**, 253–260.
- Dias, J.R. (1993) *Molecular Orbital Calculations Using Chemical Graph Theory*, Springer-Verlag, Berlin, Germany.
- Dias, J.R. (1999) Directed toward the development of a unified structure theory of polycyclic conjugated hydrocarbons: the Aufbau principle in structure/similarity studies. *J. Chem. Inf. Comput. Sci.*, **39**, 197–203.
- Dietz, A. (1995) Yet another representation of molecular structure. *J. Chem. Inf. Comput. Sci.*, **35**, 787–802.

- Diller, D.J. and Merz, K.M., Jr (2002) Can we separate active from inactive conformations? *J. Comput. Aid. Mol. Des.*, **16**, 105–112.
- Dimitrov, S., Dimitrova, G., Pavlov, T., Dimitrova, N., Patlewicz, G., Niemela, J. and Mekenyan, O. (2005) A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.*, **45**, 839–849.
- Dimitrov, S.D. and Mekenyan, O. (1997) Dynamic QSAR: least squares fits with multiple predictors. *Chemom. Intell. Lab. Syst.*, **39**, 1–9.
- Dimitrov, S.D., Mekenyan, O.G., Sinks, G.D. and Schultz, T.W. (2003) Global modeling of narcotic chemicals: ciliate and fish toxicity. *J. Mol. Struct. (Theochem)*, **622**, 63–70.
- Dimoglo, A.S. (1985) Compositional approach to electronic structure description of chemical compounds, oriented on computer analysis of structure–activity relationships. *Khim-Farm. Zh.*, **4**, 438–444.
- Dimoglo, A.S., Bersuker, I.B. and Gorbachov, M.Y. (1988) Electron-topological study of SAR of various inhibitors of α -chymotrypsin. *Khim-Farm. Zh.*, **22**, 1355–1361.
- Dimoglo, A.S., Shvets, N.M., Tetko, I.V. and Livingstone, D.J. (2001) Electronic-topological investigation of the structure–acetylcholinesterase inhibitor activity relationship in the series of *N*-benzylpiperidine derivatives. *Quant. Struct. -Act. Relat.*, **20**, 31–45.
- Dimov, D., Nedyalkova, Z., Haladjova, S., Schüürmann, G. and Mekenyan, O. (2001) QSAR modeling of antimycobacterial activity and activity against other bacteria of 3-formyl rifamycin SV derivatives. *Quant. Struct. -Act. Relat.*, **20**, 298–318.
- Dimov, N. and Osman, A. (1996) Selection of molecular descriptors used in quantitative structure–gas chromatographic retention relationships. 2. Isoalkanes and alkenes. *Anal. Chim. Acta*, **323**, 15–25.
- Dimov, N., Osman, A., Mekenyan, O. and Papazova, D. (1994) Selection of molecular descriptors used in quantitative structure gas chromatographic retention relationships. 1. Application to alkylbenzenes and naphthalenes. *Anal. Chim. Acta*, **298**, 303–317.
- Dimov, N. and Papazova, D. (1978) Calculation of the retention indices of C_5 – C_9 cycloalkanes on squalene. *J. Chromat.*, **148**, 11–15.
- Dimov, N. and Papazova, D. (1979) Correlation equations for prediction of gas chromatographic separation of hydrocarbons on squalene. *Chromatographia*, **12**, 720.
- Dimroth, K., Reichardt, C., Siepmann, T. and Bohlman, F. (1963) Über Pyridinium-*N*-Phenol-betaine und ihre Verwendung zur Charakterisierung der Polarität von Lösungsmitteln. *Liebigs Ann. Chem.*, **661**, 1–37.
- Diudea, M.V. (1994) Molecular topology. 16. Layer matrices in molecular graphs. *J. Chem. Inf. Comput. Sci.*, **34**, 1064–1071.
- Diudea, M.V. (1995a) Molecular topology. 23. Novel Schultz analogue indices. *MATCH Commun. Math. Comput. Chem.*, **32**, 85–103.
- Diudea, M.V. (1995b) Molecular topology. 21. Wiener index of dendrimers. *MATCH Commun. Math. Comput. Chem.*, **32**, 71–83.
- Diudea, M.V. (1996a) Walk numbers ^eW_M : Wiener-type numbers of higher rank. *J. Chem. Inf. Comput. Sci.*, **36**, 535–540.
- Diudea, M.V. (1996b) Wiener and hyper-Wiener numbers in a single matrix. *J. Chem. Inf. Comput. Sci.*, **36**, 833–836.
- Diudea, M.V. (1997a) Cluj matrix invariants. *J. Chem. Inf. Comput. Sci.*, **37**, 300–305.
- Diudea, M.V. (1997b) Cluj matrix, CJ_U : source of various graph descriptors. *MATCH Commun. Math. Comput. Chem.*, **35**, 169–183.
- Diudea, M.V. (1997c) Indices of reciprocal properties or Harary indices. *J. Chem. Inf. Comput. Sci.*, **37**, 292–299.
- Diudea, M.V. (1997d) Unsymmetric matrix CJ_U : source of various graph invariants. *MATCH Commun. Math. Comput. Chem.*, **35**, 169–183.
- Diudea, M.V. (1999) Valencies of property. *Croat. Chem. Acta*, **72**, 835–851.
- Diudea, M.V. (ed.) (2001) *QSPR/QSAR Studies by Molecular Descriptors*, Nova Science Publishers, Huntington, NY, p. 438.
- Diudea, M.V. (2002a) Cluj polynomials. *Studia Univ. Babes-Bolyai*, **47**, 131–139.
- Diudea, M.V. (2002b) Hosoya polynomial in tori. *MATCH Commun. Math. Comput. Chem.*, **45**, 109–122.
- Diudea, M.V. (2006) Omega polynomial. *Carpathian J. Math.*, **22**, 43–47.
- Diudea, M.V. and Bal, L. (1990) Recursive relationships for computing Y indices in some particular graphs. *Studia Univ. Babes-Bolyai*, **35**, 17–28.
- Diudea, M.V., Cigher, S. and John, P.E. (2008) Omega and related counting polynomials. *MATCH Commun. Math. Comput. Chem.*, **60**, 237–250.
- Diudea, M.V. and Gutman, I. (1998) Wiener-type topological indices. *Croat. Chem. Acta*, **71**, 21–51.
- Diudea, M.V., Gutman, I. and Jäntschi, L. (2001) *Molecular Topology*, Nova Science Publishers, Huntington, NY, p. 332.
- Diudea, M.V., Horvath, D. and Bonchev, D. (1995a) Molecular topology. 14. MOLORD algorithm and

- real number subgraph invariants. *Croat. Chem. Acta*, **68**, 131–148.
- Diudea, M.V., Horvath, D. and Graovac, A. (1995b) Molecular topology. 15. 3D distance matrices and related topological indices. *J. Chem. Inf. Comput. Sci.*, **35**, 129–135.
- Diudea, M.V., Horvath, D., Kacso, I.E., Minailiuc, O.M. and Pârv, B. (1992) Molecular topology. VIII. Centricities in molecular graphs. The MOLCEN algorithm. *J. Math. Chem.*, **11**, 259–270.
- Diudea, M.V., Ivanciuc, O., Nikolić, S. and Trinajstić, N. (1997) Matrices of reciprocal distance. Polynomials and derived numbers. *MATCH Commun. Math. Comput. Chem.*, **35**, 41–64.
- Diudea, M.V., Jäntschi, L. and Pejov, L. (2002) Topological substituent descriptors. *Leonardo Electron. J. Pract. Technol.*, **1**, 1–18.
- Diudea, M.V. and Kacso, I.E. (1991) Composition rules for some topological indices. *MATCH Commun. Math. Comput. Chem.*, **26**, 255–269.
- Diudea, M.V., Kacso, I.E. and Minailiuc, O.M. (1992) Y indices in homogeneous dendrimers. *MATCH Commun. Math. Comput. Chem.*, **28**, 61–99.
- Diudea, M.V., Kacso, I.E. and Topan, M.I. (1996) Molecular topology. 18. A QSPR/QSAR study by using new valence group carbon-related electronegativities. *Rev. Roum. Chim.*, **41**, 141–157.
- Diudea, M.V., Katona, G., Lukovits, I. and Trinajstić, N. (1998) Detour–Cluj versus Detour indices. *Croat. Chem. Acta*, **71**, 459–471.
- Diudea, M.V., Katona, G., Minailiuc, O.M. and Pârv, B. (1995) Molecular topology 24. Wiener and hyper-Wiener indices in spiro-graphs. *Russ. Chem. Bull.*, **44**, 1606–1611.
- Diudea, M.V., Katona, G. and Pârv, B. (1997) Delta number, D_Δ , of dendrimers. *Croat. Chem. Acta*, **70**, 509–517.
- Diudea, M.V., Kiss, A.A., Estrada, E. and Guevara, N. (2000) Connectivity-, Wiener- and Harary-type indices of dendrimers. *Croat. Chem. Acta*, **73**, 367–381.
- Diudea, M.V., Minailiuc, O.M. and Balaban, A.T. (1991) Molecular topology. IV. Regressive vertex degrees (new graph invariants) and derived topological indices. *J. Comput. Chem.*, **12**, 527–535.
- Diudea, M.V., Minailiuc, O.M. and Katona, G. (1996) Molecular topology. 22. Novel connectivity descriptors based on walk degrees. *Croat. Chem. Acta*, **69**, 857–871.
- Diudea, M.V., Minailiuc, O.M. and Katona, G. (1997a) Molecular topology. 26. SP indices: novel connectivity descriptors. *Rev. Roum. Chim.*, **42**, 239–249.
- Diudea, M.V., Minailiuc, O.M., Katona, G. and Gutman, I. (1997b) Szeged matrices and related numbers. *MATCH Commun. Math. Comput. Chem.*, **35**, 129–143.
- Diudea, M.V. and Pârv, B. (1988) Molecular topology. 3. A new centric connectivity index (CCI). *MATCH Commun. Math. Comput. Chem.*, **23**, 65–87.
- Diudea, M.V. and Pârv, B. (1995) Molecular topology. 25. Hyper-Wiener index of dendrimers. *J. Chem. Inf. Comput. Sci.*, **35**, 1015–1018.
- Diudea, M.V., Pârv, B. and Gutman, I. (1997a) Detour–Cluj matrix and derived invariants. *J. Chem. Inf. Comput. Sci.*, **37**, 1101–1108.
- Diudea, M.V., Pârv, B. and Topan, M. (1997b) Derived Szeged and Cluj indices. *J. Serb. Chem. Soc.*, **62**, 267–276.
- Diudea, M.V. and Pop, C.M. (1996) Molecular topology. 27. A Schultz-type index based on the Wiener matrix. *Indian J. Chem.*, **35**, 257–261.
- Diudea, M.V. and Randić, M. (1997) Matrix operator, $W(M_1, M_2, M_3)$ and Schultz-type indices. *J. Chem. Inf. Comput. Sci.*, **37**, 1095–1100.
- Diudea, M.V. and Silaghi-Dumitrescu, I. (1989a) Molecular topology. I. Valence group electronegativity as a vertex discriminator. *Rev. Roum. Chim.*, **34**, 1175–1182.
- Diudea, M.V. and Silaghi-Dumitrescu, I. (1989b) Valence group electronegativity as a vertex discriminator. *Rev. Roum. Chim.*, **34**, 1175–1182.
- Diudea, M.V., Topan, M. and Graovac, A. (1994) Molecular topology. 17. Layer matrices of walk degrees. *J. Chem. Inf. Comput. Sci.*, **34**, 1072–1078.
- Diudea, M.V. and Ursu, O. (2003) Layer matrices and distance property descriptors. *Indian J. Chem.*, **42A**, 1283–1294.
- Diudea, M.V., Vizitiu, A.E. and Janežič, D. (2007) Cluj and related polynomials applied in correlating studies. *J. Chem. Inf. Model.*, **47**, 864–874.
- Dixon, J.S. and Villar, H.O. (1998) Bioactive diversity and screening library selection via affinity fingerprinting. *J. Chem. Inf. Comput. Sci.*, **38**, 1192–1203.
- Dixon, S.L. and Jurs, P.C. (1992) Atomic charge calculations for quantitative structure–property relationships. *J. Comput. Chem.*, **13**, 492–504.
- Dixon, S.L. and Jurs, P.C. (1993) Estimation of pK_a for organic oxyacids using calculated atomic charges. *J. Comput. Chem.*, **14**, 1460–1467.
- Dixon, S.L. and Koehler, R.T. (1999) The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *J. Med. Chem.*, **42**, 2887–2900.
- Dixon, S.L. and Villar, H.O. (1999) Investigation of classification methods for the prediction of activity

- in diverse chemical libraries. *J. Comput. Aid. Mol. Des.*, **13**, 533–545.
- Dobeš, P., Kmunicec, J., Mikeš, V. and Damborsky, J. (2004) Binding of fatty acids to β -cryptogein: quantitative structure–activity relationships and design of selective protein mutants. *J. Chem. Inf. Comput. Sci.*, **44**, 2126–2132.
- Dobrynin, A.A. (1993) Degeneracy of some matrix graph invariants. *J. Math. Chem.*, **14**, 175–184.
- Dobrynin, A.A. (1995) Solving a problem connected with distances in graphs. *Graph Theory Notes, New York*, **28**, 21–23.
- Dobrynin, A.A. (1997a) A new formula for the calculation of the Wiener index of hexagonal chains. *MATCH Commun. Math. Comput. Chem.*, **35**, 75–90.
- Dobrynin, A.A. (1997b) Congruence relations for the Wiener index of hexagonal chains. *J. Chem. Inf. Comput. Sci.*, **37**, 1109–1110.
- Dobrynin, A.A. (1998a) Formula for calculating the Wiener index of catacondensed benzenoid graphs. *J. Chem. Inf. Comput. Sci.*, **38**, 811–814.
- Dobrynin, A.A. (1998b) New congruence relations for the Wiener index of cata-condensed benzenoid graphs. *J. Chem. Inf. Comput. Sci.*, **38**, 405–409.
- Dobrynin, A.A. (1999a) A simple formula for the calculation of the Wiener index of hexagonal chains. *Computers Chem.*, **23**, 43–48.
- Dobrynin, A.A. (1999b) Explicit relation between the Wiener index and the Schultz index of catacondensed benzenoid graphs. *Croat. Chem. Acta*, **72**, 869–874.
- Dobrynin, A.A. (2003) On the Wiener index decomposition for catacondensed benzenoid graphs. *Indian J. Chem.*, **42**, 1270–1271.
- Dobrynin, A.A. and Gutman, I. (1994) On a graph invariant related to the sum of all distances in a graph. *Publ. Inst. Math. (Beograd)*, **56**, 18–22.
- Dobrynin, A.A. and Gutman, I. (1996) On the Szeged index of unbranched catacondensed benzenoid molecules. *Croat. Chem. Acta*, **69**, 845–856.
- Dobrynin, A.A. and Gutman, I. (1999) The average Wiener index of trees and chemical trees. *J. Chem. Inf. Comput. Sci.*, **39**, 679–683.
- Dobrynin, A.A., Gutman, I. and Dömötör, G. (1995) A Wiener-type graph invariant for some bipartite graphs. *Appl. Math. Lett.*, **8**, 57–62.
- Dobrynin, A.A., Gutman, I. and Piottukh-Peletskaia, V.N. (1999) Hyper-Wiener index of acyclic structures. *J. Struct. Chem.*, **40**, 293–298.
- Dobrynin, A.A. and Kochetova, A.A. (1994) Degree distance of a graph: a degree analogue of the Wiener index. *J. Chem. Inf. Comput. Sci.*, **34**, 1082–1086.
- Dobrynin, A.A. and Mel'nikov, L.S. (2001) Path layer matrix for weighted graphs. *MATCH Commun. Math. Comput. Chem.*, **44**, 135–146.
- Dobrynin, A.A. and Mel'nikov, L.S. (2004) Trees, quadratic line graphs and the Wiener index. *Croat. Chem. Acta*, **77**, 477–480.
- Dobrynin, A.A. and Mel'nikov, L.S. (2005a) Wiener index for graphs and their line graphs with arbitrary large cyclomatic numbers. *Appl. Math. Lett.*, **18**, 307–312.
- Dobrynin, A.A. and Mel'nikov, L.S. (2005b) Wiener index, line graphs and the cyclomatic number. *MATCH Commun. Math. Comput. Chem.*, **53**, 209–214.
- Doherty, P.J., Hoes, R.M., Robbat, A., Jr and White, C.M. (1984) Relationship between gas chromatographic retention indices and molecular connectivities of nitrated polycyclic aromatic hydrocarbons. *Anal. Chem.*, **56**, 2697–2701.
- Doichinova, I.A., Natcheva, R.N. and Mihailova, D.N. (1994) QSAR studies of 8-substituted xanthines as adenosine receptor antagonists. *Eur. J. Med. Chem.*, **29**, 133–138.
- Domine, D., Devillers, J., Wienke, D. and Buydens, L. (1996) Test series selection from nonlinear neural mapping. *Quant. Struct. -Act. Relat.*, **15**, 395–402.
- Donovan, W.H. and Famini, G.R. (1996) Using theoretical descriptions in structure–activity relationships: retention indices of sulfur vesicants and related compounds. *J. Chem. Soc. Perkin Trans. 2*, 83–89.
- Dorransoro, I., Chana, A., Abasolo, M.I., Castro, A., Gil, C., Stud, M. and Martinez, A. (2004) CODES/ neural network model: a useful tool for *in silico* prediction of oral absorption and blood–brain barrier permeability of structurally diverse drugs. *QSAR Comb. Sci.*, **23**, 89–98.
- Dosmorov, S.V. (1982) Generation of homogeneous reaction mechanism. *Kinetics and Catalysis*, (in russian).
- Douali, L., Villemin, D., Zyad, A. and Cherqaoui, D. (2004) Artificial neural networks: non-linear QSAR studies of HEPT derivatives as HIV-1 reverse transcriptase inhibitors. *Mol. Div.*, **8**, 1–8.
- Doucet, J.P. and Panaye, A. (1998) 3D structural information: from property prediction to substructure recognition with neural networks. *SAR & QSAR Environ. Res.*, **8**, 249–272.
- Doucet, J.P., Panaye, A. and Dubois, J.-E. (1983) Topological correlations of carbon-13 chemical shifts by perturbation on a focus: DARC-PULFO method. Attenuation and inversion of α -methyl substituent effects. *J. Org. Chem.*, **48**, 3174–3182.
- Dowdy, D.L., McKone, T.E. and Hsieh, D.P. (1996) Prediction of chemical biotransfer of organic

- chemicals from cattle diet into beef and milk using the molecular connectivity index. *Environ. Sci. Technol.*, **30**, 984–989.
- Doweyko, A.M. (1991) The hypothetical active site lattice – *in vitro* and *in vivo* explorations using a three-dimensional QSAR technique. *J. Math. Chem.*, **7**, 273–285.
- Doweyko, A.M. (2004) 3D-QSAR illusions. *J. Comput. Aid. Mol. Des.*, **18**, 57–596.
- Doweyko, A.M. and Mattes, W.B. (1992) An application of 3D QSAR to the analysis of the sequence specificity of DNA alkylation by uracil mustard. *Biochemistry*, **31**, 9388–9392.
- Downs, G.M. (2003) Ring perception, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 161–177.
- Downs, G.M., Gillet, V.J., Holliday, J.D. and Lynch, M. F. (1989a) Review of ring perception algorithms for chemical graphs. *J. Chem. Inf. Comput. Sci.*, **29**, 172–187.
- Downs, G.M., Gillet, V.J., Holliday, J.D. and Lynch, M. F. (1989b) Theoretical aspects of ring perception and development of the extended set of smallest rings concept. *J. Chem. Inf. Comput. Sci.*, **29**, 187–206.
- Downs, G.M. and Willett, P. (1999) Similarity searching in databases of chemical structures. *Rev. Comput. Chem.*, **7**, 1–66.
- Doyle, J.K. and Garver, J.E. (1977) Mean distance in a graph. *Disc. Math.*, **17**, 147–154.
- Doytchinova, I., Valkova, I. and Natcheva, R. (2001) CoMFA study on adenosine A2A receptor agonists. *Quant. Struct. -Act. Relat.*, **20**, 124–129.
- Doytchinova, I.A. and Flower, D.R. (2002) A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. *J. Comput. Aid. Mol. Des.*, **16**, 535–544.
- Doytchinova, I.A., Walshe, V., Borrow, P. and Flower, D.R. (2005) Towards the chemometric dissection of peptide-HLA-A*0201 binding affinity: comparison of local and global QSAR models. *J. Comput. Aid. Mol. Des.*, **19**, 203–212.
- Draber, W. (1996) Sterimol and its role in drug research. *Z. Naturforsch.*, **51c**, 1–7.
- DRAGON (Software for molecular descriptor calculations), Ver. 5.5, Taletè s.r.l., via V. Pisani, 13, 20124 Milano, Italy, <http://www.taletè.mi.it/dragon.htm>.
- Drakulić, B.J., Juranić, Z.D., Stanojković, T.P. and Juranić, I.O. (2005) 2-[(Carboxymethyl)sulfanyl]-4-oxo-4-arylbutanoic acids selectively suppressed proliferation of neoplastic human HeLa cells. A SAR/QSAR study. *J. Med. Chem.*, **48**, 5600–5603.
- Draper, N. and Smith, H. (1998) *Applied Regression Analysis*, John Wiley & Sons, Inc., New York, pp. 706.
- Drefahl, A. and Reinhard, M. (1993) Similarity-based search and evaluation of environmentally relevant properties for organic compounds in combination with the group contribution approach. *J. Chem. Inf. Comput. Sci.*, **33**, 886–895.
- Drmanić, S.Ž., Jovanović, B.Ž. and Mišić-Vuković, M.M. (2000) A comparative LFER study of the reactivity of pyridineacetic, pyridineacetic acids N-oxide and substituted phenylacetic acids with diazodiphenylmethane in various alcohols. *J. Serb. Chem. Soc.*, **62**, 847–856.
- Dross, K., Rekker, R.F., de Vries, G. and Mannhold, R. (1998) The lipophilic behaviour of organic compounds. 3. The search for interconnections between reversed/phase chromatographic data and log P_{oct} values. *Quant. Struct. -Act. Relat.*, **17**, 549–557.
- Du, Q., Artica, G.A. and Mezey, P.G. (1997) Heuristic lipophilicity potential for computer-aided rational drug design. *J. Comput. Aid. Mol. Des.*, **11**, 503–515.
- Du, Y. and Liang, Y.-Z. (2003) Data mining for seeking accurate quantitative relationship between molecular structure and GC retention indices of alkanes by projection pursuit. *Comp. Biol. Chem.*, **27**, 339–353.
- Du, Y., Liang, Y.-Z., Li, B. and Xu, C. (2005) Orthogonalization of block variables by subspace-projection for quantitative structure–property relationship (QSPR) research. *J. Chem. Inf. Comput. Sci.*, **42**, 993–1003.
- Du, Y., Liang, Y.-Z. and Yun, D. (2002) Data mining for seeking an accurate quantitative relationship between molecular structure and GC retention indices of alkenes by projection pursuit. *J. Chem. Inf. Comput. Sci.*, **42**, 1283–1292.
- Duan, X.-M., Li, Z.-H., Hu, H.-R., Song, G.-L., Wang, W.-N., Chen, G.-H. and Fan, K.-N. (2005a) Linear regression correction to first principle theoretical calculations – improved descriptors and enlarged training set. *Chem. Phys. Lett.*, **409**, 315–321.
- Duan, X.-M., Li, Z.-H., Song, G.-L., Wang, W.-N., Chen, G.-H. and Fan, K.-N. (2005b) Neural network correction for heats of formation with a larger experimental training set and new descriptors. *Chem. Phys. Lett.*, **410**, 125–130.
- Duart, M.J., García-Domenech, R., Antón-Fos, G.M. and Gálvez, J. (2001) Optimization of a mathematical topological pattern for the prediction of antihistaminic activity. *J. Comput. Aid. Mol. Des.*, **15**, 561–572.

- Duboc, C. (1978) The correlation analysis of nucleophilicity, in *Correlation Analysis in Chemistry* (eds N.B. Chapman and J. Shorter), Plenum Press, New York, pp. 313–355.
- Dubois, J.-E. (1976) Ordered chromatic graph and limited environment concept, in *Chemical Applications of Graph Theory* (ed. A.T. Balaban), Academic Press, New York, pp. 333–370.
- Dubois, J.-E., Chrétien, J.R., Soják, L. and Rijks, J.A. (1980) Topological analysis of the behaviour of linear alkanes up to tetradecenes in gas-liquid chromatography on squalene. *J. Chromat.*, **194**, 121–134.
- Dubois, J.-E., Doucet, J.P., Panaye, A. and Fan, B.T. (1999) DARC site topological correlations: ordered structural descriptors and property evaluation, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 613–673.
- Dubois, J.-E., Laurent, D. and Aranda, A. (1973a) DARC system. XVI. Theory of topology-information. I. Method for the perturbation of ordered, concentric, limited environments (PELCO). *J. Chim. Phys. Phys-Chim. Biol.*, **70**, 1608–1615.
- Dubois, J.-E., Laurent, D. and Aranda, A. (1973b) DARC system. XVII. Theory of topology-information. II. PELCO [perturbation of ordered, concentric, limited environments] method. Procedure for the establishment of the correlation of topology-information. *J. Chim. Phys. Phys-Chim. Biol.*, **70**, 1616–1624.
- Dubois, J.-E., Laurent, D., Bost, P., Chambaud, S. and Mercier, C. (1976) Système DARC. Méthode DARC/PELCO. Stratégies de Recherche de Corrélations Appliquées à une Population d'Adamantanamines Antigrippales. *Eur. J. Med. Chem.*, **11**, 225–236.
- Dubois, J.-E., Laurent, D., Panaye, A. and Sobel, Y. (1975a) Hyperstructures formelles d'antériorité. *Comp. Rend. Acad. Sci. (Paris, French)*, **281**, 687–690.
- Dubois, J.-E., Laurent, D., Panaye, A. and Sobel, Y. (1975b) Système DARC: concept d'hyperstructure formelle. *Comp. Rend. Acad. Sci. (Paris, French)*, **280**, 851–854.
- Dubois, J.-E., Laurent, D. and Viellard, H. (1966) Système de documentation et d'automatization des recherches des corrélations (DARC). Principes généraux. *Comp. Rend. Acad. Sci. (Paris, French)*, **263**, 764–767.
- Dubois, J.-E., Laurent, D. and Viellard, H. (1967) Système DARC. Principes des recherches de corrélations et équations générale de topo-information. *Comp. Rend. Acad. Sci. (Paris, French)*, **264**, 1019–1022.
- Dubois, J.-E. and Loukianoff, M. (1993) DARC "logic method" for molal volume prediction. *SAR & QSAR Environ. Res.*, **1**, 63–75.
- Dubois, J.-E., Loukianoff, M. and Mercier, C. (1992) Topology and the quest for structural knowledge. *J. Chim. Phys.*, **89**, 1493–1506.
- Dubois, J.-E., MacPhee, J.A. and Panaye, A. (1980) Steric effects. III. Composition of the E_s parameter. Variation of alkyl steric effects with substitution. Role of conformation in determining sterically active and inactive sites. *Tetrahedron*, **36**, 919–928.
- Dubois, J.-E., Mercier, C. and Panaye, A. (1986) DARC topological system and computer aided design. *Acta Pharm. Jugosl.*, **36**, 135–169.
- Dubois, J.-E., Mercier, C. and Sobel, Y. (1979) Théorie des graphes chimiques: méthode Darc/Pelco. Préférence des corrélations de topologie-information et analyse de fiabilité. *Comp. Rend. Acad. Sci. (Paris, French)*, **289**, 89–92.
- Dubois, J.-E., Panaye, A. and Attias, R. (1987) DARC system: notions of defined and generic substructures. Filiation and coding of FREL substructure (SS) classes. *J. Chem. Inf. Comput. Sci.*, **27**, 74–82.
- Dubois, J.-E., Sicouri, G., Sobel, Y. and Picchiottino, R. (1984) Système DARC: Opérateurs localisés et co-structures de l'invariant d'une réaction. *Comp. Rend. Acad. Sci. (Paris, French)*, **298** 525–530.
- Dubois, J.-E. and Sobel, Y. (1985) DARC system for documentation and artificial intelligence in chemistry. *J. Chem. Inf. Comput. Sci.*, **25**, 326–335.
- Dubost, J.P. (1993) 2D and 3D lipophilicity parameters in QSAR, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 93–100.
- Dubus, E., Ijjaali, I., Petitot, F. and Michel, A. (2006) *In silico* classification of hERG channel blockers: a knowledge-based strategy. *ChemMedChem*, **1**, 622–630.
- Duca, J.S. and Hopfinger, A.J. (2001) Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. *J. Chem. Inf. Comput. Sci.*, **41**, 1367–1387.
- Duchowicz, P.R. and Castro, E.A. (2000) A rather simple method to calculate log P values in QSAR/QSPR studies. *Acta Chim. Sloven.*, **47**, 281–292.
- Duchowicz, P.R., Castro, E.A. and Fernández, F.M. (2006) Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR theories. *MATCH Commun. Math. Comput. Chem.*, **55**, 179–192.

- Duchowicz, P.R., Castro, E.A., Fernández, F.M. and Pérez González, M. (2005) A new search algorithm for QSPR/QSAR theories: normal boiling points of some organic molecules. *Chem. Phys. Lett.*, **412**, 376–380.
- Duchowicz, P.R., Castro, E.A. and Toropov, A.A. (2002) Improved QSPR analysis of standard entropy of acyclic and aromatic compounds using optimized correlation weights of linear graph invariants. *Computers Chem.*, **26**, 327–332.
- Duchowicz, P.R., Castro, E.A., Toropov, A.A., Nesterov, I.V. and Nabiev, O.M. (2004) QSPR modeling the aqueous solubility of alcohols by optimization of correlation weights of local graph invariants. *Mol. Div.*, **8**, 325–330.
- Duchowicz, P.R., Pérez González, M., Helguera, A. M., Soeiro Cordeiro, N.D. and Castro, E.A. (2007) Application of the replacement method as novel variable selection in QSPR. 2. Soil sorption coefficients. *Chemom. Intell. Lab. Syst.*, **88**, 197–203.
- Duchowicz, P.R., Sinani, R.G., Castro, E.A. and Toropov, A.A. (2003) Maximum topological distances based indices as molecular descriptors for QSPR. V. Modeling the free energy of hydrocarbons. *Indian J. Chem.*, **42**, 1354–1359.
- Duchowicz, P.R., Vitale, M.G., Castro, E.A., Fernández, M. and Caballero, J. (2007) QSAR analysis for heterocyclic antifungals. *Bioorg. Med. Chem.*, **15**, 2680–2689.
- Ducrot, P., Andrianjara, C.R. and Wrigglesworth, R. (2001) CoMFA and CoMSIA 3D-quantitative structure–activity relationship model on benzodiazepine derivatives, inhibitors of phosphodiesterase. IV. *J. Comput. Aid. Mol. Des.*, **15**, 767–785.
- Duewer, D.L. (1990) The Free–Wilson paradigm *redux*: significance of the Free–Wilson coefficients, insignificance of coefficient ‘uncertainties’ and statistical sins. *J. Chemom.*, **4**, 299–321.
- Duffy, J.C., Dearden, J.C. and Rostron, C. (1996) A QSAR study of antiinflammatory *N*-arylanthranilic acids. *J. Pharm. Pharmacol.*, **48**, 883–886.
- Dunbar, J.B., Jr (1997) Cluster-based selection. *Persp. Drug Disc. Des.*, **7/8**, 51–63.
- Dunn, W.J. III (1977) Molar refractivity as an independent variable in quantitative structure–activity studies. *Eur. J. Med. Chem.*, **12**, 109–112.
- Dunn, W.J., III, Koehler, M.G. and Grigoras, S. (1987) The role of solvent-accessible surface area in determining partition coefficients. *J. Med. Chem.*, **30**, 1121–1126.
- Dunn, W.J. III, and Rogers, D. (1996) Genetic partial least squares in QSAR, in *Genetic Algorithms in Molecular Modeling. Principles of QSAR and Drug Design*, Vol. 1 (ed. J. Devillers), Academic Press, London, UK, pp. 109–130.
- Dunn, W.J. III, and Wold, S. (1978) A structure–carcinogenicity study of 4-nitroquinoline 1-oxides using the SIMCA method of pattern recognition. *J. Med. Chem.*, **21**, 1001–1007.
- Dunn, W.J. III, and Wold, S. (1980) Structure–activity analyzed by pattern recognition: the asymmetric case. *J. Med. Chem.*, **23**, 595–599.
- Dunn, W.J. III, and Wold, S. (1990) Pattern recognition techniques in drug design, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 691–714.
- Dunn, W.J., III, Wold, S., Edlund, U. and Hellberg, S. (1984) Multivariate structure–activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method. *Quant. Struct. -Act. Relat.*, **3**, pp. 131–137.
- Dunn, W.J., III, Wold, S. and Martin, Y.C. (1978) Structure–activity study of β -adrenergic agents using the SIMCA method of pattern recognition. *J. Med. Chem.*, **21**, 922–930.
- Dunnivant, F.M., Elzerman, A.W., Jurs, P.C. and Hasan, M.N. (1992) Quantitative structure–property relationships for aqueous solubilities and Henry’s law constants of polychlorinated biphenyls. *Environ. Sci. Technol.*, **26**, 1567–1573.
- Duperray, B., Chastrette, M., Makabeth, M.C. and Pacheco, H. (1976a) Analyse Comparative de Corrélations Hansch, Free–Wilson et Darc–Pelco pour une Famille de Bactéricides: des Phénols Halogénés. *Eur. J. Med. Chem.*, **11**, 323–336.
- Duperray, B., Chastrette, M., Makabeth, M.C. and Pacheco, H. (1976b) Analyse de l’Activité Bactéricide de Populations d’Alcools Aliphatiques et de β -Naptols suivant les Méthodes de Hansch et Darc–Pelco: Effet d’Allongement de Chaîne. *Eur. J. Med. Chem.*, **11**, 433–437.
- Duprat, A.F., Huynh, T. and Dreyfus, G. (1998) Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of log P. *J. Chem. Inf. Comput. Sci.*, **38**, 586–594.
- Durand, P.J., Pasari, R., Baker, J.W. and Tsai, C. (1999) An efficient algorithm for similarity analysis of molecules. *Internet J. Chem.*, **2** (17), 1–16.
- Durant, J.L., Leland, B.A., Henry, D.R. and Nourse, J. G. (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 1273–1280.
- Dureja, H. and Madan, A.K. (2006) Prediction of h5-HT_{2A} receptor antagonistic activity of arylindoles:

- computational approach using topochemical descriptors. *J. Mol. Graph. Model.*, **25**, 373–379.
- Dureja, H. and Madan, A.K. (2007) Topochemical models for prediction of telomerase inhibitory activity of flavonoids. *Chem. Biol. Drug Des.*, **70**, 47–52.
- Durst, G.L. (1998) Comparative molecular-field analysis (CoMFA) of herbicidal protoporphyrinogen oxidase-inhibitors using standard steric and electrostatic fields and an alternative LUMO field. *Quant. Struct. -Act. Relat.*, **17**, 419–426.
- Dury, L., Latour, T., Leherter, L., Barberis, F. and Varcauteren, D.P. (2001) A new graph descriptor for molecules containing cycles. Application as screening criterion for searching molecular structures within large databases of organic compounds. *J. Chem. Inf. Comput. Sci.*, **41**, 1437–1445.
- Dutta, D., Dutta, R., Wild, D. and Chen, T. (2007) Ensemble feature selection: consistent descriptor subsets for multiple QSAR models. *J. Chem. Inf. Model.*, **47**, 989–997.
- Dutta, D., Guha, R., Jurs, P.C. and Chen, T. (2006) Scalable partitioning and exploration of chemical spaces using geometric hashing. *J. Chem. Inf. Model.*, **46**, 321–333.
- Dyckjaer, J.D. and Jónsdóttir, S.Ó. (2003) QSPR models based on molecular mechanics and quantum chemical calculations. 2. Thermodynamic properties of alkanes, alcohols, polyols, and ethers. *Ind. Eng. Chem. Res.*, **42**, 4241–4259.
- Dziembowska, T. (1994) Intramolecular hydrogen bonding. *Pol. J. Chem.*, **68**, 1455–1489.
- Eckert, H. and Bajorath, J. (2006a) Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J. Chem. Inf. Model.*, **46**, 2515–2526.
- Eckert, H. and Bajorath, J. (2006b) Determination and mapping of activity-specific descriptor value ranges for the identification of active compounds. *J. Med. Chem.*, **49**, 2284–2293.
- Eckert, H. and Bajorath, J. (2007a) Exploring peptide-likeness of active molecules using 2D fingerprint methods. *J. Chem. Inf. Model.*, **47**, 1366–1378.
- Eckert, H. and Bajorath, J. (2007b) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today*, **12**, 225–233.
- Eckert, H., Vogt, I. and Bajorath, J. (2006) Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: design of DynaMAD and comparison with MAD and DMC. *J. Chem. Inf. Model.*, **46**, 1623–1634.
- Edvinsson, T., Arteca, G.A. and Elvingson, C. (2003) Path–space ratio as a molecular shape descriptor of polymer conformation. *J. Chem. Inf. Comput. Sci.*, **43**, 126–133.
- Edward, J.T. (1982a) Correlation of alkane solubilities in water with connectivity index. *Can. J. Chem.*, **60**, 2573–2578.
- Edward, J.T. (1982b) The relation of physical properties of alkanes to connectivity indices: a molecular explanation. *Can. J. Chem.*, **60**, 480–485.
- Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Planes*, Society for Industrial and Applied Mathematics, Philadelphia, PA, p. 92.
- Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Ass.*, **78**, 316–331.
- Efron, B. (1987) Better bootstrap confidence intervals. *J. Am. Stat. Ass.*, **82**, 171–200.
- Efroymson, M.A. (1960) Multiple regression analysis, in *Mathematical Methods for Digital Computers* (eds. A. Ralston and H.S. Wilf), John Wiley & Sons, Inc., New York.
- Egan, W.J. and Lauri, G. (2002) Prediction of intestinal permeability. *Adv. Drug Deliv. Rev.*, **54**, 273–289.
- Egan, W.J., Merz, K.M., Jr and Baldwin, J.J. (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.*, **43**, 3867–3877.
- Egghe, L. (1987) Pratt's measure for some bibliometric distributions and its relation with the 80/20 rule. *Journal of the American Society for Information Science*, **38**, 288–297.
- Egolf, L.M. and Jurs, P.C. (1992) Estimation of autoignition temperatures of hydrocarbons, alcohols and esters from molecular structure. *Ind. Eng. Chem. Res.*, **31**, 1798–1807.
- Egolf, L.M. and Jurs, P.C. (1993a) Prediction of boiling points of organic heterocyclic compounds using regression and neural network techniques. *J. Chem. Inf. Comput. Sci.*, **33**, 616–625.
- Egolf, L.M. and Jurs, P.C. (1993b) Quantitative structure–retention and structure–odor intensity relationships for a diverse group of odor-active compounds. *Anal. Chem.*, **65**, 3119–3126.
- Egolf, L.M., Wessel, M.D. and Jurs, P.C. (1994) Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **34**, 947–956.
- Ehrenson, S., Brownlee, R.T.C. and Taft, R.W. (1973) Generalized treatment of substituent effects in the benzene series. Statistical analysis by the dual substituent parameter equation. I. *Prog. Phys. Org. Chem.*, **10**, 1–80.

- Ehresmann, B., De Groot, M.J. and Clark, T. (2005) Surface-integral QSPR models: local energy properties. *J. Chem. Inf. Model.*, **45**, 1053–1060.
- Eike, D.M., Brennecke, J.F. and Maginn, E.J. (2003) Predicting melting points of quaternary ammonium ionic liquids. *Green Chemistry*, **5**, 323–328.
- Eisenberg, D. and McLachlan, A.D. (1986) Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
- Ekins, S., Boulanger, B., Swaan, P.W. and Hupcey, M. A.Z. (2002) Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput. Aid. Mol. Des.*, **16**, 381–401.
- Ekins, S., Bravi, G., Binkley, S., Gillespie, J.S., Ring, B. J., Wikel, J.H. and Wrighton, S.A. (1999a) Three and four dimensional-quantitative structure–activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics*, **9**, 477–489.
- Ekins, S., Bravi, G., Binkley, S., Gillespie, J.S., Ring, B. J., Wikel, J.H. and Wrighton, S.A. (1999b) Three- and four-dimensional quantitative structure–activity relationship analyses of cytochrome P-450 3A4 inhibitors. *J. Pharmacol. Exp. Ther.*, **290**, 429–438.
- Ekins, S., Bravi, G., Ring, B.J., Gillespie, T.A., Gillespie, J.S., Vandenbranden, M., Wrighton, S.A. and Wikel, J.H. (1999c) Three-dimensional quantitative structure–activity relationship analyses of substrates for CYP2B6. *J. Pharmacol. Exp. Ther.*, **288**, 21–29.
- Ekins, S., Durst, G.L., Stratford, R.E., Thorner, D.A., Lewis, R., Loncharich, R.J. and Wikel, J.H. (2001) Three-dimensional quantitative structure–permeability relationship analysis for a series of inhibitors of rhinovirus replication. *J. Chem. Inf. Comput. Sci.*, **41**, 1578–1586.
- Ekins, S., Mestres, J. and Testa, B. (2007) *In silico* pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Brit. J. Pharmacol.*, **152**, 9–20.
- Ekins, S. and Rose, J. (2002) *In silico* ADME/Tox: the state of the art. *J. Mol. Graph. Model.*, **20**, 305–309.
- El Tayar, N. and Testa, B. (1993) Polar intermolecular interactions encoded in partition coefficients and their interest in QSAR, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 101–108.
- El Tayar, N., Testa, B. and Carrupt, P.-A. (1992) Polar intermolecular interactions encoded in partition coefficients: an indirect estimation of hydrogen-bond parameters of polyfunctional solutes. *J. Phys. Chem.*, **96**, 1455–1459.
- El Tayar, N., Tsai, R.-S., Carrupt, P.-A. and Testa, B. (1992) Octan-1-ol–water partition coefficients of zwitterionic α -amino acids. Determination by centrifugal partition chromatography and factorization into steric/hydrophobic and polar components. *J. Chem. Soc. Perkin Trans. 2*, 79–84.
- El Tayar, N., Tsai, R.-S., Testa, B., Carrupt, P.-A., Hansch, C., and Leo, A. (1991a) Percutaneous penetration of drugs: a quantitative structure–permeability relationship study. *J. Pharm. Sci.*, **80**, 744–749.
- El Tayar, N., Tsai, R.-S., Testa, B., Carrupt, P.-A. and Leo, A. (1991b) Partitioning of solutes in different solvent systems: the contribution of hydrogen-bonding capacity and polarity. *J. Pharm. Sci.*, **80**, 590–598.
- El-Taher, S., El-sawy, K.M. and Hilal, R. (2002) Electronic structure of some adenosine receptor antagonists. VQSAR investigation. *J. Chem. Inf. Comput. Sci.*, **42**, 386–392.
- Elango, M., Parthasarathi, R., Subramanian, V., Sarkar, U. and Chattaraj, P.K. (2005) Formaldehyde decomposition through profiles of global reactivity indices. *J. Mol. Struct. (Theochem)*, **723**, 43–52.
- Elass, A., Brocard, J., Surpateanu, G. and Vergoten, G. (1999) Conformational searching using the comparative molecular field analysis (CoMFA) method of substituted arene-tricarbonyl-chromium complexes. *J. Mol. Struct. (Theochem)*, **466**, 21–33.
- Elass, A., Vergoten, G., Legrand, D., Mazurier, J., Ellassrochard, E. and Spik, G. (1996a) Processes underlying interactions of human lactoferrin with the Jurkat human lymphoblastic T-cell line receptor. 1. Quantitative structure–affinity relationships studies. *Quant. Struct. -Act. Relat.*, **15**, 94–101.
- Elass, A., Vergoten, G., Legrand, D., Mazurier, J., Ellassrochard, E. and Spik, G. (1996b) Processes underlying interactions of human lactoferrin with the Jurkat human lymphoblastic T-cell line receptor. 2. Comparative molecular field analysis. *Quant. Struct. -Act. Relat.*, **15**, 102–107.
- Elbro, H.S., Fredeslund, A. and Rasmussen, P. (1991) Group contribution method for the prediction of liquid densities as function of temperature for solvents, oligomers, and polymers. *Ind. Eng. Chem. Res.*, **30**, 2576–2582.
- Eldred, D.V. and Jurs, P.C. (1999) Prediction of acute mammalian toxicity of organophosphorus pesticide compounds from molecular structure. *SAR & QSAR Environ. Res.*, **10**, 75–99.
- Eldred, D.V., Weikel, C.L., Jurs, P.C. and Kaiser, K.L. E. (1999) Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. *Chem. Res. Toxicol.*, **12**, 670–678.

- Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G. V. and Mee, R.P. (1997) Empirical scoring functions. I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aid. Mol. Des.*, **11**, 425–445.
- Elhallaoui, M., Elasri, M., Ouazzani, F., Mechaqrane, A. and Lakhliifi, T. (2003) Quantitative structure–activity relationships of noncompetitive antagonists of the NMDA receptor: a study of a series of MK801 derivative molecules using statistical methods and neural network. *Int. J. Mol. Sci.*, **4**, 249–262.
- Elk, S.B. (1990) A canonical ordering of polybenzenes and polymantanes using a prime number factorization technique. *J. Math. Chem.*, **4**, 55–68.
- Elk, S.B. (1995) Expansion of Matula numbers to heteroatoms and to ring compounds. *J. Chem. Inf. Comput. Sci.*, **35**, 233–236.
- Elk, S.B. and Gutman, I. (1994) Further properties derivable from the Matula numbers of an alkane. *J. Chem. Inf. Comput. Sci.*, **34**, 331–334.
- Engelhardt McClelland, H. and Jurs, P.C. (2000) Quantitative structure–property relationships for the prediction of vapor pressures of organic compounds from molecular structures. *J. Chem. Inf. Comput. Sci.*, **40**, 967–975.
- Engelhardt, H.L. and Jurs, P.C. (1997) Prediction of supercritical carbon dioxide solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **37**, 478–484.
- Engkvist, O. and Wrede, P. (2002) High-throughput *in silico* prediction of aqueous solubility based on one- and two-dimensional descriptors. *J. Chem. Inf. Comput. Sci.*, **42**, 1247–1249.
- Engkvist, O., Wrede, P. and Rester, U. (2003) Prediction of CNS activity of compound libraries using substructure analysis. *J. Chem. Inf. Comput. Sci.*, **43**, 155–160.
- English, N.J. and Carroll, D.G. (2001) Prediction of Henry's law constants by a quantitative structure–property relationship and neural networks. *J. Chem. Inf. Comput. Sci.*, **41**, 1150–1161.
- Entiger, R.C., Jackson, D.E. and Snyder, D.A. (1976) Distance in graphs. *Czech. Math. J.*, **26**, 283–296.
- Ergün, U., Baryıbcı, N., Ozan, A.T., Serhatlıoğlu, S., Ogur, E., Hardalaç, F. and Güler, I. (2004) Classification of MCA stenosis in diabetes by MLP and RBF neural network. *J. Med. Syst.*, **28**, 475–487.
- Eriksson, L., Andersson, P.L., Johansson, E. and Tysklind, M. (2002) Multivariate biological profiling and principal regions of compounds: the PCB case study. *J. Chemom.*, **16**, 497–509.
- Eriksson, L., Antti, H., Holmes, E., Johansson, E., Lundstedt, L., Shockcor, J. and Wold, S. (2003) Partial least squares (PLS) in cheminformatics, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1134–1166.
- Eriksson, L., Arnhold, T., Beck, B., Fox, T., Johansson, E. and Kriegl, J.M. (2004) Onion design and its application to a pharmaceutical QSAR problem. *J. Chemom.*, **18**, 188–202.
- Eriksson, L., Berglind, R. and Sjöström, M. (1994) A multivariate quantitative structure–activity relationship for corrosive carboxylic acids. *Chemom. Intell. Lab. Syst.*, **23**, 235–245.
- Eriksson, L., Gottfries, J., Johansson, E. and Wold, S. (2004) Time-resolved QSAR: an approach to PLS modelling of three-way biological data. *Chemom. Intell. Lab. Syst.*, **73**, 73–84.
- Eriksson, L., Hermens, J.L.M., Johansson, E., Verhaar, H.J.M. and Wold, S. (1995) Multivariate analysis of aquatic toxicity data with PLS. *Aquat. Sci.*, **57**, 217–241.
- Eriksson, L., Jaworska, J.S., Worth, A.P., Cronin, M.T. D., McDowell, R.M. and Gramatica, P. (2003) Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs. *Environ. Health Persp.*, **111**, 1361–1375.
- Eriksson, L. and Johansson, E. (1996) Multivariate design and modeling in QSAR. *Chemom. Intell. Lab. Syst.*, **34**, 1–19.
- Eriksson, L., Johansson, E., Lindgren, F., Sjöström, M. and Wold, S. (2002) Megavariate analysis of hierarchical QSAR data. *J. Comput. Aid. Mol. Des.*, **16**, 711–726.
- Eriksson, L., Johansson, E., Lindgren, F. and Wold, S. (2000a) GIF-PLS: modeling of non-linearities and discontinuities in QSAR. *Quant. Struct. -Act. Relat.*, **19**, 345–355.
- Eriksson, L., Johansson, E., Müller, M. and Wold, S. (1997) Cluster-based design in environmental QSAR. *Quant. Struct. -Act. Relat.*, **16**, 383–390.
- Eriksson, L., Johansson, E., Müller, M. and Wold, S. (2000b) On the selection of the training set in environmental QSAR analysis when compounds are clustered. *J. Chemom.*, **14**, 599–616.
- Eriksson, L., Jonsson, J. and Berglind, R. (1993) External validation of a QSAR for the acute toxicity of halogenated aliphatic hydrocarbons. *Environ. Toxicol. Chem.*, **12**, 1185–1191.
- Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Skagerberg, B., Sjöström, M., Wold, S. and Berglind, R. (1990) A strategy for ranking environmentally occurring chemicals. Part III.

- Multivariate quantitative structure–activity relationships for halogenated aliphatics. *Environ. Toxicol. Chem.*, **9**, 1339–1351.
- Eriksson, L., Jonsson, J., Sjöström, M., Wikström, C. and Wold, S. (1988) Multivariate derivation of descriptive scales for monosaccharides. *Acta Chem. Scand.*, **42**, 504–514.
- Eriksson, L., Sandström, B.E., Sjöström, M., Tysklind, M. and Wold, S. (1993) Modelling the cytotoxicity of halogenated aliphatic hydrocarbons. Quantitative structure–activity relationships for the IC₅₀ to human HeLa cells. *Quant. Struct. -Act. Relat.*, **12**, 124–131.
- Eriksson, L., Verboom, H.H. and Peijnenburg, W.J.G.M. (1996) Multivariate QSAR modelling of the rate of reductive dehalogenation of haloalkanes. *J. Chemom.*, **10**, 483–492.
- Eriksson, L., Verhaar, H.J.M. and Hermens, J.L.M. (1994) Multivariate characterization and modeling of the chemical reactivity of epoxides. *Environ. Toxicol. Chem.*, **13**, 683–691.
- Erlenmeyer, E. (1866) Studien über die s.g. aromatischen Säuren. *Ann. Chemie Pharm.*, **137**, 327–359.
- Ermondi, G., Lorenti, M. and Caron, G. (2004) Contribution of ionization and lipophilicity to drug binding to album: a preliminary step toward biodistribution prediction. *J. Med. Chem.*, **47**, 3949–3961.
- Ertepinar, H., Gok, Y., Geban, O. and Ozden, S. (1995) A QSAR study of the biological activities of some benzimidazoles and imidazopyridines against *Bacillus subtilis*. *Eur. J. Med. Chem.*, **30**, 171–175.
- Ertl, P. (1997) Simple quantum-chemical parameters as an alternative to the Hammett sigma-constant in QSAR studies. *Quant. Struct. -Act. Relat.*, **16**, 377–382.
- Ertl, P. (1998a) QSAR analysis through the World Wide Web. *Chimia*, **52**, 673–677.
- Ertl, P. (1998b) World Wide Web-based system for the calculation of substituents parameters and substituents similarity searches. *J. Mol. Graph. Model.*, **16**, 11–13.
- Ertl, P., Rhode, B. and Selzer, P. (2000) Web-based chemoinformatics for bench chemists. *Drug Discovery World*, **Fall 2000**, 45–50.
- Ertl, P. (2003) Chemoinformatics analysis of organic substituents: identification of the most common substituents. Calculation of substituents properties and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.*, **43**, 374–380.
- Ertl, P. (2008) Polar surface area, in *Molecular Drug Properties*, Vol. 37 (ed. R. Mannhold.), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 111–126.
- Ertl, P. and Jacob, O. (1997) WWW-based chemical information system. *J. Mol. Struct. (Theochem)*, **419**, 113–120.
- Ertl, P., Rohde, B. and Selzer, P. (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, **43**, 3714–3717.
- Ertl, P., Rohde, B. and Selzer, P. (2001) Calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 451–458.
- Ertl, P. and Selzer, P. (2003) Web-based calculation of molecular properties, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1336–1348.
- Esaki, T. (1980) Quantitative drug design studies. II. Development and application of new electronic substituent parameters. *J. Pharmacobiodyn.*, **3**, 562–576.
- Espeso, V.G., Molins Vara, J.J., Roy Lázaro, B., Riera Parcerisas, F. and Plavšić, D. (2000) On the Hosoya hyperindex and the molecular indices based on a new decomposition of the Hosoya Z matrix. *Croat. Chem. Acta*, **73**, 1017–1026.
- Espinosa, G., Arenas, A. and Giral, F. (2002) An integrated SOM-fuzzy ARTMAP neural system for the evaluation of toxicity. *J. Chem. Inf. Comput. Sci.*, **42**, 343–359.
- Espinosa, G., Yaffe, D., Arenas, A., Cohen, Y. and Giral, F. (2001) A fuzzy ARTMAP-based quantitative structure–property relationship (QSPR) for predicting physical properties of organic compounds. *Ind. Eng. Chem. Res.*, **40**, 2757–2766.
- Esposito, E.X., Hopfinger, A.J. and Madura, J.D. (2003) 3D- and nD-QSAR methods, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1576–1603.
- Essam, J.W., Kennedy, J.W. and Gordon, M. (1977) The graph-like state of matter. Part 8. LCGI schemes and the statistical analysis of experimental data. *J. Chem. Soc. Faraday Trans II*, **73**, 1289–1307.
- Esteban-Diez, I., González-Sáiz, J.M., Gómez-Cámara, D. and Pizarro Millán, C. (2006a) Multivariate calibration of near infrared spectra by orthogonal WAVElet correction using a genetic algorithm. *Anal. Chim. Acta*, **555**, 84–95.

- Esteban-Diez, I., González-Sáiz, J.M., Pizarro Millán, C. and Forina, M. (2006b) GA-ACE: alternating conditional expectations regression with selection of significant predictors by genetic algorithms. *Anal. Chim. Acta*, **555**, 96–106.
- Estrada, E. (1995a) Edge adjacency relationships and a novel topological index related to molecular volume. *J. Chem. Inf. Comput. Sci.*, **35**, 31–33.
- Estrada, E. (1995b) Edge adjacency relationships in molecular graphs containing heteroatoms: a new topological index related to molar volume. *J. Chem. Inf. Comput. Sci.*, **35**, 701–707.
- Estrada, E. (1995c) Graph theoretical invariant of Randić revisited. *J. Chem. Inf. Comput. Sci.*, **35**, 1022–1025.
- Estrada, E. (1995d) Three-dimensional molecular descriptors based on electron charge density weighted graphs. *J. Chem. Inf. Comput. Sci.*, **35**, 708–713.
- Estrada, E. (1996) Spectral moments of the edge adjacency matrix of molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *J. Chem. Inf. Comput. Sci.*, **36**, 844–849.
- Estrada, E. (1997) Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *J. Chem. Inf. Comput. Sci.*, **37**, 320–328.
- Estrada, E. (1998a) Modelling the diamagnetic susceptibility of organic compounds by a sub-structural graph-theoretical approach. *J. Chem. Soc. Faraday Trans.*, **94**, 1407–1410.
- Estrada, E. (1998b) Spectral moments of the edge adjacency matrix in molecular graphs. 3. Molecules containing cycles. *J. Chem. Inf. Comput. Sci.*, **38**, 23–27.
- Estrada, E. (1999a) Connectivity polynomial and long-range contributions in the molecular connectivity model. *Chem. Phys. Lett.*, **312**, 556–560.
- Estrada, E. (1999b) Edge-connectivity indices in QSPR/QSAR studies. 2. Accounting for long-range bond contributions. *J. Chem. Inf. Comput. Sci.*, **39**, 1042–1048.
- Estrada, E. (1999c) Generalized spectral moments of the iterated line graph sequence. A novel approach to QSPR studies. *J. Chem. Inf. Comput. Sci.*, **39**, 90–95.
- Estrada, E. (1999d) Novel strategies in the search of topological indices, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 403–453.
- Estrada, E. (2000) Characterization of 3D molecular structure. *Chem. Phys. Lett.*, **319**, 713–718.
- Estrada, E. (2001) Generalization of topological indices. *Chem. Phys. Lett.*, **336**, 248–252.
- Estrada, E. (2002a) Characterization of the folding degree of proteins. *Bioinformatics*, **18**, 697–704.
- Estrada, E. (2002b) Physico-chemical interpretation of molecular connectivity indices. *J. Phys. Chem. A*, **106**, 9085–9091.
- Estrada, E. (2003a) Application of a novel graph-theoretic folding degree index to the study of steroid-DB3 antibody binding affinity. *Comp. Biol. Chem.*, **27**, 305–313.
- Estrada, E. (2003b) Generalized graph matrix, graph geometry, quantum chemistry, and optimal description of physico-chemical properties. *J. Phys. Chem. A*, **107**, 7482–7489.
- Estrada, E. (2004a) A protein folding degree measure and its dependence on crystal packing, protein size, secondary structure, and domain structural class. *J. Chem. Inf. Comput. Sci.*, **44**, 1238–1250.
- Estrada, E. (2004b) Characterization of the amino acid contribution to the folding degree of proteins. *Prot. Struct. Funct. Bioinf.*, **54**, 727–737.
- Estrada, E. (2004c) Three-dimensional generalized graph matrix, Harary descriptors, and a generalized interatomic Lennard–Jones potential. *J. Phys. Chem. A*, **108**, 5468–5473.
- Estrada, E. (2006a) On the dimensionality of aromaticity criteria. *MATCH Commun. Math. Comput. Chem.*, **56**, 331–344.
- Estrada, E. (2006b) Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, **6**, 35–40.
- Estrada, E. (2007) Topological structural classes of complex networks. *Phys. Rev. E*, **75**, 016103.
- Estrada, E. (2008) How the parts organize in the whole? A top-down view of molecular descriptors and properties for QSAR and drug design. *Mini Rev. Med. Chem.*, **8**, 213–221.
- Estrada, E. and Avnir, D. (2003) Continuous symmetry numbers and entropy. *J. Am. Chem. Soc.*, **125**, 4368–4375.
- Estrada, E., Delgado, E.J., Alderete, J.B. and Jaña, G.A. (2006) Quantum-connectivity descriptors in modeling solubility of environmentally important organic compounds. *J. Comput. Chem.*, **25**, 1787–1796.
- Estrada, E. and Gonzalez, H. (2003) What are the limits of applicability for graph theoretic descriptors in QSPR/QSAR? Modeling dipole moments of aromatic compounds with TOPS-MODE descriptors. *J. Chem. Inf. Comput. Sci.*, **43**, 75–84.
- Estrada, E., Guevara, N. and Gutman, I. (1998a) Extension of edge connectivity index.

- Relationships to line graph indices and QSPR applications. *J. Chem. Inf. Comput. Sci.*, **38**, 428–431.
- Estrada, E., Guevara, N., Gutman, I. and Rodriguez, L. (1998b) Molecular connectivity indices of iterated line graphs. A new source of descriptors for QSPR and QSAR studies. *SAR & QSAR Environ. Res.*, **9**, 229–240.
- Estrada, E. and Gutierrez, Y. (1999) Modeling chromatographic parameters by a novel graph theoretical sub-structural approach. *J. Chromat.*, **858**, 187–199.
- Estrada, E. and Gutierrez, Y. (2001) The Balaban *J* index in the multidimensional space of generalized topological indices. Generalizations and QSPR improvements. *MATCH Commun. Math. Comput. Chem.*, **44**, 155–167.
- Estrada, E., Gutierrez, Y. and Gonzalez, H. (2000) Modeling diamagnetic and magnetooptic properties of organic compounds with the TOSS-MODE approach. *J. Chem. Inf. Comput. Sci.*, **40**, 1386–1399.
- Estrada, E. and Gutman, I. (1996) A topological index based on distances of edges of molecular graphs. *J. Chem. Inf. Comput. Sci.*, **36**, 850–853.
- Estrada, E. and Hatano, N. (2007) Statistical-mechanical approach to subgraph centrality in complex networks. *Chem. Phys. Lett.*, **439**, 247–251.
- Estrada, E., Ivanciuc, O., Gutman, I., Gutierrez, A. and Rodriguez, L. (1998) Extended Wiener indices. A new set of descriptors for quantitative structure–property studies. *New J. Chem.*, **22**, 819–822.
- Estrada, E. and Matamala, A.R. (2007) Generalized topological indices. Modeling gas-phase rate coefficients of atmospheric relevance. *J. Chem. Inf. Model.*, **47**, 794–804.
- Estrada, E. and Molina, E. (2001a) 3D connectivity indices in QSPR/QSAR studies. *J. Chem. Inf. Comput. Sci.*, **41**, 791–797.
- Estrada, E. and Molina, E. (2001b) Novel local (fragment-based) topological molecular descriptors for QSPR/QSAR and molecular design. *J. Mol. Graph. Model.*, **20**, 54–64.
- Estrada, E. and Molina, E. (2001c) QSPR/QSAR by graph theoretical descriptors. Beyond the frontiers, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 83–107.
- Estrada, E., Molina, E. and Perdomo-López, I. (2001a) Can 3D structural parameters be predicted from 2D (topological) molecular descriptors? *J. Chem. Inf. Comput. Sci.*, **41**, 1015–1021.
- Estrada, E., Molina, E. and Uriarte, E. (2001b) Quantitative structure–toxicity relationships using TOPS-MODE. 2. Neurotoxicity of a noncongeneric series of solvents. *SAR & QSAR Environ. Res.*, **12**, 445–459.
- Estrada, E. and Montero, L.A. (1993) Bond order weighted graphs in molecules as structure–property indices. *Mol. Eng.*, **2**, 363–373.
- Estrada, E., Paltewicz, G. and Uriarte, E. (2003) From molecular graphs to drugs. A review on the use of topological indices in drug design and discovery. *Indian J. Chem.*, **42**, 1315–1329.
- Estrada, E. and Patlewicz, G. (2004) On the usefulness of graph-theoretic descriptors in predicting theoretical parameters. Phototoxicity of polycyclic aromatic hydrocarbons (PAHs). *Croat. Chem. Acta*, **77**, 203–211.
- Estrada, E., Patlewicz, G., Chamberlain, M., Basketter, D. and Larbey, S. (2003) Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. *Chem. Res. Toxicol.*, **16**, 1226–1235.
- Estrada, E., Patlewicz, G. and Gutierrez, Y. (2004) From knowledge generation to knowledge archive. A general strategy using TOPS-MODE with DEREK to formulate new alerts for skin sensitization. *J. Chem. Inf. Comput. Sci.*, **44**, 688–698.
- Estrada, E. and Peña, A. (2000) *In silico* studies for the rational discovery of anticonvulsant compounds. *Bioorg. Med. Chem.*, **8**, 2755–2770.
- Estrada, E., Peña, A. and García-Domenech, R. (1998) Designing sedative/hypnotic compounds from a novel substructural graph-theoretical approach. *J. Comput. Aid. Mol. Des.*, **12**, 583–595.
- Estrada, E., Perdomo-López, I. and Torres-Labandeira, J.J. (2001) Combination of 2D-, 3D-connectivity and quantum chemical descriptors in QSPR. Complexation of α - and β -cyclodextrin with benzene derivatives. *J. Chem. Inf. Comput. Sci.*, **41**, 1561–1568.
- Estrada, E., Quincoces, J. and Patlewicz, G. (2004) Creating molecular diversity from antioxidants in Brazilian propolis. Combination of TOPS-MODE QSAR and virtual structure generation. *Mol. Div.*, **8**, 21–33.
- Estrada, E. and Ramirez, A. (1996) Edge adjacency relationships and molecular topographic descriptors. Definition and QSAR applications. *J. Chem. Inf. Comput. Sci.*, **36**, 837–843.
- Estrada, E. and Rodriguez, L. (1997) Matrix algebraic manipulation of molecular graphs. 2. Harary- and MTI-like molecular descriptors. *MATCH Commun. Math. Comput. Chem.*, **35**, 157–167.
- Estrada, E. and Rodriguez, L. (1999) Edge-connectivity Indices in QSPR/QSAR studies. 1.

- Comparison with other topological indices in QSPR studies. *J. Chem. Inf. Comput. Sci.*, **39**, 1037–1041.
- Estrada, E., Rodríguez, L. and Gutiérrez, A. (1997) Matrix algebraic manipulation of molecular graphs. 1. Distance and vertex-adjacency matrices. *MATCH Commun. Math. Comput. Chem.*, **35**, 145–156.
- Estrada, E. and Rodríguez-Velásquez, J.A. (2005a) Spectral measures of bipartivity in complex networks. *Phys. Rev. E*, **72**, 046105.
- Estrada, E. and Rodríguez-Velásquez, J.A. (2005b) Subgraph centrality and clustering in complex networks. *Phys. Rev. E*, **71**, 056103–1–056103/9.
- Estrada, E. and Rodríguez-Velásquez, J.A. (2006a) Atomic branching in molecules. *Int. J. Quant. Chem.*, **106**, 823–832.
- Estrada, E. and Rodríguez-Velásquez, J.A. (2006b) Subgraph centrality and clustering in complex hyper-networks. *Physica A*, **364**, 581–594.
- Estrada, E., Torres, L., Rodríguez, L. and Gutman, I. (1998) An atom-bond connectivity index: modelling the enthalpy of formation of alkanes. *Indian J. Chem.*, **37**, 849–855.
- Estrada, E. and Uriarte, E. (2001a) Quantitative structure–toxicity relationships using TOPS-MODE. 1. Nitrobenzene toxicity to *Tetrahymena pyriformis*. *SAR & QSAR Environ. Res.*, **12**, 309–324.
- Estrada, E. and Uriarte, E. (2001b) Recent advances on the role of topological indices in drug discovery research. *Curr. Med. Chem.*, **8**, 1573–1588.
- Estrada, E. and Uriarte, E. (2005) Folding degrees of azurins and pseudoazurins. Implications for structure and function. *Comp. Biol. Chem.*, **29**, 345–353.
- Estrada, E., Uriarte, E., Gutierrez, Y. and Gonzalez, H. (2003) Quantitative structure–toxicity relationships using TOPS-MODE. 3. Structural factors influencing the permeability of commercial solvents through living human skin. *SAR & QSAR Environ. Res.*, **14**, 145–163.
- Estrada, E., Uriarte, E., Montero, A., Teijeira, M., Santana, L. and De Clercq, E. (2000) A novel approach for the virtual screening and rational design of anticancer compounds. *J. Med. Chem.*, **43**, 1975–1985.
- Estrada, E., Uriarte, E. and Vilar, S. (2006) Effect of protein backbone folding on the stability of protein–ligand complexes. *J. Proteome Res.*, **5**, 105–111.
- Estrada, E., Vilar, S., Uriarte, E. and Gutierrez, Y. (2002) *In silico* studies toward the discovery of new anti-HIV nucleoside compounds with the use of TOPS-MODE and 2D/3D connectivity indices. 1. Pyrimidyl derivatives. *J. Chem. Inf. Comput. Sci.*, **42**, 1194–1203.
- Euler, L. (1741) Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis, Petropolitanae*, **8**, 128–140.
- Euler, L. (1758) Elementa doctrinae solidorum and demonstratio nonnullarum insignium proprietatum quibus solida heddis planis inclusa sunt praedita.
- Evans, B.E., Rittle, K.E., Bock, M.G., DiPardo, R.M., Freidinger, R.M., Whitter, W.L., Lundell, G.F., Veber, D.F., Anderson, P.S., Chang, R.S.L., Lotti, V. J., Cerino, D.J., Chen, T.B., Kling, P.J., Kunkel, K. A., Springer, J.P. and Hirshfield, J. (1988) Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.*, **31**, 2235–2246.
- Evans, L.A., Lynch, M.F. and Willett, P. (1978) Structural search codes for online compound registration. *J. Chem. Inf. Comput. Sci.*, **18**, 146–149.
- Evers, A., Hessler, G., Matter, H. and Klabunde, T. (2005) Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. *J. Med. Chem.*, **48**, 5448–5465.
- Ewing, D.F. (1978) Correlation of NMR chemical shifts with Hammett σ values and analogous parameters, in *Correlation Analysis in Chemistry* (eds N.B. Chapman and J. Shorter), Plenum Press, New York, pp. 357–396.
- Exner, O. (1966) Additive physical properties. I. General relations and problems of statistical nature. *Collect. Czech. Chem. Comm.*, **31**, 3222–3251.
- Exner, O. (1973) The enthalpy–entropy relationship. *Prog. Phys. Org. Chem.*, **10**, 411–482.
- Exner, O. (1975) *Dipole Moments in Organic Chemistry*, George Thieme Publishers, Stuttgart, Germany, p. 156.
- Exner, O. (1978) A critical compilation of substituent constants, in *Correlation Analysis in Chemistry* (ed. N.B. Chapman and J. Shorter), Plenum Press, New York, pp. 439–540.
- Exner, O., Ingr, M. and Cársky, P. (1997) *Ab initio* calculations of substituent constants: a reinvestigation. *J. Mol. Struct. (Theochem)*, **397**, 231–238.
- Eyring, H., Walter, J. and Kimball, G. (1944) *Quantum Chemistry*, John Wiley & Sons, Inc., New York.
- Fabian, W.M.F. and Timofei, S. (1996) Comparative molecular field analysis (CoMFA) of dye–fibre affinities. Part 2. Symmetrical bisazo dyes. *J. Mol. Struct. (Theochem)*, **362**, 155–162.
- Fabian, W.M.F., Timofei, S. and Kurunczi, L. (1995) Comparative molecular field analysis (CoMFA), semiempirical (AM1) molecular orbital and

- multiconformational minimal steric difference (MTD) calculations of anthraquinone dye fiber affinities. *J. Mol. Struct. (Theochem)*, **340**, 73–81.
- Fabic-Petrac, I., Jerman-Blazic, B. and Batagelj, V. (1991) study of computation, relatedness and activity prediction of topological indices. *J. Math. Chem.*, **8**, 121–134.
- Fajtlowicz, S., John, P.E. and Sachs, H. (2005) On maximum matchings and eigenvalues of benzenoid graphs. *Croat. Chem. Acta*, **78**, 195–201.
- Faller, B. and Ertl, P. (2007) Computational approaches to determine drug solubility. *Adv. Drug Deliv. Rev.*, **59**, 533–545.
- Famini, G.R., Aguiar, D., Payne, M.A., Rodriguez, R. and Wilson, L.Y. (2002) Using the theoretical linear energy solvation energy relationship to correlate and predict nasal pungency thresholds. *J. Mol. Graph. Model.*, **20**, 277–280.
- Famini, G.R., Ashman, W.P., Mickiewicz, A.P. and Wilson, L.Y. (1992) Using theoretical descriptors in quantitative-structure-activity relationships: opian receptor activity of fentanyl compounds. *Quant. Struct. -Act. Relat.*, **11**, 162–170.
- Famini, G.R., Kassel, R.J., King, J.W. and Wilson, L.Y. (1991) Using theoretical descriptors in quantitative structure-activity relationships: comparison with the molecular transform. *Quant. Struct. -Act. Relat.*, **10**, 344–349.
- Famini, G.R., Loumbev, V.P., Frykman, E.K. and Wilson, L.Y. (1998) Using theoretical descriptors in a correlation analysis of adenosine activity. *Quant. Struct. -Act. Relat.*, **17**, 558–564.
- Famini, G.R., Marquez, B.C. and Wilson, L.Y. (1993) Using theoretical descriptors in quantitative structure-activity relationships: gas phase acidity. *J. Chem. Soc. Perkin Trans. 2*, 773–782.
- Famini, G.R. and Penski, C.A. (1992) Using theoretical descriptors in quantitative structure-activity relationships: some physico-chemical properties. *J. Phys. Org. Chem.*, **5**, 395–408.
- Famini, G.R. and Wilson, L.Y. (1993) Using theoretical descriptors in structure-activity relationships solubility in supercritical CO₂. *J. Phys. Org. Chem.*, **6**, 539–544.
- Famini, G.R. and Wilson, L.Y. (1994a) Using theoretical descriptors in linear solvation energy relationships, in *Quantitative Treatments of Solute/Solvent Interactions: Theoretical and Computational Chemistry* (eds P. Politzer and J.S. Murray), Elsevier, Amsterdam, The Netherlands, pp. 213–241.
- Famini, G.R. and Wilson, L.Y. (1994b) Using theoretical descriptors in quantitative structure-property relationships: 3-carboxy-benzisoxazole decarboxylation kinetics. *J. Chem. Soc. Perkin Trans. 2*, 1641–1650.
- Fan, W., El Tayar, N., Testa, B. and Kier, L.B. (1990) Water-dragging effect: a new experimental hydration parameter related to hydrogen-bond-donor acidity. *J. Phys. Chem.*, **94**, 4764–4766.
- Fan, W., Tsai, R.-S., El Tayar, N., Carrupt, P.-A. and Testa, B. (1994) Solute-water interactions in the organic phase of a biphasic system. 2. Effects of organic phase and temperature on the “water-dragging” effect. *J. Phys. Chem.*, **98**, 329–333.
- Fan, Y., Shi, L.M., Kohn, K.W., Pommier, Y. and Weinstein, J.N. (2001) Quantitative structure-antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies. *J. Med. Chem.*, **44**, 3254–3263.
- Fanelli, F., Menziani, M.C., Carotti, A. and De Benedetti, P.G. (1993) Theoretical quantitative structure-activity analysis of quinuclidine-based muscarinic cholinergic receptor ligands. *J. Mol. Struct. (Theochem)*, **283**, 63–71.
- Fang, H., Tong, W., Shi, L.M., Blair, R., Perkins, R., Branham, W., Hass, B.S., Xie, Q., Dial, S.L., Moland, C.L. and Sheehan, D.M. (2001) Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.*, **14**, 280–294.
- Fang, H., Tong, W., Welsh, W.J. and Sheehan, D.M. (2003) QSAR models in receptor-mediated effects: the nuclear receptor superfamily. *J. Mol. Struct. (Theochem)*, **622**, 113–125.
- Farkas, O. and Héberger, K. (2005) Comparison of ridge regression, partial least-squares, pairwise correlation, forward and best subset selection methods for prediction of retention indices for aliphatic alcohols. *J. Chem. Inf. Model.*, **45**, 339–346.
- Farkas, O., Héberger, K. and Zenkevich, I.G. (2004) Quantitative structure-retention relationships XIV. Prediction of gas chromatographic retention indices for saturated O-, N-, and S-heterocyclic compounds. *Chemom. Intell. Lab. Syst.*, **72**, 173–184.
- Farkas, O., Jakus, J. and Héberger, K. (2004) Quantitative structure-antioxidant activity relationships of flavonoid compounds. *Molecules*, **9**, 1079–1088.
- Farnum, M.A., DesJarlais, R.L. and Agrafiotis, D.K. (2003) Molecular diversity, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1640–1686.
- Fatemi, M.H. (2002) Simultaneous modeling of the Kovats retention indices on OV-1 and SE-54

- stationary phases using artificial neural networks. *J. Chromat.*, **955**, 273–280.
- Fatemi, M.H. (2003) Quantitative structure–property relationship studies of migration index in microemulsion electrokinetic chromatography using artificial neural networks. *J. Chromat.*, **1002**, 221–229.
- Fatemi, M.H. (2004) Prediction of the electrophoretic mobilities of some carboxylic acids from theoretically derived descriptors. *J. Chromat.*, **1038**, 231–237.
- Fauchère, J.L. and Pliška, V. (1983) Hydrophobic parameters of amino acids side chain from the partitioning of *N*-acetyl-amino-acid amides. *Eur. J. Med. Chem.*, **4**, 369–375.
- Fauchère, J.L., Quarendon, P. and Kaetterer, L. (1988) Estimating and representing hydrophobicity potential. *J. Mol. Graph.*, **6**, 203–206.
- Faulon, J.-L. (1998) Isomorphism, automorphism partitioning, and canonical labeling can be solved in polynomial-time for molecular graphs. *J. Chem. Inf. Comput. Sci.*, **38**, 432–444.
- Faulon, J.-L., Churchwell, C.J. and Visco, D.P. (2003) The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.*, **43**, 721–734.
- Faulon, J.-L., Collins, M.J. and Carr, R.D. (2004) The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *J. Chem. Inf. Comput. Sci.*, **44**, 427–436.
- Faulon, J.-L., Visco, D.P. and Pophale, R.S. (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.*, **43**, 707–720.
- Favaron, O., Mahéo, M. and Saclé, J.-F. (2003) The Randić index and other Graffiti parameters of graphs. *MATCH Commun. Math. Comput. Chem.*, **47**, 7–23.
- Fechner, U., Franke, L., Renner, S., Schneider, P. and Schneider, G. (2003) Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aid. Mol. Des.*, **17**, 687–698.
- Fechner, U., Paetz, J. and Schneider, G. (2005) Comparison of three holographic fingerprint descriptors and their binary counterparts. *QSAR Comb. Sci.*, **24**, 961–967.
- Fechner, U. and Schneider, G. (2004a) Evaluation of distance metrics for ligand-based similarity searching. *ChemBioChem*, **5**, 538–540.
- Fechner, U. and Schneider, G. (2004b) Optimization of a pharmacophore-based correlation vector descriptor for similarity searching. *QSAR Comb. Sci.*, **23**, 19–22.
- Fechner, U. and Schneider, G. (2006) Flux (1): a virtual synthesis scheme for fragment-based *de novo* design. *J. Chem. Inf. Model.*, **46**, 699–707.
- Fechner, U. and Schneider, G. (2007) Flux (2): comparison of molecular mutation and crossover operators for ligand-based *de novo* design. *J. Chem. Inf. Model.*, **47**, 656–667.
- Fedorowicz, A., Singh, H., Soderholm, S. and Demchuk, E. (2005) Structure–activity models for contact sensitization. *Chem. Res. Toxicol.*, **18**, 954–969.
- Fedorowicz, A., Zheng, L., Singh, H., and Demchuk, E. (2004) QSAR study of skin sensitization using local lymph node assay data. *Int. J. Mol. Sci.*, **5**, 56–66.
- Feher, M. and Schmidt, J.M. (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.*, **43**, 218–227.
- Feldman, A. and Hodes, L. (1975) An efficient design for chemical structure searching. I. The screens. *J. Chem. Inf. Comput. Sci.*, **15**, 147–152.
- Feng, J., Lurati, L., Ouyang, H., Robinson, T., Wang, Y., Yuan, S. and Young, S.S. (2003) Predictive toxicology: benchmarking molecular descriptors and statistical methods. *J. Chem. Inf. Comput. Sci.*, **43**, 1463–1470.
- Feng, L., Han, S., Wang, L.-S., Wang, Z.-T., and Zhang, Z. (1996a) Determination and estimation of partitioning properties for phenylthio-carboxylates. *Chemosphere*, **32**, 353–360.
- Feng, L., Han, S., Zhao, Y.-H., Wang, L.-S. and Chen, J. (1996b) Toxicity of organic chemicals to fathead minnow: a united quantitative structure–activity relationship model and its application. *Chem. Res. Toxicol.*, **9**, 610–613.
- Ferguson, A.M., Heritage, T.W., Jonathon, P., Pack, S. E., Phillips, L., Rogan, J. and Snaith, P.J. (1997) EVA: a new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput. Aid. Mol. Des.*, **11**, 143–152.
- Ferguson, J. (1939) The use of chemical potentials as indices of toxicity. *Proc. Roy. Soc. London B*, **127**, 387–404.
- Fernández, F.M., Duchowicz, P.R. and Castro, E.A. (2004) About orthogonal descriptors in QSPR/QSAR theories. *MATCH Commun. Math. Comput. Chem.*, **51**, 39–57.
- Ferreira, M.M.C. (2001) Polycyclic aromatic hydrocarbons: a QSPR study. *Chemosphere*, **44**, 125–146.
- Ferreira, M.M.C. and Kiralj, R. (2004) QSAR study of β -lactam antibiotic efflux by the bacterial multidrug resistance pump AcrB. *J. Chemom.*, **18**, 242–252.

- Ferreira, R. (1963a) Principle of electronegativity equalization. Part 1. Bond moments and force constants. *Trans. Faraday Soc.*, **59**, 1064–1074.
- Ferreira, R. (1963b) Principle of electronegativity equalization. Part 2. Bond dissociation energies. *Trans. Faraday Soc.*, **59**, 1075–1079.
- Fichera, M., Cruciani, G., Bianchi, A. and Musumarra, G. (2000) A 3D-QSAR study on the structural requirements for binding to CB1 and CB2 cannabinoid receptors. *J. Med. Chem.*, **43**, 2300–2309.
- Fichert, T., Yazdani, M. and Proudfoot, J.R. (2003) A structure–permeability study of small drug-like molecules. *Bioorg. Med. Chem. Lett.*, **13**, 719–722.
- Figeys, H.P. (1970) Quantum-chemical studies on the aromaticity of conjugated systems-II. Aromatic and anti-aromatic annulenes: the $(4n + 2)$ - π electron rule. *Tetrahedron*, **26**, 5225–5234.
- Figueiredo, L.J.deO. and Garrido, F.M.S. (2001) Chemometric analysis of nonlinear optical chromophores structure and thermal stability. *J. Mol. Struct. (Theochem)*, **539**, 75–81.
- Figueras, J. (1992) Automorphism and equivalence classes. *J. Chem. Inf. Comput. Sci.*, **32**, 153–157.
- Figueras, J. (1993) Morgan revisited. *J. Chem. Inf. Comput. Sci.*, **33**, 717–718.
- Figueras, J. (1996) Ring perception using breadth-first search. *J. Chem. Inf. Comput. Sci.*, **36**, 986–991.
- Filimonov, D., Poroikov, V.V., Borodina, Y. and Glorizova, T. (1999) Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with other descriptors. *J. Chem. Inf. Comput. Sci.*, **39**, 666–670.
- Filimonov, D.A. and Poroikov, V.V. (1996) PASS: computerized prediction of biological activity spectra for chemical substances, in *Bioactive Compound Design: Possibilities for Industrial Use* (eds M.G. Ford, R. Greenwood, G.T. Brooks, and R. Franke.), Bios Scientific Publishers, Portsmouth, UK, pp. 47–56.
- Filip, P.A., Balaban, T.-S. and Balaban, A.T. (1987) A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability. *J. Math. Chem.*, **1**, 61–83.
- Filipponi, E., Cecchetti, V., Tabarrini, O., Bonelli, D. and Fravolini, A. (2000) Chemometric rationalization of the structural and physico-chemical basis for selective cyclooxygenase-2 inhibition: toward more specific ligands. *J. Comput. Aid. Mol. Des.*, **14**, 277–291.
- Filipponi, E., Cruciani, G., Tabarrini, O., Cecchetti, V. and Fravolini, A. (2001) QSAR study and VolSurf characterization of anti-HIV quinolone library. *J. Comput. Aid. Mol. Des.*, **15**, 203–217.
- Finizio, A., Sicbaldi, F. and Vighi, M. (1995) Evaluation of molecular connectivity indices as a predictive method of $\log K_{ow}$ for different classes of chemicals. *SAR & QSAR Environ. Res.*, **3**, 71–80.
- Finizio, A., Vighi, M. and Sandroni, D. (1997) Determination of n -octanol/water partition coefficient (K_{ow}) of pesticide critical review and comparison of methods. *Chemosphere*, **34**, 131–161.
- Fisanick, W., Cross, K.P., Forman, J.C. and Rusinko, A. III (1993) Experimental system for similarity and 3D searching of CAS registry substances. 1. 3D substructure searching. *J. Chem. Inf. Comput. Sci.*, **33**, 548–559.
- Fischer, H., Gottschlich, R. and Seelig, A. (1998) Blood–brain barrier permeation: molecular parameters governing passive diffusion. *J. Membr. Biol.*, **165**, 201–211.
- Fischer, H., Kansy, M., Potthast, M. and Csato, M. (2001) Prediction of *in vitro* phospholipidosis of drugs by means of their amphiphilic properties, in *Rational Approaches in Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 286–292.
- Fischer, J.R. and Rarey, M. (2007) SwiFT: an index structure for reduced graph descriptors in virtual screening and clustering. *J. Chem. Inf. Model.*, **47**, 1341–1353.
- Fisher, S.W., Lydy, M.J., Barger, J. and Landrum, P.F. (1993) Quantitative structure–activity relationships for predicting the toxicity of pesticides in aquatic systems with sediment. *Environ. Toxicol. Chem.*, **12**, 1307–1318.
- Fitch, W.L., McGregor, M., Katritzky, A.R., Lomaka, A., Petrukhin, R. and Karelson, M. (2002) Prediction of ultraviolet spectral absorbance using quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.*, **42**, 830–840.
- Flapan, E. (1995) Intrinsic chirality. *J. Mol. Struct. (Theochem)*, **336**, 157–164.
- Fleischer, R., Froberg, P., Büge, A., Nuhn, P. and Wiese, M. (2000) QSAR analysis of substituted 2-phenylhydrazonoacetamides acting as inhibitors of 15-lipoxygenase. *Quant. Struct. -Act. Relat.*, **19**, 162–172.
- Fleming, I. (1990) *Frontier Orbitals and Organic Chemical Reactions*, John Wiley & Sons, Inc., New York.
- Fligner, M.A., Verducci, J.S. and Blower, P.E. (2002) A modification of the Jaccard–Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics*, **44**, 110–119.
- Floersheim, P., Nozulak, J. and Weber, H.P. (1993) Experience with comparative molecular field

- analysis, in *Trends in QSAR and Molecular Modelling* 92 (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 227–232.
- Flory, P.J. (1969) *Statistical Mechanics of Chain Molecules*, Wiley-Interscience, New York.
- Flower, D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.*, **38**, 379–386.
- Folkers, G., Merz, A. and Rognan, D. (1993a) CoMFA as a tool for active site modelling, in *Trends in QSAR and Molecular Modelling* 92 (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 233–244.
- Folkers, G., Merz, A. and Rognan, D. (1993b) CoMFA: scope and limitations, in *3D QSAR in Drug Design: Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 583–618.
- Fomenko, A., Filimonov, D., Sobolev, B. and Poroikov, V.V. (2006) Prediction of protein functional specificity without an alignment. *OMICS*, **10**, 56–65.
- Fontaine, F., Pastor, M., Gutiérrez de Terán, H., Lozano, J.J. and Sanz, F. (2003) Use of alignment-free molecular descriptors in diversity analysis and optimal sampling of molecular libraries. *Mol. Div.*, **6**, 135–147.
- Fontaine, F., Pastor, M. and Sanz, F. (2004) Incorporating molecular shape into the alignment-free grid-independent descriptors. *J. Med. Chem.*, **47**, 2805–2815.
- Ford, G.P., Katritzky, A.R. and Topsom, R.D. (1978) Substituents effects in olefinic systems, in *Correlation Analysis in Chemistry* (eds N.B. Chapman and J. Shorter), Plenum Press, New York, pp. 269–311.
- Ford, M., Phillips, L. and Stevens, A. (2004) Optimising the EVA descriptor for prediction of biological activity. *Org. Biomol. Chem.*, **2**, 3301–3311.
- Ford, M.G., Greenwood, R., Brooks, G.T. and Franke, R. (1996) *Bioactive Compound Design: Possibilities for Industrial Use*, Bios Scientific Publishers, Portsmouth, UK, p. 186.
- Forina, M., Boggia, R., Mosti, L. and Fossa, P. (1997) Zupan's descriptors in QSAR applied to the study of a new class of cardiotonic agents. *Il Farmaco*, **52**, 411–419.
- Forsyth, D.A. (1973) Semiempirical models for substituent effects in electrophilic aromatic substitution and side-chain reactions. *J. Am. Chem. Soc.*, **95**, 3594–3603.
- Fossa, P., Menozzi, G. and Mosti, L. (2001) An updated topographical model for phosphodiesterase 4 (PDE4) catalytic site. *Quant. Struct. -Act. Relat.*, **20**, 17–22.
- Foster, J.P. and Weinhold, F. (1980) Natural hybrid orbitals. *J. Am. Chem. Soc.*, **102**, 7211–7218.
- Foster, R., Hyde, R.M. and Livingstone, D.J. (1978) Substituent constant for drug design studies based on properties of organic electron donor–acceptor complexes. *J. Pharm. Sci.*, **67**, 1310–1313.
- Fowler, P.W. (2002) Resistance distances in fullerene graphs. *Croat. Chem. Acta*, **75**, 401–408.
- Fradera, X., Amat, L., Besalú, E. and Carbó, R. (1997) Application of molecular quantum similarity to QSAR. *Quant. Struct. -Act. Relat.*, **16**, 25–32.
- Frank, I.E. (1987) Intermediate least squares regression method. *Chemom. Intell. Lab. Syst.*, **1**, 233–242.
- Frank, I.E. and Friedman, J.H. (1989) Classification: oldtimers and newcomers. *J. Chemom.*, **3**, 463–475.
- Frank, I.E. and Friedman, J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–135.
- Frank, I.E. and Todeschini, R. (1994) *The Data Analysis Handbook*, Elsevier, Amsterdam, The Netherlands, p. 366.
- Franke, L., Schwarz, O., Müller-Kuhrt, L., Hoernig, C., Fisher, L., George, S., Tanrikulu, Y., Schneider, P., Werz, O., Steinhilber, D. and Schneider, G. (2007) Identification of natural-product-derived inhibitors of 5-lipoxygenase activity by ligand-based virtual screening. *J. Med. Chem.*, **50**, 2640–2646.
- Franke, R. (1984) *Theoretical Drug Design Methods*, Elsevier, Amsterdam, The Netherlands.
- Franke, R. and Buschauer, A. (1992) Quantitative structure–activity relationships in histamine H₂-agonists related to impromidine and arpromidine. *Eur. J. Med. Chem.*, **27**, 443–448.
- Franke, R. and Buschauer, A. (1993) Interaction terms in Free–Wilson analysis: a QSAR of histamine H₂-agonists, in *Trends in QSAR and Molecular Modelling* 92 (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 160–162.
- Franke, R., Gruska, A. and Presber, W. (1994) Combined factor and QSAR analysis for antibacterial and pharmacokinetic data from parallel biological measurements. *Pharmazie*, **49**, 600–605.
- Franke, R., Hübel, S. and Streich, W.J. (1985) Substructural QSAR approaches and topological pharmacophores. *Environ. Health Persp.*, **61**, 239–255.
- Franke, R., Rose, S.V., Hyde, R.M. and Gruska, A. (1995) The use of indicator variables in QSARs of chiral compounds, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and

- F. Manaut), Prous Science, Barcelona, Spain, pp. 113–116.
- Franke, R. and Streich, W.J. (1985a) Topological pharmacophores, new methods and their application to a set of antimalarials. Part 2. Results from LOGANA. *Quant. Struct. -Act. Relat.*, **4**, 51–63.
- Franke, R. and Streich, W.J. (1985b) Topological pharmacophores, new methods and their application to a set of antimalarials. Part 3. Results from LOCON. *Quant. Struct. -Act. Relat.*, **4**, 63–69.
- Franot, C., Roberts, D.W., Basketter, D.A., Benezra, C. and Lepoittevin, J.-P. (1994) Structure–activity relationships for contact allergenic potential of γ,γ -dimethyl- γ -butyrolactone derivatives. 2. Quantitative structure–skin sensitization relationships for α -substituted- α -methyl- γ,γ -dimethyl- γ -butyrolactones. *Chem. Res. Toxicol.*, **7**, 307–312.
- Frater, F., Bonchev, D. and Enchev, V. (1980) A theoretical information approach to ring and total aromaticity in ground and excited states. *Croat. Chem. Acta*, **53**, 545–554.
- Free, S.M. and Wilson, J.W. (1964) A mathematical contribution to structure–activity studies. *J. Med. Chem.*, **7**, 395–399.
- Freeland, R.G., Funk, S.A., O’Korn, L.J. and Wilson, G.A. (1979) The chemical abstract service chemical registry system. II. Augmented connectivity molecular formula. *J. Chem. Inf. Comput. Sci.*, **19**, 94–98.
- Freeman, L.C. (1977) A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–41.
- Freeman, L.C. (1979) Centrality in social networks: conceptual clarification. *Social Networks*, **1**, 215–239.
- Freidig, A.P. and Hermens, J.L.M. (2000) Narcosis and chemical reactivity QSARs for acute fish toxicity. *Quant. Struct. -Act. Relat.*, **19**, 547–553.
- Frèrejacque, M. (1939) Condensation d’une molécule organique. *Bull. Soc. Chim. Fran. (French)*, **6**, 1008–1111.
- Friedman, J.H. (1988) Multivariate adaptive regression splines. Report, Laboratory of Computational Statistics, Department of Statistics, Stanford, CA.
- Frierson, M.R., Klopman, G. and Rosenkranz, H.S. (2006) Structure–activity relationships (SARs) among mutagens and carcinogens: a review. *Envir. Mutag.*, **8**, 283–327.
- Frimurer, T.M., Bywater, R., Nærum, L., Lauritsen, L. N. and Brunak, S. (2000) Improving the odds in discriminating “drug-like” from “non drug-like” compounds. *J. Chem. Inf. Comput. Sci.*, **40**, 1315–1324.
- Fröhlich, H., Wegner, J.K., Sieker, F. and Zell, A. (2006) Kernel functions for attributed molecular graphs: a new similarity-based approach to ADME prediction in classification and regression. *QSAR Comb. Sci.*, **25**, 317–326.
- Fuchs, R., Abraham, M.H., Kamlet, M.J. and Taft, R. W. (1982) *J. Phys. Org. Chem.*, **2**, 559.
- Fujita, S. (1988) Logical perception of ring-opening, ring-closure, and rearrangement reactions based on imaginary transition structures, selection of the essential set of essential rings (ESER). *J. Chem. Inf. Comput. Sci.*, **28**, 1–9.
- Fujita, S. (1996) The sphericity concept for an orbit of bonds, formulation of chirogenic sites in a homospheric orbit and of bond-differentiating chiral reactions with applications to C_{60} -adducts. *J. Chem. Inf. Comput. Sci.*, **36**, 270–285.
- Fujita, T. (1978) Steric effects in quantitative structure–activity relationships. *Prot. Struct. Funct. Gen.*, **50**, 987–994.
- Fujita, T. (1981) The *ortho* effect in quantitative structure–activity correlations. *Anal. Chim. Acta*, **133**, 667–676.
- Fujita, T. (1983) Substituent effects in the partition coefficient of disubstituted benzenes: bidirectional Hammett-type relationships. *Prog. Phys. Org. Chem.*, **14**, 75–113.
- Fujita, T. (1990) The extrathermodynamic approach to drug design, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 497–560.
- Fujita, T. (1995) Quantitative structure–activity analysis and database aided bioisosteric structural transformation procedure as methodologies of agrochemical design. *ACS Symp. Ser.*, **606**, 13–34.
- Fujita, T. (1997) Recent success stories leading to commercializable bioactive compounds with the aid of traditional QSAR procedures. *Quant. Struct. -Act. Relat.*, **16**, 107–112.
- Fujita, T. and Ban, T. (1971) Structure–activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. *J. Med. Chem.*, **14**, 148–152.
- Fujita, T. and Iwamura, H. (1983) Applications of various steric constants to quantitative analysis of structure–activity relationships, in *Steric Effects in Drug Design*, Topics in Current Chemistry, Vol. 114 (eds M. Charton and I. Motoc), Springer-Verlag, Berlin, Germany, pp. 119–157.
- Fujita, T., Iwasa, J. and Hansch, C. (1964) A new substituent constant, π , derived from partition coefficients. *J. Am. Chem. Soc.*, **86**, 5175–5180.
- Fujita, T. and Nishioka, T. (1976) The analysis of the *ortho* effect. *Prog. Phys. Org. Chem.*, **12**, 49–89.

- Fujita, T., Nishioka, T. and Nakajima, M. (1977) Hydrogen-bonding parameter and its significance in quantitative structure–activity studies. *J. Med. Chem.*, **20**, 1071–1081.
- Fujita, T., Takayama, C. and Nakajima, M. (1973) The nature and composition of Taft–Hancock steric constants. *J. Org. Chem.*, **38**, 1623–1630.
- Fujiwara, H., Da, Y.-Z. and Ito, K. (1991) The energy aspect of oil water partition and its application to the analysis of quantitative structure–activity relationships of aliphatic alcohols in the liposome water partition system. *Bull. Chem. Soc. Jap.*, **64**, 3707–3712.
- Fukui, K. (1982) Role of frontier orbitals in chemical reactions. *Science*, **218**, 747–754.
- Fukui, K., Yonezawa, Y. and Shingu, H. (1954) Theory of substitution in conjugated molecules. *Bull. Chem. Soc. Jap.*, **27**, 423–427.
- Fukunaga, J.Y., Hansch, C. and Steller, E.E. (1976) Inhibition of dihydrofolate reductase: structure–activity correlations of quinazolines. *J. Med. Chem.*, **19**, 605–611.
- Fuller, F.B. (1971) The writhing number of a space curve. *Proc. Nat. Acad. Sci. USA*, **68**, 815.
- Funar-Timofei, S., Kurunczi, L. and Iliescu, S. (2005) Structure–property study of some phosphorus-containing polymers by computational methods. *Polym. Bull.*, **54**, 443–449.
- Funar-Timofei, S. and Schüürmann, G. (2002) Comparative molecular field analysis (CoMFA) of anionic azo dye–fiber affinities. I. Gas-phase molecular orbital descriptors. *J. Chem. Inf. Comput. Sci.*, **42**, 788–795.
- Funar-Timofei, S., Suzuki, T., Paier, J.A., Steinreiber, A., Faber, K. and Fabian, W.M.F. (2003) Quantitative structure–activity relationships for the enantioselectivity of oxirane ring-opening catalyzed by epoxide hydrolases. *J. Chem. Inf. Comput. Sci.*, **43**, 934–940.
- Furet, P., Sele, A. and Cohen, N.C. (1988) 3D molecular lipophilicity potential profiles: a new tool in molecular modeling. *J. Mol. Graph.*, **6**, 182–189.
- Furtula, B., Gutman, I., Tomović, Ž., Vesel, A. and Pesek, I. (2002) Wiener-type topological indices of phenylenes. *Indian J. Chem.*, **41**, 1767–1772.
- Gabanyi, Z., Surjan, P.R. and Naray-Szabo, G. (1982) Application of topological molecular transforms to rational drug design. *Eur. J. Med. Chem.*, **17**, 307–311.
- Gago, F., Pastor, M., Perez-Butragueño, J., Lopez, R., Alvarez-Builla, J. and Elguero, J. (1994) Hydrophobicity of heterocycles determination of the π values of substituents on *N*-phenylpyrazoles. *Quant. Struct. -Act. Relat.*, **13**, 165–171.
- Gaillard, P., Carrupt, P.-A. and Testa, B. (1994a) The conformation-dependent lipophilicity of morphine glucuronides as calculated from their molecular lipophilicity potential. *Bioorg. Med. Chem. Lett.*, **4**, 737–742.
- Gaillard, P., Carrupt, P.-A., Testa, B. and Boudon, A. (1994b) Molecular lipophilicity potential, a tool in 3D QSAR: method and applications. *J. Comput. Aid. Mol. Des.*, **8**, 83–96.
- Gakh, A.A. and Burnett, M.N. (2001) Modular chemical descriptor language (MCDL): composition, connectivity, and supplementary modules. *J. Chem. Inf. Comput. Sci.*, **41**, 1494–1499.
- Gakh, A.A., Gakh, E.G., Sumpter, B.G. and Noid, D. W. (1994) Neural network-graph theory approach to the prediction of the physical properties of organic compounds. *J. Chem. Inf. Comput. Sci.*, **34**, 832–839.
- Galanakis, D., Calder, J.A., Ganellin, C.R., Owen, C.S. and Dunn, P.M. (1995) Synthesis and quantitative structure–activity relationships of dequalinium analogs as K^+ channel blockers investigation into the role of the substituent at position 4 of the quinoline ring. *J. Med. Chem.*, **38**, 3536–3546.
- Gallegos Saliner, A. and Gironés, X. (2005) Topological quantum similarity measures: applications in QSAR. *J. Mol. Struct. (Theochem)*, **727**, 97–106.
- Gallegos Saliner, A., Patlewicz, G. and Worth, A.P. (2008) A review of (Q)SAR models for skin and eye irritation and corrosion. *QSAR Comb. Sci.*, **27**, 49–59.
- Gallegos, A., Robert, D., Gironés, X. and Carbó-Dorca, R. (2001) Structure–toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J. Comput. Aid. Mol. Des.*, **15**, 67–80.
- Gallo, R. (1983) Treatment of steric effects. *Prog. Phys. Org. Chem.*, **14**, 115–163.
- Gálvez, J. (1998) On a topological interpretation of electronic and vibrational energies. *J. Mol. Struct. (Theochem)*, **429**, 255–264.
- Gálvez, J. (2003) Prediction of molecular volume and surface of alkanes by molecular topology. *J. Chem. Inf. Comput. Sci.*, **43**, 1231–1239.
- Gálvez, J., De Julián-Ortiz, V. and García-Domenech, R. (2005) Application of molecular topology to the prediction of potency and selection of novel insecticides active against malaria vectors. *J. Mol. Struct. (Theochem)*, **727**, 107–113.
- Gálvez, J., García, A.E., De Julián-Ortiz, V. and Soler, R. (1995) A topological approach to drug design, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*

- (eds F. Sanz J. Giraldo and F. Manaut), Prous Science, Barcelona, Spain, pp. 163–166.
- Gálvez, J., García, R., Salabert, M.T. and Soler, R. (1994) Charge indexes: new topological descriptors. *J. Chem. Inf. Comput. Sci.*, **34**, 520–525.
- Gálvez, J., García-Domenech, R. and de Gregorio Alapont, C. (2000) Indices of differences of path lengths: novel topological descriptors derived from electronic interferences in graphs. *J. Comput. Aid. Mol. Des.*, **14**, 679–687.
- Gálvez, J., García-Domenech, R., de Gregorio Alapont, C., De Julián-Ortiz, V. and Popa, L. (1996) Pharmacological distribution diagrams: a tool for *de novo* drug design. *J. Mol. Graph.*, **14**, 272–276.
- Gálvez, J., García-Domenech, R. and De Julián-Ortiz, V. (2006) Assigning wave functions to graphs: a way to introduce novel topological indices. *MATCH Commun. Math. Comput. Chem.*, **56**, 509–518.
- Gálvez, J., García-Domenech, R., De Julián-Ortiz, V. and Soler, R. (1994) Topological approach to analgesia. *J. Chem. Inf. Comput. Sci.*, **34**, 1198–1203.
- Gálvez, J., García-Domenech, R., De Julián-Ortiz, V. and Soler, R. (1995) Topological approach to drug design. *J. Chem. Inf. Comput. Sci.*, **35**, 272–284.
- Gálvez, J., Gomez-Lechón, M.J., García-Domenech, R. and Castell, J.V. (1996) New cytostatic agents obtained by molecular topology. *Bioorg. Med. Chem. Lett.*, **6**, 2301–2306.
- Gálvez, J., Julian-Ortiz, J.V. and García-Domenech, R. (2001) General topological patterns of known drugs. *J. Mol. Graph. Model.*, **20**, 84–94.
- GAMESS, Mark Gordon's Quantum Theory Group, Iowa State University, IO.
- Gamper, A.M., Winger, R.H., Liedl, K.R., Sottriffer, C. A., Varga, J.M., Kroemer, R.T. and Rode, B.M. (1996) Comparative molecular field analysis of haptens docked to the multispecific antibody IgE (Lb4). *J. Med. Chem.*, **39**, 3882–3888.
- Gan, H.H., Pasquali, S. and Schlick, T. (2003) Exploring the repertoire of RNA secondary motifs using graph theory: implications for RNA design. *Nucleic Acids Res.*, **31**, 2926–2943.
- Gancia, E., Bravi, G., Mascagni, P. and Zaliani, A. (2000) Global 3D-QSAR methods: MS-WHIM and autocorrelation. *J. Comput. Aid. Mol. Des.*, **14**, 293–306.
- Gancia, E., Montana, J.G. and Manallack, D.T. (2001) Theoretical hydrogen bonding parameters for drug design. *J. Mol. Graph. Model.*, **19**, 349–362.
- Gangal, R. (2002) DivCalc: a utility for diversity analysis and compound sampling. *Molecules*, **7**, 657–661.
- Gani, R., Harper, P.M. and Hostrup, M. (2005) Automatic creation of missing groups through connectivity index for pure-component property prediction. *Ind. Eng. Chem. Res.*, **44**, 7262–7269.
- Gantchev, T.G., Ali, H. and Vanlier, J.E. (1994) Quantitative structure–activity relationships comparative molecular field analysis (QSAR/CoMFA) for receptor binding properties of halogenated estradiol derivatives. *J. Med. Chem.*, **37**, 4164–4176.
- Gao, C., Govind, R. and Tabak, H.H. (1992) Application of the group contribution method for predicting the toxicity of organic chemicals. *Environ. Toxicol. Chem.*, **11**, 631–636.
- Gao, H. (2001) Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *J. Chem. Inf. Comput. Sci.*, **41**, 402–407.
- Gao, H. and Bajorath, J. (1999) Comparison of binary and 2D QSAR analyses using inhibitors of human carbonic anhydrase II as a test case. *Mol. Div.*, **4**, 115–130.
- Gao, H., Lajiness, M.S. and Van Drie, J. (2002) Enhancement of binary QSAR analysis by a GA-based variable selection method. *J. Mol. Graph. Model.*, **20**, 259–268.
- Gao, H., Williams, C., Labute, P. and Bajorath, J. (1999) Binary quantitative structure–activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.*, **39**, 164–168.
- Gao, J. and Zhang, X. (2006a) Analysis of RNA secondary structures based on 4D representation. *MATCH Commun. Math. Comput. Chem.*, **56**, 249–259.
- Gao, J. and Zhang, X. (2006b) Similarity analysis of RNA secondary structures based on 4D representation. *MATCH Commun. Math. Comput. Chem.*, **56**, 249–259.
- Gao, Y. and Hosoya, H. (1988) Topological index and thermodynamics properties. IV. Size dependency of the structure–activity correlation of alkanes. *Bull. Chem. Soc. Jap.*, **61**, 3093–3102.
- Garcia, E. (2002) QCODES: fast topological descriptors for macromolecules. *J. Chem. Inf. Comput. Sci.*, **42**, 1370–1377.
- Garcia, E., Lopezdecerain, A., Martinez Merino, V. and Monge, A. (1992) Quantitative structure mutagenic activity relationships of triazino indole derivatives. *Mut. Res.*, **268**, 1–9.
- García, G.C., Ruiz, I.L. and Gómez-Nieto, M.A. (2002) Cyclical conjunction: an efficient operator for the extraction of cycles from a graph. *J. Chem. Inf. Comput. Sci.*, **42**, 1415–1424.
- García-Domenech, R., de Gregorio Alapont, C., De Julián-Ortiz, V., Gálvez, J. and Popa, L. (1997) Molecular connectivity to find β -blockers with low toxicity. *Bioorg. Med. Chem. Lett.*, **7**, 567–572.

- García-Domenech, R. and De Julián-Ortiz, V. (1998) Antimicrobial activity characterization in a heterogeneous group of compounds. *J. Chem. Inf. Comput. Sci.*, **38**, 445–449.
- García-Domenech, R., Gálvez, J., De Julián-Ortiz, V. and Pogliani, L. (2008) Some new trends in chemical graph theory. *Chem. Rev.*, **108**, 1127–1169.
- García-Domenech, R., Gálvez, J., Moliner, R. and García-March, F.J. (1991) Prediction and interpretation of some pharmacological properties of cephalosporins using molecular connectivity. *Drug Invest.*, **3**, 344–350.
- García-Domenech, R., García-March, F.J., Soler, R., Gálvez, J., Antón-Fos, G.M. and De Julián-Ortiz, V. (1996) New analgesics designed by molecular topology. *Quant. Struct. -Act. Relat.*, **15**, 201–207.
- García-Domenech, R., Rios-Santamarina, I., Catalá, A., Calabuig, C., del Castillo, L. and Gálvez, J. (2003) Application of molecular topology to the prediction of antifungal activity for a set of dication-substituted carbazoles, furans and benzimidazoles. *J. Mol. Struct. (Theochem)*, **624**, 97–107.
- García-March, F.J., Antón-Fos, G.M., Perez-Gimenez, F., Salabert-Salvador, M.T., Cercos-del-Pozo, R.A. and De Julián-Ortiz, V. (1996) Prediction of chromatographic properties for a group of natural phenolic derivatives by molecular topology. *J. Chromat.*, **719**, 45–51.
- Gardiner, E.J., Gillet, V.J., Willett, P. and Cosgrove, D. A. (2007) Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *J. Chem. Inf. Model.*, **47**, 354–366.
- Gardner, R.J. (1980) Correlation of bitterness thresholds of amino acids and peptides with molecular connectivity. *J. Sci. Food Agric.*, **31**, 23.
- Garg, S. and Achenie, L.E.K. (2001) Mathematical programming assisted drug design for nonclassical antifolates. *Biotechnol. Prog.*, **17**, 412–418.
- Gargallo, R., Sottriffer, C.A., Liedl, K.R. and Rode, B. M. (1999) Application of multivariate data analysis methods to comparative molecular field analysis (CoMFA) data: proton affinities and pK_a prediction for nucleic acids components. *J. Comput. Aid. Mol. Des.*, **13**, 611–623.
- Gargas, M.L. and Seybold, P.G. (1988) Modeling the tissue solubilities and metabolic rate constant (V_{max}) of halogenated methanes, ethanes, and ethylenes. *Toxicol. Lett.*, **43**, 235–256.
- Garkani-Nejad, Z., Karlovits, M., Demuth, W., Stimpfl, T., Vycudilik, W., Jalali-Heravi, M. and Varmuza, K. (2004) Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds. *J. Chromat.*, **1028**, 287–295.
- Garrone, A., Marengo, E., Fornatto, E. and Gasco, A. (1992) A study on pK_a (app) and partition coefficient of substituted benzoic acids in SDS anionic micellar system. *Quant. Struct. -Act. Relat.*, **11**, 171–175.
- Gasteiger, J. (1979) A representation of π systems for efficient computer manipulation. *J. Chem. Inf. Comput. Sci.*, **19**, 111–115.
- Gasteiger, J. (1988) Empirical methods for the calculation of physico-chemical data of organic compounds, in *Physical Property Prediction in Organic Chemistry* (eds C. Jochum, M.G. Hicks and J. Sunkel), Springer-Verlag, Berlin, Germany, pp. 119–138.
- Gasteiger, J. (ed.) *Software Development in Chemistry*, Vol. 10, Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main, Germany, pp. 434.
- Gasteiger, J. (1998) Making the computer understand chemistry. *Internet J. Chem.*, **1**, article 33.
- Gasteiger, J. (2001) Data mining in drug design, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 459–474.
- Gasteiger, J. (2003a) A hierarchy of structure representations, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1034–1061.
- Gasteiger, J. (ed.) (2003b) *Handbook of Chemoinformatics: From Data to Knowledge*, 4 Vols, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 1870.
- Gasteiger, J. (2003c) Physico-chemical effects in the representation of molecular structures for drug designing. *Mini Rev. Med. Chem.*, **3**, 789–796.
- Gasteiger, J. (2006) The central role of chemoinformatics. *Chemom. Intell. Lab. Syst.*, **82**, 200–209.
- Gasteiger, J. and Engel, T. (eds) (2003) *Chemoinformatics: A Textbook*, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 650.
- Gasteiger, J., Hutchings, M.G., Christoph, B., Gann, L., Hiller, C., Löw, P., Marsili, M., Saller, H. and Yuki, K. (1987) A new treatment of chemical reactivity: development of EROS, an expert system for reaction prediction and synthesis design. *Top. Curr. Chem.*, **137**, 19–73.
- Gasteiger, J. and Jochum, C. (1979) An algorithm for the perception of synthetically important rings. *J. Chem. Inf. Comput. Sci.*, **19**, 244–255.
- Gasteiger, J. and Li, X. (1994) Representation of the electrostatic potentials of muscarinic and nicotinic

- agonists with artificial neuronal nets. *Angew. Chem. Int. Ed. Engl.*, **33**, 643–646.
- Gasteiger, J., Li, X., Rudolph, C.J., Sadowski, J. and Zupan, J. (1994a) Representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.*, **116**, 4608–4620.
- Gasteiger, J., Li, X. and Uschold, A. (1994b) The beauty of molecular surfaces as revealed by self-organizing neural networks. *J. Mol. Graph.*, **12**, 90–97.
- Gasteiger, J. and Marsili, M. (1980) Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron*, **36**, 3219–3228.
- Gasteiger, J., Röse, P. and Saller, H. (1988) Multidimensional explorations into chemical reactivity: the reactivity space. *J. Mol. Graph.*, **6**, 87–92.
- Gasteiger, J., Sadowski, J., Schuur, J., Selzer, P., Steinhauer, L. and Steinhauer, V. (1996) Chemical information in 3D space. *J. Chem. Inf. Comput. Sci.*, **36**, 1030–1037.
- Gasteiger, J., Schuur, J., Selzer, P., Steinhauer, L. and Steinhauer, V. (1997) Finding the 3D structure of a molecule in its IR spectrum. *Fresen. J. Anal. Chem.*, **359**, 50–55.
- Gates, M.A. (1985) Simple DNA sequence representations. *Nature*, **316**, 219.
- Gaudio, A.C. and Montanari, C.A. (2002) HEPT derivatives as nonnucleoside inhibitors of HIV-1 reverse transcriptase: QSAR studies agree with the crystal structures. *J. Comput. Aid. Mol. Des.*, **16**, 287–295.
- GAUSSIAN03, Gaussian, Inc., Pittsburgh, PA.
- Gautzsch, R. and Zinn, P. (1992a) List operations on chemical graphs. 1. Basic list structures and operations. *J. Chem. Inf. Comput. Sci.*, **32**, 541–550.
- Gautzsch, R. and Zinn, P. (1992b) List operations on chemical graphs. 2. Combining basic list operations. *J. Chem. Inf. Comput. Sci.*, **32**, 551–555.
- Gautzsch, R. and Zinn, P. (1994) List operations on chemical graphs. 5. Implementation of breadth-first molecular path generation and application in the estimation of retention index data and boiling points. *J. Chem. Inf. Comput. Sci.*, **34**, 791–800.
- Gautzsch, R. and Zinn, P. (1996) Use of incremental models to estimate the retention indexes of aromatic compounds. *Chromatographia*, **43**, 163–176.
- Gavezzotti, A. (1983) The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and solid-state organic reactivity. *J. Am. Chem. Soc.*, **105**, 5220–5225.
- Gawlik, B.M., Sotiriou, N., Feicht, E.A., Schulte-Hostede, S. and Kettrup, A. (1997) Alternatives for the determination of the soil adsorption coefficient, k_{oc} , of non-ionic organic compounds: a review. *Chemosphere*, **34**, 2525–2551.
- Gayoso, J. and Kimri, S. (1990a) Tentative unification of quantum theories of polyacene carcinogenesis. I. Theory of the M, L, and B regions. *Int. J. Quant. Chem.*, **38**, 461–486.
- Gayoso, J. and Kimri, S. (1990b) Tentative unification of quantum theories of polyacene carcinogenesis. II. Role of the K region in the metabolic activation process leading to the ultimate carcinogen. Theory of the M, L, and BK regions. *Int. J. Quant. Chem.*, **38**, 487–495.
- Geach, J., Walters, C.J., James, B., Caviness, K.E. and Hefferlin, R.A. (2002) Global molecular identification from graphs. Main-group triatomic molecules. *Croat. Chem. Acta*, **75**, 383–400.
- Geary, R.C. (1954) The contiguity ratio and statistical mapping. *Incorp. Statist.*, **5**, 115–145.
- Gedeck, P., Rohde, B. and Bartels, C. (2006) QSAR: how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.*, **46**, 1924–1936.
- Geerlings, P., De Proft, F. and Langenaeker, W. (2003) Conceptual density functional theory. *Chem. Rev.*, **103**, 1793–1873.
- Geerlings, P., De Proft, F. and Martin, J.M.L. (1996) Density-functional theory concepts and techniques for studying molecular charge distributions and related properties, in *Recent Developments and Applications of Modern Density Functional Theory*, Vol. 4 (ed. J.M. Seminario), Elsevier, Amsterdam, The Netherlands, pp. 773–809.
- Geerlings, P., Langenaeker, W., De Proft, F. and Baeten, A. (1996) Molecular electrostatic potentials vs. DFT descriptors of reactivity, in *Molecular Electrostatic Potentials: Concepts and Applications*, Vol. 3 (eds. J.S. Murray and K. Sen), Elsevier, Amsterdam, The Netherlands, pp. 587–617.
- Geladi, P., Berglund, A. and Wold, S. (1997) Comments on “multivariate prediction for QSAR” by Heinz Schmidli. *Chemom. Intell. Lab. Syst.*, **37**, 135–137.
- Geladi, P. and Kowalski, B.R. (1986) Partial least squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1–17.
- Geladi, P. and Tosato, M.L. (1990) Multivariate latent variable projection methods: SIMCA and PLS, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 145–152.

- Gelius, U. (1974) Binding energies and chemical shifts in ESCA. *Phys. Scripta*, **9**, 133–147.
- Gerstl, Z. and Helling, C.S. (1987) Evaluation of molecular connectivity as a predictive method for the adsorption of pesticides by soils. *J. Environ. Sci. Health*, **22**, 55–69.
- Ghafourian, T. and Cronin, M.T.D. (2004) Comparison of electrotopological-state indices versus atomic charge and superdelocalisability indices in a QSAR study of the receptor binding properties of halogenated estradiol derivatives. *Mol. Div.*, **8**, 343–355.
- Ghose, A.K. and Crippen, G.M. (1982) Quantitative structure–activity relationship by distance geometry: quinazolines as dihydrofolate reductase inhibitors. *J. Med. Chem.*, **25**, 892–899.
- Ghose, A.K. and Crippen, G.M. (1983) Combined distance geometry analysis of dihydrofolate reductase inhibition by quinazolines and triazines. *J. Med. Chem.*, **26**, 996–1010.
- Ghose, A.K. and Crippen, G.M. (1984) General distance geometry three-dimensional receptor model for diverse dihydrofolate reductase inhibitors. *J. Med. Chem.*, **27**, 901–914.
- Ghose, A.K. and Crippen, G.M. (1985a) Geometrically feasible binding modes of a flexible ligand molecule at the receptor site. *J. Comput. Chem.*, **6**, 350–359.
- Ghose, A.K. and Crippen, G.M. (1985b) Use of physico-chemical parameters in distance geometry and related three-dimensional quantitative structure–activity relationships: a demonstration using *Escherichia coli* dihydrofolate reductase inhibitors. *J. Med. Chem.*, **28**, 333–346.
- Ghose, A.K. and Crippen, G.M. (1986) Atomic physico-chemical parameters for three-dimensional-structure-directed quantitative structure–activity relationships. 1. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.*, **7**, 565–577.
- Ghose, A.K. and Crippen, G.M. (1987) Atomic physico-chemical parameters for three-dimensional-structure-directed quantitative structure–activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.*, **27**, 21–35.
- Ghose, A.K. and Crippen, G.M. (1990) The distance geometry approach to modeling receptor sites, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 715–733.
- Ghose, A.K., Logan, M.E., Treasurywala, A.M., Wang, H., Wahl, R.C., Tomczuk, B.E., Gowravaram, M.R., Jaeger, E.P. and Wendoloski, J.J. (1995) Determination of pharmacophoric geometry for collagenase inhibitors using a novel computational method and its verification using molecular dynamics, NMR, and X-ray crystallography. *J. Am. Chem. Soc.*, **117**, 4671–4682.
- Ghose, A.K., Pritchett, A. and Crippen, G.M. (1988) Atomic physico-chemical parameters for three dimensional structure directed quantitative structure–activity relationships iii: modeling hydrophobic interactions. *J. Comput. Chem.*, **9**, 80–90.
- Ghose, A.K., Viswanadhan, V.N. and Wendoloski, J.J. (1998) prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A*, **102**, 3762–3772.
- Ghose, A.K., Viswanadhan, V.N. and Wendoloski, J.J. (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.*, **1**, 55–68.
- Ghose, A.K. and Wendoloski, J.J. (1998) Pharmacophore modelling: methods, experimental verification and applications, in *3D QSAR in Drug Design*, Vol. 2 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 253–271.
- Ghoshal, N., Mukhopadhyay, S.N., Ghoshal, T.K. and Achari, B. (1993) Quantitative structure–activity relationship studies using artificial neural networks. *Indian J. Chem.*, **32**, 1045–1050.
- Ghuloum, A.M., Sage, C.R. and Jain, A.N. (1999) Molecular hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J. Med. Chem.*, **42**, 1739–1748.
- Gibson, S., McGuire, R. and Rees, D.C. (1996) Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments. *J. Med. Chem.*, **39**, 4065–4072.
- Gieleciak, R. and Polanski, J. (2007) Modeling robust QSAR. 2. Iterative variable elimination schemes for CoMSA: application for modeling benzoic acid pK_a values. *J. Chem. Inf. Model.*, **47**, 547–556.
- Giese, B. and Beckhaus, H.-D. (1978) Front strain of π and σ radicals. *Angew. Chem. Int. Ed. Engl.*, **17**, 594–595.
- Gifford, E.M., Johnson, M.A., Kaiser, D.G. and Tsai, C.-C. (1992) Visualizing relative occurrences in metabolic transformations of xenobiotics using structure–activity maps. *J. Chem. Inf. Comput. Sci.*, **32**, 591–599.

- Gilat, G. (1994) On quantifying chirality: obstacles and problems toward unification. *J. Math. Chem.*, **15**, 197–205.
- Gillet, V.J., Downs, G.M., Holliday, J.D., Lynch, M.F. and Dethlefsen, W. (1991) Computer storage and retrieval of generic chemical structures in patents. 13. Reduced graph generation. *J. Chem. Inf. Comput. Sci.*, **31**, 260–270.
- Gillet, V.J., Willett, P. and Bradshaw, J. (1998) Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.*, **38**, 165–179.
- Gillet, V.J., Willett, P. and Bradshaw, J. (2003) Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.*, **43**, 338–345.
- Gillet, V.J., Willett, P., Bradshaw, J. and Green, D.V.S. (1999) Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci.*, **39**, 169–177.
- Gilson, M.K., Gilson, H.S.R. and Potter, M.J. (2003) Fast assignment of accurate partial atomic charges: an electronegativity equalization method that accounts for alternate resonance forms. *J. Chem. Inf. Comput. Sci.*, **43**, 1982–1997.
- Gimarc, B.M. and Parr, R.G. (1965) The quantum theory of valence. *Annu. Rev. Phys. Chem.*, **16**, 451–480.
- Gineityte, V. (1998) Spectral meaning of coefficients within the adjacency matrix eigenfunctions of chemical graphs of alkanes. *Croat. Chem. Acta*, **71**, 673–688.
- Gini, C. (1909) Il diverso accrescimento delle classi sociali e la concentrazione della ricchezza. *Giornale degli Economisti (Italian)*, **11**, 37.
- Gini, G., Lorenzini, M., Benfenati, E., Grasso, P. and Bruschi, M. (1999) Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network. *J. Chem. Inf. Comput. Sci.*, **39**, 1076–1080.
- Gini, G., Testaguzza, V., Benfenati, E. and Todeschini, R. (1998) Hybrid toxicology expert system: architecture and implementation of a multi-domain hybrid expert system for toxicology. *Chemom. Intell. Lab. Syst.*, **43**, 135–145.
- Ginn, C.M., Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W. (1997) Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.*, **37**, 23–27.
- Ginn, C.M., Willett, P. and Bradshaw, J. (2000) Combination of molecular similarity measures using data fusion. *Persp. Drug Disc. Des.*, **20**, 1–16.
- Giordanetto, F., Fossa, P., Menozzi, G. and Mosti, L. (2003) *In silico* rationalization of the structural and physico-chemical requirements for photobiological activity in angelicine derivatives and their hetero-analogues. *J. Comput. Aid. Mol. Des.*, **17**, 53–64.
- Giraud, E., Luttmann, C., Lavelle, F., Riou, J.-F., Meilliet, P. and Laoui, A. (2000) Multivariate data analysis using D-optimal designs, partial least squares, and response surface modeling: a directional approach for the analysis of farnesyltransferase inhibitors. *J. Med. Chem.*, **43**, 1807–1816.
- Gironés, X., Amat, L. and Carbó-Dorca, R. (2002) Modeling large macromolecular structures using promolecular densities. *J. Chem. Inf. Comput. Sci.*, **42**, 847–852.
- Gironés, X., Amat, L., Robert, D. and Carbó-Dorca, R. (2000) Use of electron–electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput. Aid. Mol. Des.*, **14**, 477–485.
- Gironés, X. and Carbó-Dorca, R. (2002a) Molecular quantum similarity-based QSARs for binding affinities of several steroid sets. *J. Chem. Inf. Comput. Sci.*, **42**, 1185–1193.
- Gironés, X. and Carbó-Dorca, R. (2002b) Using molecular quantum similarity measures under stochastic transformation to describe physical properties of molecular systems. *J. Chem. Inf. Comput. Sci.*, **42**, 317–325.
- Gironés, X. and Carbó-Dorca, R. (2003) Molecular basis of LFER. Modeling of the electronic substituent effect using fragment quantum self-similarity measures. *J. Chem. Inf. Comput. Sci.*, **43**, 2033–2038.
- Gironés, X., Gallegos, A. and Carbó-Dorca, R. (2000) Modeling antimalarial activity: application of kinetic energy density quantum similarity measures as descriptors in QSAR. *J. Chem. Inf. Comput. Sci.*, **40**, 1400–1407.
- Gironés, X., Gallegos, A. and Carbó-Dorca, R. (2001) Antimalarial activity of synthetic 1,2,4-trioxanes and cyclic peroxy ketals, a quantum similarity study. *J. Comput. Aid. Mol. Des.*, **15**, 1053–1063.
- Gironés, X., Gallegos, A. and Carbó-Dorca, R. (2002) Modeling antimalarial activity: application of kinetic energy density quantum similarity measures as descriptors in QSAR. *J. Chem. Inf. Comput. Sci.*, **40**, 1400–1407.
- Giuffreda, M.G., Bruschi, M. and Lüthi, H.P. (2004) Electron delocalization in linearly π -conjugated systems: a concept for quantitative analysis. *Chem. Eur. J.*, **10**, 5671–5680.
- Givèhchi, A., Bender, A. and Glen, R.C. (2006) Analysis of activity space by fragment fingerprints,

- 2D descriptors, and multitarget dependent transformation of 2D descriptors. *J. Chem. Inf. Model.*, **46**, 1078–1083.
- Givehchi, A. and Schneider, G. (2004) Impact of descriptor vector scaling on the classification of drugs and nondrugs with artificial neural networks. *J. Mol. Model.*, **10**, 204–211.
- Givehchi, A. and Schneider, G. (2005) Multi-space classification for predicting GPCR-ligands. *Mol. Div.*, **9**, 371–383.
- Gladstone, J.H. and Dale, T.P. (1858) On the influence of temperature on the refraction of light. *Phil. Trans. Roy. Soc. (London)*, **148**, 887–894.
- Glen, R.C. and Adams, S.E. (2006) Similarity metrics and descriptor spaces: which combinations to choose? *QSAR Comb. Sci.*, **25**, 1133–1142.
- Glennon, R.A. and Kier, L.B. (1978) LSD analogs as serotonin antagonists: a molecular connectivity SAR analysis. *Eur. J. Med. Chem.*, **13**, 219.
- Glennon, R.A., Kier, L.B. and Shulgin, A.T. (1979) Molecular connectivity analysis of hallucinogenic mescaline analogs. *J. Pharm. Sci.*, **68**, 906.
- Glossman-Mitnik, D. (2005) G3-B3 calculation of the molecular structure and descriptors of isomeric thiadiazoles. *J. Mol. Struct.*, **725**, 27–30.
- Godden, J.W. and Bajorath, J. (2000) Shannon entropy: a novel concept in molecular descriptor and diversity analysis. *J. Mol. Graph. Model.*, **18**, 73–76.
- Godden, J.W. and Bajorath, J. (2001) Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.*, **41**, 1060–1066.
- Godden, J.W. and Bajorath, J. (2002) Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J. Chem. Inf. Comput. Sci.*, **42**, 87–93.
- Godden, J.W. and Bajorath, J. (2003) An information-theoretic approach to descriptor selection for database profiling and QSAR modeling. *QSAR Comb. Sci.*, **22**, 487–497.
- Godden, J.W., Furr, J.R. and Bajorath, J. (2003) Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci.*, **43**, 182–188.
- Godden, J.W., Furr, J.R., Xue, L., Stahura, F.L. and Bajorath, J. (2004) Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *J. Chem. Inf. Comput. Sci.*, **44**, 21–29.
- Godden, J.W., Stahura, F.L. and Bajorath, J. (2000) Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.*, **40**, 796–800.
- Godden, J.W., Xue, L. and Bajorath, J. (2000) Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients. *J. Chem. Inf. Comput. Sci.*, **40**, 163–166.
- Godden, J.W., Xue, L. and Bajorath, J. (2002a) Classification of biologically active compounds by median partitioning. *J. Chem. Inf. Comput. Sci.*, **42**, 1263–1269.
- Godden, J.W., Xue, L., Kitchen, D.B., Stahura, F.L., Schermerhorn, E.J. and Bajorath, J. (2002b) Median partitioning: a novel method for the selection of representative subsets from large compound pools. *J. Chem. Inf. Comput. Sci.*, **42**, 885–893.
- Godfrey, M. (1978) Theoretical models for interpreting linear correlations in organic chemistry, in *Correlation Analysis in Chemistry* (eds N.B. Chapman and J. Shorter.), Plenum Press, New York, pp. 85–117.
- Godha, K., Mori, I., Ohta, D. and Kikuchi, T. (2000) A CoMFA analysis with conformational propensity: an attempt to analyze the SAR of a set of molecules with different conformational flexibility using a 3D-QSAR method. *J. Comput. Aid. Mol. Des.*, **14**, 265–275.
- Godsil, C.D. and Gutman, I. (1999) Wiener index, graph spectrum, line graph. *Acta Chim. Hung. - Mod. Chem.*, **136**, 503–510.
- Goel, A. and Madan, A.K. (1995) Structure–activity study on antiinflammatory pyrazole carboxylic acid hydrazide analogs using molecular connectivity indices. *J. Chem. Inf. Comput. Sci.*, **35**, 510–514.
- Gola, J., Obrezanova, O., Champness, E. and Segall, M. (2006) ADMET property prediction: the state of the art and current challenges. *QSAR Comb. Sci.*, **25**, 1172–1180.
- Golberg, L. (ed.) (1983) *Structure–Activity Correlation as a Predictive Tool in Toxicology: Fundamentals, Methods, and Applications*, Hemisphere Publishing, Washington, DC.
- Golbraikh, A. (2000) Molecular dataset diversity indices and their applications to comparison of chemical databases and QSAR analysis. *J. Chem. Inf. Comput. Sci.*, **40**, 414–425.
- Golbraikh, A., Bonchev, D. and Tropsha, A. (2001a) Novel chirality descriptors derived from molecular topology. *J. Chem. Inf. Comput. Sci.*, **41**, 147–158.
- Golbraikh, A., Bonchev, D. and Tropsha, A. (2002) Novel ZE-isomerism descriptors derived from

- molecular topology and their application to QSAR analysis. *J. Chem. Inf. Comput. Sci.*, **42**, 769–787.
- Golbraikh, A., Bonchev, D., Xiao, Y.-D., and Tropsha, A. (2001b) Novel chiral topological descriptors and their applications to QSAR, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 219–223.
- Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H. and Tropsha, A. (2003) Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aid. Mol. Des.*, **17**, 241–253.
- Golbraikh, A. and Tropsha, A. (2002a) Beware of q^2 ! *J. Mol. Graph. Model.*, **20**, 269–276.
- Golbraikh, A. and Tropsha, A. (2002b) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput. Aid. Mol. Des.*, **16**, 357–369.
- Golbraikh, A. and Tropsha, A. (2002c) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Div.*, **5**, 231–243.
- Golbraikh, A. and Tropsha, A. (2003) QSAR modeling using chirality descriptors derived from molecular topology. *J. Chem. Inf. Comput. Sci.*, **43**, 144–154.
- Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Massachusetts, MA.
- Goldman, B.B. and Wipke, W.T. (2000) Quadratic shape descriptors. 1. Rapid superposition of dissimilar molecules using geometrically invariant surface descriptors. *J. Chem. Inf. Comput. Sci.*, **40**, 644–658.
- Golender, V.E., Drboglav, V.V. and Rozenblit, A.B. (1981) Graph potentials method and its application for chemical information processing. *J. Chem. Inf. Comput. Sci.*, **21**, 196–204.
- Goll, E.S. and Jurs, P.C. (1999a) Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model. *J. Chem. Inf. Comput. Sci.*, **39**, 974–983.
- Goll, E.S. and Jurs, P.C. (1999b) Prediction of vapor pressures of hydrocarbons and halohydrocarbons from molecular structure with a computational neural network model. *J. Chem. Inf. Comput. Sci.*, **39**, 1081–1089.
- Golovanov, I.B. and Tsygankova, I.G. (2000) Estimation of physico-chemical properties from the structure–property relationship: a new approach. *Quant. Struct. -Act. Relat.*, **19**, 554–564.
- GOLPE, Ver. 4.5, Multivariate Infometric Analysis s.r.l., Viale dei Castagni 16, Perugia, Italy.
- Golub, G.H. and van Loan, C.F. (1983) *Matrix Computations*, John Hopkins University Press, Baltimore, MD.
- Gombar, V.K. (1999) Reliable assessment of log P of compounds of pharmaceutical relevance. *SAR & QSAR Environ. Res.*, **10**, 371–380.
- Gombar, V.K., Borgstedt, H.H., Enslein, K., Hart, J.B. and Blake, B.W. (1991) A QSAR model of teratogenesis. *Quant. Struct. -Act. Relat.*, **10**, 306–332.
- Gombar, V.K. and Enslein, K. (1990) Quantitative structure–activity relationship (QSAR) studies using electronic descriptors calculated from topological and molecular orbital (MO) methods. *Quant. Struct. -Act. Relat.*, **9**, 321–325.
- Gombar, V.K. and Enslein, K. (1996) Assessment of n -octanol/water partition coefficient: when is the assessment reliable? *J. Chem. Inf. Comput. Sci.*, **36**, 1127–1134.
- Gombar, V.K., Enslein, K. and Blake, B.W. (1995) Assessment of developmental toxicity potential of chemicals by quantitative structure–toxicity relationship models. *Chemosphere*, **31**, 2499–2510.
- Gombar, V.K. and Jain, D.V.S. (1987) Quantification of molecular shape and its correlation with physico-chemical properties. *Indian J. Chem.*, **26**, 554–555.
- Gombar, V.K., Kumar, A. and Murthy, M.S. (1987) Quantitative structure–activity relationships. Part IX. A modified connectivity index as structure quantifier. *Indian J. Chem.*, **26**, 1168–1170.
- González Díaz, H., Bonet, I., Terán, C., De Clercq, E., Bello, R., García, M.M., Santana, L. and Uriarte, E. (2007) ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.*, **42**, 580–585.
- González Díaz, H., Gia, O., Uriarte, E., Hernandez, I., Ramos, R., Chaviano, M., Seijo, S., Castillo, J.A., Morales, L., Santana, L., Akpaloo, D., Molina, E., Cruz, M., Torres, L.A. and Cabrera, M.A. (2003) Markovian chemicals “*in silico*” design (MARCH-INSIDE), a promising approach for computer-aided molecular design. I. Discovery of anticancer compounds. *J. Mol. Model.*, **9**, 395–407.
- González Díaz, H., Marrero, Y., Hernandez, I., Bastida, I., Tenorio, E., Nasco, O., Uriarte, E., Castañedo, N., Cabrera, M.A., Aguila, E., Marrero, O., Morales, A. and Pérez, M. (2003) 3D-MEDNEs: an alternative “*in silico*” technique for chemical research in toxicology. 1. Prediction of chemically induced agranulocytosis. *Chem. Res. Toxicol.*, **16**, 1318–1327.
- González Díaz, H., Olazabal, E., Castañedo, N., Sánchez, I.H., Morales, A., Serrano, H.S.,

- González, J. and Ramos de Armas, R. (2002) Markovian chemicals 'in silico' design (MARCH-INSIDE), a promising approach for computer aided molecular design. II. Experimental and theoretical assessment of a novel method for virtual screening of fasciolicides. *J. Mol. Model.*, **8**, 237–245.
- González Díaz, H., Sánchez, I.H. and Uriarte, E. (2003) Symmetry considerations in Markovian chemicals 'in silico' design (MARCH-INSIDE). I. Central chirality codification, classification of ACE inhibitors and prediction of σ -receptor antagonist activities. *Comp. Biol. Chem.*, **27**, 217–227.
- González Díaz, H., Torres-Gómez, L.A., Guevara, Y., Almeida, M.S., Molina, R., Castañedo, N., Santana, L. and Uriarte, E. (2005) Markovian chemicals 'in silico' design (MARCH-INSIDE), a promising approach for computer-aided molecular design. III. 2.5D indices for the discovery of antibacterials. *J. Mol. Model.*, **11**, 116–123.
- González Díaz, H. and Uriarte, E. (2005) Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. *Biopolymers*, **77**, 296–303.
- Good, A.C. (1992) The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.*, **10**, 144–151.
- Good, A.C. (1995) 3D molecular similarity indices and their application in QSAR studies, in *Molecular Similarity in Drug Design* (ed. P.M. Dean), Chapman & Hall, London, UK, pp. 24–56.
- Good, A.C., Cho, S.J. and Mason, J.S. (2004) Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening. *J. Comput. Aid. Mol. Des.*, **18**, 523–527.
- Good, A.C., Ewing, T.J.A., Gschwend, D.A. and Kuntz, I.D. (1995) New molecular shape descriptors: application in database screening. *J. Comput. Aid. Mol. Des.*, **9**, 1–12.
- Good, A.C., Hodgkin, E.E. and Richards, W.G. (1992) Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.*, **32**, 188–191.
- Good, A.C. and Kuntz, I.D. (1995) Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput. Aid. Mol. Des.*, **9**, 373–379.
- Good, A.C., Peterson, S.J. and Richards, W.G. (1993) QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.*, **36**, 2929–2937.
- Good, A.C. and Richards, W.G. (1993) Rapid evaluation of shape similarity using Gaussian functions. *J. Chem. Inf. Comput. Sci.*, **33**, 112–116.
- Good, A.C. and Richards, W.G. (1998) Explicit calculation of 3D molecular similarity, in *3D QSAR in Drug Design*, Vol. 2 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 321–338.
- Good, A.C., So, S.-S. and Richards, W.G. (1993) Structure–activity relationships from molecular similarity matrices. *J. Med. Chem.*, **36**, 433–438.
- Goodford, P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–857.
- Goodford, P.J. (1995) The properties of force fields, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, Barcelona, Spain, pp. 199–205.
- Goodford, P.J. (1996) Multivariate characterization of molecules for QSAR analysis. *J. Chemom.*, **10**, 107–117.
- Goodford, P.J. (2006) The basic principles of GRID, in *Molecular Interaction Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 3–26.
- Gopinathan, M.S. and Jug, K. (1983a) Valency. I. A quantum chemical definition and properties. *Theor. Chim. Acta*, **63**, 497–509.
- Gopinathan, M.S. and Jug, K. (1983b) Valency. II. applications to molecules with first-row atoms. *Theor. Chim. Acta*, **63**, 511–527.
- Gordeeva, E.V., Katritzky, A.R., Shcherbukhin, V.V. and Zefirov, N.S. (1993) Rapid conversion of molecular graphs to three-dimensional using the MOLGEO program. *J. Chem. Inf. Comput. Sci.*, **33**, 102–111.
- Gordeeva, E.V., Molchanova, M.S. and Zefirov, N.S. (1990) General methodology and computer program for the exhaustive restoring of chemical structures by molecular connectivity indices. Solution of the inverse problem in QSAR/QSPR. *Tetrahedron Comput. Methodol.*, **3**, 389–415.
- Gordon, M. and Kennedy, J.W. (1973) The graph-like state of matter. Part 2. LCGI schemes for the thermodynamics of alkanes and the theory of inductive inference. *J. Chem. Soc. Faraday Trans II*, **69**, 484–504.
- Gordon, M. and Scantlebury, G.R. (1964) Non-random polycondensation: statistical theory of the substitution effect. *Trans. Faraday Soc.*, **60**, 604–621.
- Gordon, M.C. (1980) Quantitative structure–activity relationships by distance geometry: systematic

- analysis of dihydrofolate reductase inhibitors. *J. Math. Chem.*, **23**, 599–606.
- Gordon, P.A. (2001a) Statistical associating fluid theory. 1. Application toward describing isoparaffins. *Ind. Eng. Chem. Res.*, **40**, 2947–2955.
- Gordon, P.A. (2001b) Statistical associating fluid theory. 2. Estimation of parameters to predict Lube-ranged isoparaffin properties. *Ind. Eng. Chem. Res.*, **40**, 2956–2965.
- Gordy, W. (1946) A new method of determining electronegativity from other atomic properties. *Phys. Rev.*, **69**, 604–607.
- Gordy, W. (1947) Dependence of bond order and bond energy upon bond length. *J. Chim. Phys.*, **15**, 305–310.
- Gordy, W. (1951) Interpretation of nuclear quadrupole couplings in molecules. *J. Chim. Phys.*, **19**, 792–793.
- Görgényi, M., Dewulf, J., Van Langenhove, H. and Király, Z. (2005) Solubility of volatile organic compounds in aqueous ammonia solution. *Chemosphere*, **59**, 1083–1090.
- Gorše, M. and Žerovnik, J. (2004) A remark on modified Wiener indices. *MATCH Commun. Math. Comput. Chem.*, **50**, 109–116.
- Gough, J. and Hall, F.M. (1999a) Modeling antileukemic activity of carboquinones with electrotopological state and chi indices. *J. Chem. Inf. Comput. Sci.*, **39**, 356–361.
- Gough, J. and Hall, L.H. (1999b) Modeling the toxicity of amide herbicides using the electrotopological state. *Environ. Toxicol. Chem.*, **18**, 1069–1075.
- Gough, K.M., Belohorcova, K. and Kaiser, K.L. (1994) Quantitative structure–activity relationships (QSARs) of *Photobacterium phosphoreum* toxicity of nitrobenzene derivatives. *Sci. Total Environ.*, **142**, 179–190.
- Govers, H., Rupert, C. and Aiking, H. (1984) Quantitative structure–activity relationships for polycyclic aromatic hydrocarbons: correlation between molecular connectivity, physico-chemical properties, bioconcentration and toxicity in *Daphnia pulex*. *Chemosphere*, **13**, 227–236.
- Govers, H.A.J. (1990) Prediction of environmental behaviour and effects of polycyclic aromatic hydrocarbons by PAR and QSAR, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 411–432.
- Gozalbes, R., Doucet, J.P. and Derouin, F. (2002) Application of topological descriptors in QSAR and drug design: history and new trends. *Curr. Drug Targets Infect. Disord.*, **2**, 93–102.
- Gozalbes, R., Gálvez, J., García-Domenech, R. and Derouin, F. (1999) Molecular search of new active drugs against *Toxoplasma gondii*. *SAR & QSAR Environ. Res.*, **10**, 47–60.
- Graffis, C.A. and Ballantine, D.S. (2002) Characterization of phosphorus-containing gas chromatographic stationary phases by linear solvation energy relationships. *J. Chromat.*, **946**, 185–196.
- Graham, C., Gealy, R., Macina, O.T., Karol, M.H. and Rosenkranz, H.S. (1996) QSAR for allergic contact dermatitis. *Quant. Struct. -Act. Relat.*, **15**, 224–229.
- Graham, D.J. (2002) Information and organic molecules: structure considerations via integer statistics. *J. Chem. Inf. Comput. Sci.*, **42**, 215–221.
- Graham, D.J. and Schacht, D.V. (2000) Base information content in organic formulas. *J. Chem. Inf. Comput. Sci.*, **40**, 942–946.
- Graham, R.L., Hoffman, A.J. and Hosoya, H. (1977) On the distance matrix of a direct graph. *J. Graph Theory*, **1**, 85–88.
- Graham, R.L. and Lovasz, L. (1978) Distance matrix polynomials of trees. *Adv. Math.*, **29**, 60–88.
- Gramatica, P. (2001) QSAR approach to the evaluation of chemicals. *Chemistry Today*, 18–24.
- Gramatica, P. (2006) WHIM descriptors of shape. *QSAR Comb. Sci.*, **25**, 327–332.
- Gramatica, P. (2007) Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.*, **26**, 694–701.
- Gramatica, P., Battaini, F. and Papa, E. (2004) QSAR prediction of physico-chemical properties of esters. *Fresen. Environ. Bull.*, **13**, 1258–1262.
- Gramatica, P., Consolaro, F. and Pozzi, S. (2001) QSAR approach to POPs screening for atmospheric persistence. *Chemosphere*, **43**, 655–664.
- Gramatica, P., Consonni, V. and Pavan, M. (2003) Prediction of aromatic amines mutagenicity from theoretical molecular descriptors. *SAR & QSAR Environ. Res.*, **14**, 237–250.
- Gramatica, P., Consonni, V. and Todeschini, R. (1999) QSAR study on the tropospheric degradation of organic compounds. *Chemosphere*, **38**, 1371–1378.
- Gramatica, P., Corradi, M. and Consonni, V. (2000) Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere*, **41**, 763–777.
- Gramatica, P. and Di Guardo, A. (2002) Screening of pesticides for environmental partitioning tendency. *Chemosphere*, **47**, 947–956.
- Gramatica, P., Giani, E. and Papa, E. (2007) Statistical external validation and consensus modeling: a QSPR case study for K_{oc} prediction. *J. Mol. Graph. Model.*, **25**, 755–766.

- Gramatica, P., Navas, N. and Todeschini, R. (1998) 3D-Modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs). *Chemom. Intell. Lab. Syst.*, **40**, 53–63.
- Gramatica, P., Navas, N. and Todeschini, R. (1999) Classification of organic solvents and modelling of their physico-chemical properties by chemometric methods using different sets of molecular descriptors. *TRAC*, **18**, 461–471.
- Gramatica, P. and Papa, E. (2003) QSAR modeling of bioconcentration factor by theoretical molecular descriptors. *QSAR Comb. Sci.*, **22**, 374–385.
- Gramatica, P. and Papa, E. (2005) An update of the BCF QSAR model based on theoretical molecular descriptors. *QSAR Comb. Sci.*, **24**, 953–960.
- Gramatica, P. and Papa, E. (2007) Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. *Environ. Sci. Technol.*, **41**, 2833–2839.
- Gramatica, P., Papa, E. and Pozzi, S. (2004) Prediction of POP environmental persistence and long range transport by QSAR and chemometric approaches. *Fresen. Environ. Bull.*, **13**, 1204–1209.
- Gramatica, P., Pilutti, P. and Papa, E. (2002) Ranking of volatile organic compounds for tropospheric degradability by oxidants: a QSPR approach. *SAR & QSAR Environ. Res.*, **13**, 743–753.
- Gramatica, P., Pilutti, P. and Papa, E. (2003a) Predicting the NO₃ radical tropospheric degradability of organic pollutants by theoretical molecular descriptors. *Atmos. Environ.*, **37**, 3115–3124.
- Gramatica, P., Pilutti, P. and Papa, E. (2003b) QSAR prediction of ozone tropospheric degradation. *QSAR Comb. Sci.*, **22**, 364–373.
- Gramatica, P., Pilutti, P. and Papa, E. (2004a) A tool for the assessment of VOC degradability by tropospheric oxidants starting from chemical structure. *Atmos. Environ.*, **38**, 6167–6175.
- Gramatica, P., Pilutti, P. and Papa, E. (2004b) Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *J. Chem. Inf. Comput. Sci.*, **44**, 1794–1802.
- Gramatica, P., Pilutti, P. and Papa, E. (2005) Ranking of phenols for abiotic oxidation in aqueous environment: a QSPR approach. *Ann. Chim. (Rome)*, **95**, 199–209.
- Gramatica, P., Pilutti, P. and Papa, E. (2007) Approaches for externally validated QSAR modelling of nitrated polycyclic aromatic hydrocarbon mutagenicity. *SAR & QSAR Environ. Res.*, **18**, 169–178.
- Gramatica, P., Pozzi, S., Consonni, V. and Di Guardo, A. (2002) Classification of environmental pollutants for global mobility potential. *SAR & QSAR Environ. Res.*, **13**, 205–217.
- Gramatica, P., Santagostino, A., Bolzacchini, E. and Rindone, B. (2002) Atmospheric monitoring, toxicology and QSAR modelling of nitrophenols. *Fresen. Environ. Bull.*, **11**, 757–762.
- Gramatica, P., Vighi, M., Consolaro, F., Todeschini, R., Finizio, A. and Faust, M. (2001) QSAR approach for the selection of congeneric compounds with a similar toxicological mode of action. *Chemosphere*, **42**, 873–883.
- Grant, J.A., Gallardo, M.A. and Pickup, B.T. (1996) A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.*, **17**, 1653–1666.
- Grant, J.A. and Pickup, B.T. (1995) Gaussian description of molecular shape. *J. Phys. Chem.*, **99**, 3503–3510.
- Graovac, A. and Gutman, I. (1978) The determinant of the adjacency matrix of the graph of a conjugated molecule. *Croat. Chem. Acta*, **51**, 133–140.
- Graovac, A. and Gutman, I. (1979) The determinant of the adjacency matrix of a molecular graph. *MATCH Commun. Math. Comput. Chem.*, **6**, 49–73.
- Graovac, A., Gutman, I. and Trinajstić, N. (1977) *Topological Approach to the Chemistry of Conjugated Molecules*, Springer-Verlag, Berlin, Germany, p. 123.
- Graovac, A., Gutman, I., Trinajstić, N. and Živković, T. (1972) Graph theory and molecular orbitals: application of Sachs theorem. *Theor. Chim. Acta*, **26**, 67.
- Graovac, A., Juvan, M., Mohar, B. and Žerovnik, J. (1999) Computing the determinant and the algebraic structure count in polygraphs. *Croat. Chem. Acta*, **72**, 853–867.
- Graovac, A. and Pisanski, T. (1991) On the Wiener index of a graph. *J. Math. Chem.*, **8**, 53–62.
- Graovac, A., Vukicević, D., Ježek, D. and Žerovnik, J. (2005) Simplified computation of matchings in polygraphs. *Croat. Chem. Acta*, **78**, 283–287.
- Grassy, G. and Lahana, R. (1993) Statistical analysis and shape recognition: applications to MD simulations, conformational analysis and structure–activity relationships, in *Trends in QSAR and Molecular Modelling* 92 (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 216–219.
- Grassy, G., Trape, P., Bompart, J., Calas, B. and Auzou, G. (1995) Variable mapping of structure–activity relationships. Application to 17-spirolactone derivatives with mineralocorticoid activity. *J. Mol. Graph.*, **13**, 356–367.

- Gratteri, P., Cruciani, G., Scapechi, S. and Romanelli, M.N. (2001) Grid independent descriptors (GRIND) in the rational design of muscarinic antagonists, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 241–243.
- Gratteri, P., Romanelli, M.N., Cruciani, G., Bonaccini, C. and Melani, F. (2004) GRIND-derived pharmacophore model for a series of α -tropanyl derivative ligands of the sigma-2 receptor. *J. Comput. Aid. Mol. Des.*, **18**, 361–374.
- Grdadolnik, S.G. and Mierke, D.F. (1997) Structural characterization of the molecular dimer of the peptide antibiotic vancomycin by distance geometry in four spatial dimensions. *J. Chem. Inf. Comput. Sci.*, **37**, 1044–1047.
- Greco, G., Novellino, E., Fiorini, I., Nacci, V., Campiani, G., Ciani, S.M., Garofalo, A., Bernasconi, P. and Mennini, T. (1994a) A comparative molecular field analysis model for 6-arylpyrrolo[2,1-d][1,5]benzothiazepines binding selectively to the mitochondrial benzodiazepine receptor. *J. Med. Chem.*, **37**, 4100–4108.
- Greco, G., Novellino, E. and Martin, Y.C. (1997) Approaches to three-dimensional quantitative structure–activity relationships, in *Reviews in Computational Chemistry*, Vol. 11 (eds K.F.B. Lipkowitz and D.B. Boyd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 183–240.
- Greco, G., Novellino, E., Pellicchia, M., Silipo, C. and Vittoria, A. (1993) Use of the hydrophobic substituent constant in a comparative molecular field analysis (CoMFA) on a set of anilides inhibiting the hill reaction. *SAR & QSAR Environ. Res.*, **1**, 301–334.
- Greco, G., Novellino, E., Pellicchia, M., Silipo, C. and Vittoria, A. (1994b) Effects of variable sampling on CoMFA coefficient contour maps. *J. Mol. Graph.*, **12**, 67–68.
- Greco, G., Novellino, E., Pellicchia, M., Silipo, C. and Vittoria, A. (1994c) Effects of variable selection on CoMFA coefficient contour maps in a set of triazines inhibiting DHFR. *J. Comput. Aid. Mol. Des.*, **8**, 97–112.
- Greco, G., Novellino, E., Silipo, C. and Vittoria, A. (1991) Comparative molecular field analysis on a set of muscarinic agonists. *Quant. Struct. -Act. Relat.*, **10**, 289–299.
- Green, A.L. (1956) A simple approximation to the resonance energies of aromatic molecules. *J. Chem. Soc.*, 1886–1888.
- Green, S.M. and Marshall, G.R. (1995) 3D-QSAR: a current perspective. *Trends Pharmacol. Sci.*, **16**, 285–291.
- Gregori-Puigjané, E. and Mestres, J. (2006) SHED: Shannon entropy descriptors from topological feature distributions. *J. Chem. Inf. Model.*, **46**, 1615–1622.
- Grgas, B., Nikolić, S., Paulić, N. and Raos, N. (1999) Estimation of stability constants of copper(II) chelates with N-alkylated amino acids using topological indices. *Croat. Chem. Acta*, **72**, 885–895.
- Grigoras, S. (1990) A structural approach to calculate physical properties of pure organic substances: the critical temperature, critical volume and related properties. *J. Comput. Chem.*, **11**, 493–510.
- Grigorov, M., Weber, J., Tronchet, J.M.J., Jefford, C. W., Milhous, W.K. and Maric, D. (1997) A QSAR study of the antimalarial activity of some synthetic 1,2,4-trioxanes. *J. Chem. Inf. Comput. Sci.*, **37**, 124–130.
- Grob, C.A. (1985) 95. Inductive charge dispersal in quinuclidinium ions. *Helv. Chim. Acta*, **68**, 882–886.
- Grob, C.A. and Schlageter, M.G. (1976) 31. The derivation of inductive substituent constants from pK_a values of 4-substituted quinuclidines. Polar effects. Part I. *Helv. Chim. Acta*, **59**, 264–276.
- Groditzky, J.A. and Coats, J.R. (2002) QSAR evaluation of monoterpenoids' insecticidal activity. *J. Agr. Food Chem.*, **50**, 4576–4580.
- Gross, K.C., Seybold, P.G., Peralta-Inga, Z., Murray, J. S. and Politzer, P. (2001) Comparison of quantum chemical parameters and Hammett constants in correlating pK_a values of substituted anilines. *J. Org. Chem.*, **66**, 6919–6925.
- Grossman, S.C. (1985) Chemical ordering of molecules: a graph theoretical approach to structure–property studies. *Int. J. Quant. Chem.*, **28**, 1–16.
- Grossman, S.C., Jerman-Blazic Dzonova, B. and Randić, M. (1985) A graph theoretical approach to quantitative structure–activity relationship. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **12**, 123–139.
- Grover, M., Gulati, M., Singh, B. and Singh, S. (2000) Correlation of penicillin structure with rate constants for basic hydrolysis. *Pharm. Pharmacol. Comm.*, **6**, 355–363.
- Grüber, C. and Buss, V. (1989) Quantum-mechanically calculated properties for the development of quantitative structure–activity relationships (QSARs). pK_a -values of phenols and aromatic and aliphatic carboxylic acids. *Chemosphere*, **19**, 1595–1609.
- Grunenberg, J. and Herges, R. (1995) Prediction of chromatographic retention values (R_M) and partition coefficients ($\log P_{oct}$) using a combination of semiempirical self-consistent reaction field

- calculations and neural networks. *J. Chem. Inf. Comput. Sci.*, **35**, 905–911.
- Grünheidt Borges, E. and Takahashi, Y. (2001) QSAR study of anti-ulcer compounds using calculated parameters. *J. Mol. Struct. (Theochem)*, **539**, 245–251.
- Grünheidt Borges, E. and Takahata, Y. (2002) The 4-indolyl-2-guanidinothiazoles QSAR study of anti-ulcer activity using quantum descriptors. *J. Mol. Struct. (Theochem)*, **580**, 263–270.
- Grunwald, E. and Winstein, S. (1948) The correlation of solvolysis rates. *J. Am. Chem. Soc.*, **70**, 846–854.
- Guevara, N. (1999) Fragmental graphs. A novel approach to generate a new family of descriptors. Applications to QSPR studies. *J. Mol. Struct. (Theochem)*, **493**, 29–36.
- Guha, R. and Jurs, P.C. (2005a) Determining the validity of a QSAR model: a classification approach. *J. Chem. Inf. Model.*, **45**, 65–73.
- Guha, R. and Jurs, P.C. (2005b) Interpreting computational neural network QSAR models: a measure of descriptor importance. *J. Chem. Inf. Model.*, **45**, 800–806.
- Guha, R., Serra, J.R. and Jurs, P.C. (2004) Generation of QSAR sets with a self-organizing map. *J. Mol. Graph. Model.*, **23**, 1–14.
- Guha, R., Stanton, D.T. and Jurs, P.C. (2005) Interpreting computational neural network quantitative structure–activity relationship models: a detailed interpretation of the weights and biases. *J. Chem. Inf. Model.*, **45**, 1109–1121.
- Guha, R. and Ven Drie, J.H. (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.*, **48**, 646–658.
- Gund, P., Barry, D.C., Blaney, J.M. and Cohen, N.C. (1988) Guidelines for publications in molecular modeling related to medicinal chemistry. *J. Med. Chem.*, **31**, 2230–2234.
- Gundertofte, K. and Jørgensen, F.S. (eds) (2000) *Molecular Modeling and Prediction of Bioactivity*, Kluwer Academic/Plenum Publishers, New York, p. 502.
- Guo, X. and Nandy, A. (2003) Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy. *Chem. Phys. Lett.*, **369**, 361–366.
- Guo, X. and Randić, M. (1999) Trees with the same topological index *J. SAR & QSAR Environ. Res.*, **10**, 381–394.
- Guo, X., Randić, M. and Basketter, D. (2001) A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Lett.*, **350**, 106–112.
- Guo, X., Randić, M. and Klein, D.J. (1996) Analytical expressions for the count of LM-conjugated circuits of benzenoid hydrocarbons. *Int. J. Quant. Chem.*, **60**, 943–958.
- Guo, X. and Zhang, F. (1993) An efficient algorithm for generating all Kekulé patterns of a generalized benzenoid system. *J. Math. Chem.*, **12**, 163–172.
- Gupta, M.K. and Prabhakar, Y.S. (2005) Topological descriptors in modeling the antimalarial activity of 4-(3',5'-disubstituted anilino)quinolines. *J. Chem. Inf. Model.*, **46**, 93–102.
- Gupta, S., Metthew, S., Abreu, P.M. and Aires-de-Sousa, J. (2006) QSAR analysis of phenolic antioxidants using MOLMAP descriptors of local properties. *Bioorg. Med. Chem.*, **14**, 1199–1206.
- Gupta, S., Singh, M. and Madan, A.K. (1999) Superpendentic index: a novel topological descriptor for predicting biological activity. *J. Chem. Inf. Comput. Sci.*, **39**, 272–277.
- Gupta, S., Singh, M. and Madan, A.K. (2000) Connective eccentricity index: a novel topological descriptor for predicting biological activity. *J. Mol. Graph. Model.*, **18**, 18–25.
- Gupta, S., Singh, M. and Madan, A.K. (2001a) Applications of graph theory: relationships of molecular connectivity index and atomic molecular connectivity index with anti-HSV activity. *J. Mol. Struct. (Theochem)*, **571**, 147–152.
- Gupta, S., Singh, M. and Madan, A.K. (2001b) Predicting anti-HIV activity: computational approach using a novel topological descriptor. *J. Comput. Aid. Mol. Des.*, **15**, 671–678.
- Gupta, S., Singh, M. and Madan, A.K. (2002) Eccentric distance sum: a novel graph invariant for predicting biological and physical properties. *J. Math. Anal. Appl.*, **275**, 386–401.
- Gupta, S., Singh, M. and Madan, A.K. (2003) Novel topochemical descriptors for predicting anti-HIV activity. *Indian J. Chem.*, **42**, 1414–1425.
- Gupta, S.P. and Sharma, M.K. (1986) Molecular connectivity in Hückel's molecular orbital theory. II. Parametrization of resonance integral. *MATCH Commun. Math. Comput. Chem.*, **21**, 123–132.
- Gupta, S.P., Singh, P. and Bindal, M.C. (1983) QSAR studies on hallucinogens. *Chem. Rev.*, **83**, 633–649.
- Gurden, S.P., Westerhuis, J.A., Bro, R. and Smilde, A. K. (2001) A comparison of multiway regression and scaling methods. *Chemom. Intell. Lab. Syst.*, **59**, 121–136.
- Gussio, R., Pattabiraman, N., Zaharevitz, D.W., Kellogg, G.E., Topol, I.A., Rice, W.G., Schaeffer, C. A., Erickson, J.W. and Burt, S.K. (1996) All atom models for the nonnucleoside binding site of HIV-1 reverse transcriptase complexed with inhibitors. A 3D QSAR approach. *J. Med. Chem.*, **39**, 1645–1650.

- Gustafson, D.I. (1989) Groundwater ubiquity score: a simple method for assessing pesticide leachability. *Environ. Toxicol. Chem.*, **8**, 339–357.
- Gute, B.D. and Basak, S.C. (1997) Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. *SAR & QSAR Environ. Res.*, **7**, 117–131.
- Gute, B.D. and Basak, S.C. (2001) Molecular similarity-based estimation of properties: a comparison of three structure spaces. *J. Mol. Graph. Model.*, **20**, 95–109.
- Gute, B.D., Basak, S.C., Mills, D. and Hawkins, D.M. (2002) Tailored similarity spaces for the prediction of physico-chemical properties. *Internet Electron. J. Mol. Des.*, **1**, 374–387.
- Gute, B.D., Grunwald, G.D. and Basak, S.C. (1999) Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): a hierarchical QSAR approach. *SAR & QSAR Environ. Res.*, **10**, 1–15.
- Gute, B.D., Grunwald, G.D., Mills, D. and Basak, S.C. (2001) Molecular similarity based estimation of properties: a comparison of structure spaces and property spaces. *SAR & QSAR Environ. Res.*, **11**, 363–382.
- Gutman, I. (1974) Some topological properties of benzenoid systems. *Croat. Chem. Acta*, **46**, 209–215.
- Gutman, I. (1976) Empirical parameters for donor and acceptor properties of solvents. *Electrochim. Acta*, **21**, 661–670.
- Gutman, I. (1977) New applications of the Dewar index. *Croat. Chem. Acta*, **49**, 635–641.
- Gutman, I. (1978a) The energy of a graph. *Ber. Math. Statist. Sect. Forschungszentrum (Graz, German)* **103**, 1–22.
- Gutman, I. (1978b) Topological formulas for free-valency index. *Croat. Chem. Acta*, **51**, 29–33.
- Gutman, I. (1979) The matching polynomial. *MATCH Commun. Math. Comput. Chem.*, **6**, 75–91.
- Gutman, I. (1983) Characteristic and matching polynomials of benzenoid hydrocarbons. *J. Chem. Soc. Faraday Trans II*, **79**, 337–345.
- Gutman, I. (1985) Topological properties of benzenoid systems. XLI. Carbon–carbon bond types and connectivity indices of benzenoid hydrocarbons. *J. Serb. Chem. Soc.*, **50**, 451–455.
- Gutman, I. (1986) Topological properties of benzenoid systems. XLVIII. Two contradictory formulas for total π -electron energy and their reconciliation. *MATCH Commun. Math. Comput. Chem.*, **21**, 317–324.
- Gutman, I. (1987a) Acyclic conjugated molecules, trees and their energies. *J. Math. Chem.*, **1**, 123–144.
- Gutman, I. (1987b) Wiener numbers of benzenoid hydrocarbons: two theorems. *Chem. Phys. Lett.*, **136**, 134–136.
- Gutman, I. (1988a) Hosoya index of a class of benzenoid hydrocarbons. *J. Serb. Chem. Soc.*, **53**, 129–132.
- Gutman, I. (1988b) On the Hosoya index of very large molecules. *MATCH Commun. Math. Comput. Chem.*, **23**, 95–104.
- Gutman, I. (1988c) Topological properties of benzenoid systems. LI. Hosoya index of molecules containing a polyacene fragment. *Z. Naturforsch.*, **43a**, 939–942.
- Gutman, I. (1990) A property of the simple topological index. *MATCH Commun. Math. Comput. Chem.*, **25**, 131–140.
- Gutman, I. (1991a) Polynomials in graph theory, in *Chemical Graph Theory: Introduction and Fundamentals* (eds D. Bonchev and D.H. Rouvray), Abacus Press/Gordon and Breach Science Publishers, New York, pp. 133–176.
- Gutman, I. (1991b) Topological properties of benzenoid systems. Merrifield–Simmons indices and independence polynomials of unbranched catafusenes. *Rev. Roum. Chim.*, **36**, 379–388.
- Gutman, I. (1992a) Remark on the moment expansion of total π -electron energy. *Theor. Chim. Acta*, **83**, 313–318.
- Gutman, I. (1992b) Some analytical properties of the independence and matching polynomials. *MATCH Commun. Math. Comput. Chem.*, **28**, 139–150.
- Gutman, I. (1993a) A new method for the calculation of the Wiener number of acyclic molecules. *J. Mol. Struct. (Theochem)*, **285**, 137–142.
- Gutman, I. (1993b) Calculating the Wiener number: the Doyle–Graver method. *J. Serb. Chem. Soc.*, **58**, 745–750.
- Gutman, I. (1994a) Formula for the Wiener number of trees and its extension to graphs containing cycles. *Graph Theory Notes, New York*, **27**, 9–15.
- Gutman, I. (1994b) Selected properties of the Schultz molecular topological index. *J. Chem. Inf. Comput. Sci.*, **34**, 1087–1089.
- Gutman, I. (1995) The topological indices of linear phenylenes. *J. Serb. Chem. Soc.*, **60**, 99–104.
- Gutman, I. (1997) A property of the Wiener number and its modifications. *Indian J. Chem.*, **36**, 128–132.
- Gutman, I. (1998) Permanents of adjacency matrices and their dependence on molecular structure. *Polycycl. Aromat. Comp.*, **12**, 281–287.
- Gutman, I. (2001) The energy of a graph: old and new results, in *Algebraic Combinatorics and Applications* (eds A. Betten, A. Kohnert, R. Laue and A.

- Wassermann), Springer-Verlag, Berlin, Germany, pp. 196–211.
- Gutman, I. (2002a) Molecular graphs with minimal and maximal Randić indices. *Croat. Chem. Acta*, **75**, 357–369.
- Gutman, I. (2002b) Relation between hyper-Wiener and Wiener index. *Chem. Phys. Lett.*, **364**, 352–356.
- Gutman, I. (2002c) Two theorems on connectivity indices. *J. Serb. Chem. Soc.*, **67**, 99–102.
- Gutman, I. (2003a) Hyper-Wiener index and Laplacian spectrum. *J. Serb. Chem. Soc.*, **68**, 949–952.
- Gutman, I. (2003b) Relation between the Laplacian and the ordinary characteristic polynomial. *MATCH Commun. Math. Comput. Chem.*, **47**, 133–140.
- Gutman, I. (2004) A new hyper-Wiener index. *Croat. Chem. Acta*, **77**, 61–64.
- Gutman, I. (2005) Topology and stability of conjugated hydrocarbons. The dependence of total π -electron energy on molecular topology. *J. Serb. Chem. Soc.*, **70**, 441–456.
- Gutman, I. (2006) Chemical graph theory: the mathematical connection. *Adv. Quant. Chem.*, **51**, 125–138.
- Gutman, I., Araujo, O. and Morales, D.A. (2000a) Bounds for the Randić connectivity index. *J. Chem. Inf. Comput. Sci.*, **40**, 593–598.
- Gutman, I., Araujo, O. and Morales, D.A. (2000b) Estimating the connectivity index of a saturated hydrocarbon. *Indian J. Chem.*, **39**, 381–385.
- Gutman, I., Araujo, O. and Rada, J. (2000c) An identity for Randić's connectivity index and its applications. *Acta Chim. Hung. -Mod. Chem.*, **137**, 653–658.
- Gutman, I., Bonchev, D., Seitz, W.A. and Gordeeva, E. V. (1995) Complementing the proof of the limit of relative atomic moments. *J. Chem. Inf. Comput. Sci.*, **35**, 894–895.
- Gutman, I., Bosanac, S. and Trinajstić, N. (1978) Graph theory and molecular orbitals. XX. Local and long range contributions to bond order. *Croat. Chem. Acta*, **51**, 293–298.
- Gutman, I. and Cioslowski, J. (1987) Bounds for the Hosoya index. *Z. Naturforsch.*, **42a**, 438–440.
- Gutman, I. and Cyvin, S.J. (1988a) All-benzenoid systems: topological properties of benzenoid systems. LVII. *MATCH Commun. Math. Comput. Chem.*, **23**, 175–178.
- Gutman, I. and Cyvin, S.J. (1988b) Hosoya index of fused molecules. *MATCH Commun. Math. Comput. Chem.*, **23**, 89–94.
- Gutman, I. and Das, K.C. (2004) The first Zagreb index 30 years after. *MATCH Commun. Math. Comput. Chem.*, **50**, 83–92.
- Gutman, I. and Diudea, M.V. (1998) Defining Cluj matrices and Cluj invariants. *J. Serb. Chem. Soc.*, **63**, 497–504.
- Gutman, I. and Dobrynin, A.A. (1998) The Szeged index: a success story. *Graph Theory Notes, New York*, **34**, 37–44.
- Gutman, I. and Dömötör, G. (1994) Wiener number of polyphenyls and phenylenes. *Z. Naturforsch.*, **49a**, 1040–1044.
- Gutman, I. and El-Basil, S. (1986) Fibonacci graphs. *MATCH Commun. Math. Comput. Chem.*, **20**, 81–94.
- Gutman, I. and Estrada, E. (1996) Topological indices based on line graph of the molecular graph. *J. Chem. Inf. Comput. Sci.*, **36**, 541–543.
- Gutman, I., Estrada, E. and Ivanciuc, O. (1999) Some properties of the Wiener polynomial trees. *Graph Theory Notes, New York*, **36**, 7–13.
- Gutman, I., Estrada, E. and Rodríguez-Velásquez, J.A. (2007) On a graph-spectrum-based structure descriptor. *Croat. Chem. Acta*, **80**, 151–154.
- Gutman, I. and Furtula, B. (2003) Hyper-Wiener index vs. Wiener index. Two highly correlated structure-descriptors. *Monatsh. Chem.*, **134**, 975–981.
- Gutman, I., Furtula, B. and Arsic, B. (2004a) On structure descriptors related with intramolecular energy of alkanes. *Z. Naturforsch.*, **59a**, 694–698.
- Gutman, I., Furtula, B., Arsic, B. and Bošković, Ž. (2004b) On the relation between Zenkevich and Wiener indices of alkanes. *J. Serb. Chem. Soc.*, **69**, 265–271.
- Gutman, I., Furtula, B. and Belic, J. (2003) Note on the hyper-Wiener index. *J. Serb. Chem. Soc.*, **68**, 943–948.
- Gutman, I., Furtula, B., Toropov, A.A. and Toropova, A.P. (2005) The graph of atomic orbitals and its basic properties. 2. Zagreb indices. *MATCH Commun. Math. Comput. Chem.*, **53**, 225–230.
- Gutman, I., Furtula, B., Vidović, D. and Hosoya, H. (2004c) A concealed property of the topological index. *Z. Bull. Chem. Soc. Jap.*, **77**, 491–496.
- Gutman, I., Graovac, A. and Mohar, B. (1982) On the existence of a Hermitian matrix whose characteristic polynomial is the matching polynomial of a molecular graph. *MATCH Commun. Math. Comput. Chem.*, **13**, 129–150.
- Gutman, I., Hosoya, H. and Babic, D. (1996) Topological indices and graph polynomials of some macrocyclic belt-shaped molecules. *J. Chem. Soc. Faraday Trans.*, **92**, 625–628.
- Gutman, I., Hosoya, H., Uraković, G. and Ristić, L. (1992) Two variants of the topological index and the relations between them. *Bull. Chem. Soc. Jap.*, **65**, 14–18.

- Gutman, I., Indulal, G. and Todeschini, R. (2008) Generalizing the McClelland bounds for total π -electron energy. *Z. Naturforsch.*, **63a**, 280–282.
- Gutman, I. and Jovašević, V. (1998) Wiener indices of benzenoid hydrocarbons containing two linear polyacene fragments. *J. Serb. Chem. Soc.*, **63**, 31–40.
- Gutman, I., Kennedy, J.W. and Quintas, L.V. (1990) Wiener numbers of random benzenoid chains. *Chem. Phys. Lett.*, **173**, 403–408.
- Gutman, I., Khadikar, P.V., Rajput, P.V. and Karmarkar, S. (1995) The Szeged index of polyacenes. *J. Serb. Chem. Soc.*, **60**, 759–764.
- Gutman, I. and Klavžar, S. (1995) An algorithm for the calculation of the Szeged index of benzenoid hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **35**, 1011–1014.
- Gutman, I. and Klavžar, S. (1997) Bounds for the Schultz molecular topological index of benzenoid systems in terms of the Wiener index. *J. Chem. Inf. Comput. Sci.*, **37**, 741–744.
- Gutman, I. and Klavžar, S. (1998) Relations between Wiener numbers of benzenoid hydrocarbons and phenylenes. *Acta Chim. Hung.-Mod. Chem.*, **135**, 45–55.
- Gutman, I., Klavžar, S., Petrovšek, M. and Žigert, P. (2001) On Hosoya polynomials of benzenoid graphs. *MATCH Commun. Math. Comput. Chem.*, **43**, 49–66.
- Gutman, I., Kolakovic, N. and Cyvin, S.J. (1989a) Hosoya index of some polymers. *MATCH Commun. Math. Comput. Chem.*, **24**, 105–117.
- Gutman, I., Kolakovic, N., Graovac, A. and Babic, D. (1989b) A method for calculation of the Hosoya index of polymers, in *MATH/CHEM/COMP 1988* (ed. A. Graovac), Elsevier, Amsterdam, The Netherlands, pp. 141–154.
- Gutman, I. and Körtévlyesi, T. (1995) Wiener indices and molecular surfaces. *Z. Naturforsch.*, **50a**, 669–671.
- Gutman, I. and Kruszewski, J. (1985) On the occurrence of eigenvalue one in the graph spectrum of benzenoid systems. *Nouv. J. Chim.*, **9**, 669–670.
- Gutman, I., Lee, S.-L., Chu, C.H. and Luo, Y.-R. (1994) Chemical applications of the Laplacian spectrum of molecular graphs: studies of the Wiener number. *Indian J. Chem.*, **33**, 603–608.
- Gutman, I. and Lepovic, M. (2001) Choosing the exponent in the definition of the connectivity index. *J. Serb. Chem. Soc.*, **66**, 605–611.
- Gutman, I., Lepovic, M., Vidović, D. and Clark, L.H. (2002) Exponent-dependent properties of the connectivity index. *Indian J. Chem.*, **41**, 457–461.
- Gutman, I., Linert, W., Lukovits, I. and Dobrynin, A. A. (1997) Trees with external hyper-Wiener index: mathematical basis and chemical applications. *J. Chem. Inf. Comput. Sci.*, **37**, 349–354.
- Gutman, I., Linert, W., Lukovits, I. and Tomović, Ž. (2000a) On the multiplicative Wiener index and its possible chemical applications. *Monatsh. Chem.*, **131**, 421–427.
- Gutman, I., Linert, W., Lukovits, I. and Tomović, Ž. (2000b) The multiplicative version of the Wiener index. *J. Chem. Inf. Comput. Sci.*, **40**, 113–116.
- Gutman, I., Luo, Y.L. and Lee, S.-L. (1993) The mean isomer degeneracy of the Wiener index. *J. Chin. Chem. Soc.*, **40**, 195–198.
- Gutman, I. and Marković, S. (1993) Benzenoid graphs with equal maximum eigenvalues. *J. Math. Chem.*, **13**, 213–215.
- Gutman, I., Marković, S., Popovic, L., Spalević, Z. and Pavlović, L. (1997) The relation between the Wiener indices of phenylenes and their hexagonal squeezes. *J. Serb. Chem. Soc.*, **62**, 207–210.
- Gutman, I. and Marković, Z. (1986) Truncated Hosoya index. *J. Serb. Chem. Soc.*, **51**, 455–458.
- Gutman, I. and Marković, Z. (1987) Approximate formulas for Hosoya's topological index. *Bull. Chem. Soc. Jap.*, **60**, 2611–2614.
- Gutman, I., Marković, Z. and Marković, S. (1987) A simple method for the approximate calculation of Hosoya's index. *Chem. Phys. Lett.*, **134**, 139–142.
- Gutman, I. and Medeleanu, M. (1998) On the structure-dependence of the largest eigenvalue of the distance matrix of an alkane. *Indian J. Chem.*, **37**, 569–573.
- Gutman, I., Miljković, O., Caporossi, G. and Hansen, P. (1999) Alkanes with small and large Randić connectivity index. *Chem. Phys. Lett.*, **306**, 366–372.
- Gutman, I., Milun, M. and Trinajstić, N. (1977) Graph theory and molecular orbitals. 19. Nonparametric resonance energies of arbitrary conjugated systems. *J. Am. Chem. Soc.*, **99**, 1692–1704.
- Gutman, I. and Mizoguchi, N. (1990) A property of the circuit characteristic polynomial. *J. Math. Chem.*, **5**, 81–82.
- Gutman, I. and Mohar, B. (1996) The quasi-Wiener and the Kirchhoff indices coincide. *J. Chem. Inf. Comput. Sci.*, **36**, 982–985.
- Gutman, I. and Plath, P.J. (2001) On molecular graphs and digraphs of annulenes and their spectra. *J. Serb. Chem. Soc.*, **66**, 237–241.
- Gutman, I., Plavšić, D., Šoškić, M., Landeka, I. and Graovac, A. (1997) On the calculation of the path numbers ^1Z , ^2Z and the Hosoya Z index. *Croat. Chem. Acta*, **70**, 941–954.
- Gutman, I. and Polansky, O.E. (1986a) A regularity for the boiling points of alkanes and its mathematical

- modeling. *Z. Phys. Chemie (German)*, **267**, 1152–1158.
- Gutman, I. and Polansky, O.E. (1986b) *Mathematical Concepts in Organic Chemistry*, Springer, Berlin, Germany.
- Gutman, I. and Polansky, O.E. (1986c) Wiener numbers of polyacenes and related benzenoid molecules. *MATCH Commun. Math. Comput. Chem.*, **20**, 115–123.
- Gutman, I., Popovic, L., Estrada, E. and Bertz, S.H. (1998) The line graph model. Predicting physico-chemical properties of alkanes. *Acta Chim. Hung. - Mod. Chem.*, **135**, 147–155.
- Gutman, I., Popovic, L., Mishra, B.K., Kuanar, M., Estrada, E. and Guevara, N. (1997) Application of line graphs in physical chemistry. Predicting the surface tensions of alkanes. *J. Serb. Chem. Soc.*, **62**, 1025–1029.
- Gutman, I. and Potgieter, J.H. (1997) Wiener index and intermolecular forces. *J. Serb. Chem. Soc.*, **62**, 185–192.
- Gutman, I. and Pyka, A. (1997) New topological indices for distinguishing between enantiomers and stereoisomers: a mathematical analysis. *J. Serb. Chem. Soc.*, **62**, 261–265.
- Gutman, I., Radenković, S., Furtula, B., Mansour, T. and Schork, M. (2007) Relating Estrada index with spectral radius. *J. Serb. Chem. Soc.*, **72**, 1321–1327.
- Gutman, I. and Randić, M. (1977) Algebraic characterization of skeletal branching. *Chem. Phys. Lett.*, **47**, 15–19.
- Gutman, I. and Rosenfeld, V.R. (1996) Spectral moments of polymer graphs. *Theor. Chim. Acta*, **93**, 191–197.
- Gutman, I., Rosenfeld, V.R. and Marković, Z. (1987) Approximate formula for Hosoya's topological index. *J. Serb. Chem. Soc.*, **52**, 139–144.
- Gutman, I. and Rouvray, D.H. (1990) A new theorem for the Wiener molecular branching index of trees with perfect matchings. *Computers Chem.*, **14**, 29–32.
- Gutman, I., Rücker, C. and Rücker, G. (2001) On walks in molecular graphs. *J. Chem. Inf. Comput. Sci.*, **41**, 739–745.
- Gutman, I., Ruscic, B., Trinajstić, N. and Wilcox, C.F., Jr (1975) Graph theory and molecular orbitals. XII. Acyclic polyenes. *J. Chim. Phys.*, **62**, 3399–3405.
- Gutman, I. and Shalabi, A. (1984) Topological properties of benzenoid systems. Part XXIX. On Hosoya's topological index. *Z. Naturforsch.*, **39a**, 797–799.
- Gutman, I. and Soldatović, T. (2001) (n,m) -Type approximations for total π -electron energy of benzenoid hydrocarbons. *MATCH Commun. Math. Comput. Chem.*, **44**, 169–182.
- Gutman, I. and Soltés, L. (1991) The range of the Wiener index and its mean isomer degeneracy. *Z. Naturforsch.*, **46a**, 865–868.
- Gutman, I., Stanković, S., Durdević, J. and Furtula, B. (2007) On the cycle-dependence of topological resonance energy. *J. Chem. Inf. Model.*, **47**, 776–781.
- Gutman, I. and Tomović, Ž. (2000a) More on the line graph model for predicting physico-chemical properties of alkanes. *Acta Chim. Hung. -Mod. Chem.*, **137**, 439–445.
- Gutman, I. and Tomović, Ž. (2000b) On the application of line graphs in quantitative structure–property studies. *J. Serb. Chem. Soc.*, **65**, 577–580.
- Gutman, I. and Tomović, Ž. (2000c) Relation between distance-based topological indices. *J. Chem. Inf. Comput. Sci.*, **40**, 1333–1336.
- Gutman, I., Tomović, Ž., Mishra, B.K. and Kuanar, M. (2001) On the use of iterated line graphs in quantitative structure–property studies. *Indian J. Chem.*, **40**, 4–11.
- Gutman, I., Toropov, A.A. and Toropova, A.P. (2005) The graph of atomic orbitals and its basic properties. 1. Wiener index. *MATCH Commun. Math. Comput. Chem.*, **53**, 215–224.
- Gutman, I. and Trinajstić, N. (1972) Graph theory and molecular orbitals. Total π -electron energy of alternant hydrocarbons. *Chem. Phys. Lett.*, **17**, 535–538.
- Gutman, I. and Trinajstić, N. (1973a) Graph theory and molecular orbitals. *Top. Curr. Chem.*, **42**, 49–93.
- Gutman, I. and Trinajstić, N. (1973b) Graph theory and molecular orbitals. VIII. Kekulé structure and permutations. *Croat. Chem. Acta*, **45**, 539–545.
- Gutman, I. and Trinajstić, N. (1976) Graph theory and molecular orbitals. XVII. On the self-polarizability of the atom. *J. Chim. Phys.*, **65**, 3796–3797.
- Gutman, I. and Vidović, D. (2002a) The largest eigenvalues of adjacency and Laplacian matrices, and ionization potentials of alkanes. *Indian J. Chem.*, **41**, 893–896.
- Gutman, I. and Vidović, D. (2002b) Two early branching indices and the relation between them. *Theor. Chem. Acc.*, **108**, 98–102.
- Gutman, I., Vidović, D., Čmiljanović, N., Milosavljević, S. and Radenković, S. (2003a) Graph energy: a useful molecular structure-descriptor. *Indian J. Chem.*, **42**, 1309–1311.
- Gutman, I., Vidović, D. and Furtula, B. (2002a) Coulson function and Hosoya index. *Chem. Phys. Lett.*, **355**, 378–382.
- Gutman, I., Vidović, D. and Furtula, B. (2003b) Chemical applications of the Laplacian spectrum. VII. Studies of the Wiener and Kirchhoff indices. *Indian J. Chem.*, **42**, 1272–1278.

- Gutman, I., Vidović, D., Furtula, B. and Vesel, A. (2003c) Relations between topological indices of large chemical trees. *Indian J. Chem.*, **42**, 1241–1245.
- Gutman, I., Vidović, D., Furtula, B. and Zenkevich, I. G. (2003d) Wiener-type indices and internal molecular energy. *J. Serb. Chem. Soc.*, **68**, 401–408.
- Gutman, I., Vidović, D. and Hosoya, H. (2002b) The relation between the eigenvalue sum and the topological index Z revisited. *Bull. Chem. Soc. Jap.*, **75**, 1723–1727.
- Gutman, I., Vidović, D. and Nedić, A. (2002c) Ordering of alkane isomers by means of connectivity indices. *J. Serb. Chem. Soc.*, **67**, 87–97.
- Gutman, I., Vidović, D. and Popovic, L. (1998) Graph representation of organic molecules. Cayley's pterograms vs. his kenograms. *J. Chem. Soc. Faraday Trans.*, **94**, 857–860.
- Gutman, I., Vidović, D. and Stevanović, D. (2002d) Chemical applications of the Laplacian spectrum. VI. On the largest Laplacian eigenvalue of alkanes. *J. Serb. Chem. Soc.*, **67**, 407–413.
- Gutman, I., Vukicević, D. and Žerovnik, J. (2004) A class of modified Wiener indices. *Croat. Chem. Acta*, **77**, 103–109.
- Gutman, I., Yeh, Y.-N., Lee, S.-L. and Chen, J.-C. (1994) Wiener numbers of dendrimers. *MATCH Commun. Math. Comput. Chem.*, **30**, 103–115.
- Gutman, I., Yeh, Y.-N., Lee, S.-L. and Luo, Y.-L. (1993) Some recent results in the theory of the Wiener number. *Indian J. Chem.*, **32**, 651–661.
- Gutman, I. and Zenkevich, I.G. (2002) Wiener index and vibrational energy. *Z. Naturforsch.*, **57a**, 824–828.
- Gutman, I. and Zhou, B. (2006) Laplacian energy of a graph. *Linear Algebra and Its Applications*, **414**, 29–37.
- Gutman, I. and Žerovnik, J. (2002) Corroborating a modification of the Wiener index. *Croat. Chem. Acta*, **75**, 603–612.
- Gutmann, V. (1978) *The Donor–Acceptor Approach to Molecular Interactions*, Plenum Press, New York.
- Gutowsky, H.S., McCall, D.W., McGarvey, B.R. and Meyer, L.H. (1952) Electron distribution in molecules. I. ^{19}F nuclear magnetic shielding and substituent effects in some benzene derivatives. *J. Am. Chem. Soc.*, **74**, 4809–4817.
- Güzel, Y. (1996) Investigation of the relationship between the inhibitory activity of glycolic acid oxidase and its chemical structure: electron-topological approach. *J. Mol. Struct.*, **366**, 131–137.
- Güzel, Y., Saripinar, E. and Yilidrim, I. (1997) Electron-topological (ET) investigation of structure–antagonist activity of a series of dibenzo [a,d]cycloalkenimines. *J. Mol. Struct. (Theochem)*, **418**, 83–91.
- Ha, Z., Ring, Z. and Liu, S. (2005) Quantitative structure–property relationship (QSPR) models for boiling points, specific gravities, and refraction indices of hydrocarbons. *Energy & Fuels*, **19**, 152–163.
- Hadaruga, D.I., Muresan, S., Bologa, C., Chiriac, A., Simon, Z., Cofar, L. and Naray-Szabo, G. (1999) QSAR for cycloaliphatic alcohols with qualitatively defined sandalwood odour characteristics. *Quant. Struct. -Act. Relat.*, **18**, 253–261.
- Hadjipavloulitina, D. and Hansch, C. (1994) Quantitative structure–activity relationships of the benzodiazepines: a review and reevaluation. *Chem. Rev.*, **94**, 1483–1505.
- Hadzi, D. and Jerman-Blazic, B. (eds) (1987) *QSAR in Drug Design and Toxicology*, Elsevier, Amsterdam, The Netherlands.
- Hadzi, D., Kidrić, J., Koller, J. and Mavri, J. (1990) The role of hydrogen bonding in drug–receptor interactions. *J. Mol. Struct.*, **237**, 139–150.
- Haeberlein, M. and Brinck, T. (1997) Prediction of water–octanol partition coefficients using theoretical descriptors derived from the molecular surface area and the electrostatic potential. *J. Chem. Soc. Perkin Trans. 2*, 289–294.
- Hagadone, T.R. (1992) Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.*, **32**, 515–521.
- Hage, P. and Harary, F. (1995) Eccentricity and centrality in networks. *Social Networks*, **17**, 57–63.
- Haggarty, S.J., Clemons, P.A. and Schreiber, S.L. (2003) Chemical genomic profiling of biological networks using graph theory and combinations of small molecule perturbations. *J. Am. Chem. Soc.*, **125**, 10543–10545.
- Hahn, M. (1995) Receptor surface models. 1. Definition and construction. *J. Med. Chem.*, **38**, 2080–2090.
- Hahn, M. (1997) Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci.*, **37**, 80–86.
- Hahn, M. and Rogers, D. (1995) Receptor surface models. 2. Application to quantitative structure–activity relationships studies. *J. Med. Chem.*, **38**, 2091–2102.
- Hahn, M. and Rogers, D. (1998) Receptor surface models, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ ESCOM, Dordrecht, The Netherlands, pp. 117–133.
- Hajduk, P.J., Mendoza, R., Petros, A.M., Huth, J.R., Bures, M., Fesik, S.W. and Martin, Y.C. (2003)

- Ligand binding to domain-3 of human serum albumin: a chemometric analysis. *J. Comput. Aid. Mol. Des.*, **17**, 93–102.
- Hakimi, S.L. and Yau, S.S. (1965) Distance matrix of a graph and its realizability. *Quarterly Applied Mathematics*, **12**, 305–317.
- Halfon, E. (1989) Comparison of an index function and a vectorial approach method for ranking of waste disposal sites. *Environ. Sci. Technol.*, **23**, 600–609.
- Halfon, E., Galassi, S., Brüggemann, R. and Provini, A. (1996) Selection of priority properties to assess environmental hazard of pesticides. *Chemosphere*, **33**, 1543–1562.
- Halfon, E. and Reggiani, M.G. (1986) On ranking chemicals for environmental hazard. *Environ. Sci. Technol.*, **20**, 1173–1179.
- Hall, G.G. (1955) The bond orders of alternant hydrocarbons molecules. *Proc. Roy. Soc. London A*, **229**, 251–259.
- Hall, G.G. (1957) The bond orders of some conjugated molecules. *Trans. Faraday Soc.*, **53**, 573–581.
- Hall, G.G. (1981) Eigenvalues of molecular graphs. *Bull. Inst. Math. Appl.*, **17**, 70–72.
- Hall, G.G. (1986) The evaluation of moments for polycyclic hydrocarbons. *Theor. Chim. Acta*, **70**, 323–332.
- Hall, G.G. (1992) Eigenvalue distributions for the graphs of alternant hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **32**, 11–13.
- Hall, G.G. (1993) Eigenvalue distributions in alternant hydrocarbons. *J. Math. Chem.*, **13**, 191–203.
- Hall, L.H. (1990) Computational aspects of molecular connectivity and its role in structure–property modeling, in *Computational Chemical Graph Theory* (ed. D.H. Rouvray), Nova Science Publishers, New York, pp. 202–233.
- Hall, L.H. (1995) Experimental design in synthesis planning and structure–property correlations. total response surface optimization, in *Chemometrics Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 91–102.
- Hall, L.H. and Aaserud, D. (1989) Structure–activity models for molar refraction of alkylsilanes based on molecular connectivity. *Quant. Struct. -Act. Relat.*, **8**, 296–304.
- Hall, L.H., Dailey, R.S. and Kier, L.B. (1993) Design of molecules from quantitative structure–activity relationship models. 3. Role of higher order path counts: path 3. *J. Chem. Inf. Comput. Sci.*, **33**, 598–603.
- Hall, L.H. and Kier, L.B. (1977a) A molecular connectivity study of electron density in alkanes. *Tetrahedron*, **33**, 1953–1957.
- Hall, L.H. and Kier, L.B. (1977b) Structure–activity studies using valence molecular connectivity. *J. Pharm. Sci.*, **66**, 642–644.
- Hall, L.H. and Kier, L.B. (1978a) A comparative analysis of molecular connectivity, Hansch, Free–Wilson and Darc–Pelco methods in the SAR of halogenated phenols. *Eur. J. Med. Chem.*, **13**, 89–92.
- Hall, L.H. and Kier, L.B. (1978b) Molecular connectivity and substructure analysis. *J. Pharm. Sci.*, **67**, 1743–1747.
- Hall, L.H. and Kier, L.B. (1981) The relation of molecular connectivity to molecular volume and biological activity. *Eur. J. Med. Chem.*, **16**, 399–407.
- Hall, L.H. and Kier, L.B. (1984) Molecular connectivity of phenols and their toxicity to fish. *Bull. Environ. Contam. Toxicol.*, **32**, 354–362.
- Hall, L.H. and Kier, L.B. (1986) Molecular connectivity and total response surface optimization. *J. Mol. Struct. (Theochem)*, **134**, 309–316.
- Hall, L.H. and Kier, L.B. (1990) Determination of topological equivalence in molecular graphs from the topological state. *Quant. Struct. -Act. Relat.*, **9**, 115–131.
- Hall, L.H. and Kier, L.B. (1991) The molecular connectivity chi indexes and kappa shape indexes in structure–property modeling, in *Reviews in Computational Chemistry*, Vol. 2 (eds K.B. Lipkowitz and D.B. Boyd), VCH Publishers, New York, pp. 367–422.
- Hall, L.H. and Kier, L.B. (1992a) Binding of salicylamides: QSAR analysis with electrotopological state indexes. *Med. Chem. Res.*, **2**, 497–502.
- Hall, L.H. and Kier, L.B. (1992b) *Enumeration, Topological Indexes and Molecular Properties in Alkanes* (eds S. Patei and Z. Rapoport), John Wiley & Sons, Ltd, Chichester, UK, pp. 186–213.
- Hall, L.H. and Kier, L.B. (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.*, **35**, 1039–1045.
- Hall, L.H. and Kier, L.B. (2000) The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.*, **40**, 784–791.
- Hall, L.H. and Kier, L.B. (2001) Issues in representation of molecular structure. The development of molecular connectivity. *J. Mol. Graph. Model.*, **20**, 4–18.
- Hall, L.H., Kier, L.B. and Brown, B.B. (1995) Molecular similarity based on novel atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.*, **35**, 1074–1080.

- Hall, L.H., Kier, L.B. and Frazer, J.W. (1993) Design of molecules from quantitative structure–activity relationship models. 2. Derivation and proof of information transfer relating equations. *J. Chem. Inf. Comput. Sci.*, **33**, 148–152.
- Hall, L.H., Kier, L.B. and Murray, W.J. (1975) Molecular connectivity. II. Relationship to water solubility and boiling point. *J. Pharm. Sci.*, **64**, 1974–1977.
- Hall, L.H., Maynard, E.L. and Kier, L.B. (1989) QSAR investigation of benzene toxicity to fathead minnow using molecular connectivity. *Environ. Toxicol. Chem.*, **8**, 783–788.
- Hall, L.H., Mohney, B. and Kier, L.B. (1991a) The electrotopological state: an atom index for QSAR. *Quant. Struct. -Act. Relat.*, **10**, 43–48.
- Hall, L.H., Mohney, B. and Kier, L.B. (1991b) The electrotopological state: structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.*, **31**, 76–82.
- Hall, L.H., Mohney, B. and Kier, L.B. (1993) Comparison of electrotopological state indexes with molecular orbital parameters: inhibition of MAO by hydrazides. *Quant. Struct. -Act. Relat.*, **12**, 44–48.
- Hall, L.H. and Story, C.T. (1996) Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.*, **36**, 1004–1014.
- Hall, L.H. and Story, C.T. (1997) Boiling point of a set of alkanes, alcohols and chloroalkanes: QSAR with atom type electrotopological state indices using artificial neural networks. *SAR & QSAR Environ. Res.*, **6**, 139–161.
- Hall, L.H. and Vaughn, T.A. (1997) QSAR of phenol toxicity using E-state and kappa shape indices. *Med. Chem. Res.*, **7**, 407–416.
- Hall, L.M., Hall, L.H. and Kier, L.B. (2003) QSAR modeling of β -lactam binding to human serum proteins. *J. Comput. Aid. Mol. Des.*, **17**, 103–118.
- Halova, J., Strouf, O., Zak, P., Sochozova, A., Uchida, N., Yuzuri, T., Sakakibara, K. and Hirota, M. (1998) QSAR of catechol analogs against malignant melanoma using fingerprint descriptors. *Quant. Struct. -Act. Relat.*, **17**, 37–39.
- Ham, N.S. (1958) Mobile bond orders in the resonance and molecular orbital theories. *J. Chim. Phys.*, **29**, 1229–1231.
- Ham, N.S. and Ruedenberg, K. (1958a) Energy levels, atom populations, bond populations in the LCAO MO model and in the FE MO model. A quantitative analysis. *J. Chim. Phys.*, **29**, 1199–1214.
- Ham, N.S. and Ruedenberg, K. (1958b) Mobile bond orders in conjugated systems. *J. Chim. Phys.*, **29**, 1215–1229.
- Hamerton, I., Howlin, B.J. and Larwood, V. (1995) Development of quantitative structure–property relationships for poly(arylene ether)s. *J. Mol. Graph.*, **13**, 14–17.
- Hammett, L.P. (1935) Reaction rates and indicator acidities. *Chem. Rev.*, **17**, 67–79.
- Hammett, L.P. (1937) The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.*, **59**, 96–103.
- Hammett, L.P. (1938) Linear free energy relationships in rate and equilibrium phenomena. *Trans. Faraday Soc.*, **34**, 156–165.
- Hammett, L.P. (1940) *Physical Organic Chemistry*, McGraw-Hill, New York.
- Hammett, L.P. (1970) *Physical Organic Chemistry: Reaction Rates, Equilibria and Mechanism*, McGraw-Hill, New York.
- Hamori, E. (1983) H curves, a novel method of representation of nucleotides series especially suited for long DNA sequences. *J. Biol. Chem.*, **258**, 1318–1327.
- Hamori, E. (1985) Novel DNA sequence representation. *Nature*, **314**, 585–586.
- Hamori, E. (1989) Graphical representation of long DNA sequences by methods of H curves, current results and future aspects. *BioTechniques*, **7**, 710–720.
- Han, C.R. (1990) The calculation of the ionic group electronegativities and neutral group electronegativities. *Acta Chim. Sin.*, **48**, 627–631.
- Hancock, C.K. and Falls, C.P. (1961) A Hammett–Taft polar-steric equation for the saponification rates of *m*- and *p*-substituted alkyl benzoates. *J. Am. Chem. Soc.*, **83**, 4214–4216.
- Hancock, C.K., Meyers, E.A. and Yager, B.J. (1961) Quantitative separation of hyperconjugation effects from steric substituent constants. *J. Am. Chem. Soc.*, **83**, 4211–4213.
- Hancock, T., Put, R., Coomans, D., Vander Heyden, Y. and Everingham, Y. (2005) A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. *Chemom. Intell. Lab. Syst.*, **76**, 185–196.
- Hand, D.J. (1981) *Discrimination and Classification*, John Wiley & Sons, Ltd, Chichester, UK.
- Hand, D.J. (1997) *Construction and Assessment of Classification Rules*, John Wiley & Sons, Ltd, Chichester, UK, p. 214.
- Handschuh, S., Wagener, M. and Gasteiger, J. (1998) Superimposition of three-dimensional chemical

- structures allowing for conformational flexibility by a hybrid method. *J. Chem. Inf. Comput. Sci.*, **38**, 220–232.
- Hann, M.M., Leach, A.R. and Harper, G. (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.*, **41**, 856–864.
- Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.*, **8**, 255–263.
- Hannay, N.B. and Smyth, C.P. (1946) The dipole moment of hydrogen fluoride and the ionic character of bonds. *J. Am. Chem. Soc.*, **68**, 171–173.
- Hannongbua, S., Lawtrakul, L. and Limtrakul, J. (1996a) Structure–activity correlation study of HIV-1 inhibitors: electronic and molecular parameters. *J. Comput. Aid. Mol. Des.*, **10**, 145–152.
- Hannongbua, S., Lawtrakul, L., Sottriffer, C.A. and Rode, B.M. (1996b) Comparative molecular field analysis of HIV-1 reverse transcriptase inhibitors in the class of 1((2-hydroxyethoxy)-methyl)-6-(phenylthio)thymine. *Quant. Struct. -Act. Relat.*, **15**, 389–394.
- Hannongbua, S., Pungpo, P., Limtrakul, J. and Wolschann, P. (1999) Quantitative structure–activity relationships and comparative molecular field analysis of TIBO derivatised HIV-1 reverse transcriptase inhibitors. *J. Comput. Aid. Mol. Des.*, **13**, 563–577.
- Hansch, C. (1969) Quantitative approach to biochemical structure–activity relationships. *Acc. Chem. Res.*, **2**, 232–239.
- Hansch, C. (1970) Steric parameters in structure–activity correlations. Cholinesterase inhibitors. *J. Org. Chem.*, **35**, 620–621.
- Hansch, C. (1971) Quantitative structure–activity relationships in drug design, in *Drug Design*, Vol. 1 (ed. E.J. Ariens), Academic Press, New York, pp. 271–342.
- Hansch, C. (1978) Recent advances in biochemical QSAR, in *Correlation Analysis in Chemistry* (eds N. B. Chapman and J. Shorter), Plenum Press, New York, pp. 397–438.
- Hansch, C. (1993) Quantitative structure–activity relationships and the unnamed science. *Acc. Chem. Res.*, **2**, 147–153.
- Hansch, C. (1995a) Comparative QSAR understanding hydrophobic interactions. *ACS Symp. Ser.*, **606**, 254–262.
- Hansch, C. (1995b) Comparative quantitative structure–activity relationship insect versus vertebrate cholinesterase. *ACS Symp. Ser.*, **589**, 281–291.
- Hansch, C. and Anderson, S.M. (1967) Structure–activity relation in barbiturates and its similarity to that in other narcotics. *J. Math. Chem.*, **10**, 745–753.
- Hansch, C. and Calef, D.F. (1976) Structure–activity relationships in papain–ligand interactions. *J. Org. Chem.*, **41**, 1240–1243.
- Hansch, C. and Clayton, J.M. (1973) Lipophilic character and biological activity of drugs. II. The parabolic case. *J. Pharm. Sci.*, **62**, 1–21.
- Hansch, C., Deutsch, E.W. and Smith, R.N. (1965) The use of substituent constants and regression analysis in the study of enzymatic reaction mechanisms. *J. Am. Chem. Soc.*, **87**, 2738–2742.
- Hansch, C. and Dunn, W.J. III (1972) Linear relationships between lipophilic character and biological activity of drugs. *J. Pharm. Sci.*, **61**, 1–19.
- Hansch, C. and Fujita, T. (1964) ρ – σ – π analysis, a method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, **86**, 1616–1626.
- Hansch, C. and Fujita, T. (1995) Status of QSAR at the end of the twentieth century, in *Classical and 3D-QSAR in Agrochemistry, ACS Symposium Series 606* (eds C. Hansch and T. Fujita), American Chemical Society, Washington, DC, pp. 1–11.
- Hansch, C. and Gao, H. (1997) Comparative QSAR: radical reactions of benzene derivatives in chemistry and biology. *Chem. Rev.*, **97**, 2995–3060.
- Hansch, C., Gao, H. and Hoekman, D. (1998) A generalized approach to comparative QSAR, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 285–368.
- Hansch, C., Hoekman, D. and Gao, H. (1996) Comparative QSAR toward a deeper understanding of chemicobiological interactions. *Chem. Rev.*, **96**, 1045–1075.
- Hansch, C., Hoekman, D., Leo, A., Zhang, L.T. and Li, P. (1995) The expanding role of quantitative structure–activity relationships (QSAR) in toxicology. *Toxicol. Lett.*, **79**, 45–53.
- Hansch, C., Kim, K.H. and Sarma, R.H. (1973) Structure–activity relationship in benzamides inhibiting alcohol dehydrogenase. *J. Am. Chem. Soc.*, **95**, 6447–6449.
- Hansch, C. and Kurup, A. (2003) QSAR of chemical polarizability and nerve toxicity. 2. *J. Chem. Inf. Comput. Sci.*, **43**, 1647–1651.
- Hansch, C., Kurup, A., Garg, R. and Gao, H. (2001) Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem. Rev.*, **101**, 619–672.

- Hansch, C., Kutter, E. and Leo, A. (1969) Homolytic constants in the correlation of chloramphenicol structure with activity. *J. Med. Chem.*, **12**, 746–749.
- Hansch, C. and Leo, A. (1979) *Substituent Constants for Correlation Analysis in Chemistry and Biology*, John Wiley & Sons, Inc., New York, p. 352.
- Hansch, C. and Leo, A. (1995) Exploring QSAR, Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC.
- Hansch, C., Leo, A. and Hoekman, D. (1995) Exploring QSAR. Hydrophobic, Electronic, and Steric Constants, American Chemical Society, Washington, DC, p. 557.
- Hansch, C., Leo, A. and Nikaitani, D. (1972) On the additive–constitutive character of partition coefficients. *J. Org. Chem.*, **37**, 3090–3092.
- Hansch, C., Leo, A. and Taft, R.W. (1991) A survey of Hammett substituent constants and resonance and field parameters. *Chem. Rev.*, **91**, 165–195.
- Hansch, C., Leo, A., Unger, S.H., Kim, K.H., Nikaitani, D. and Lien, E.J. (1973) “Aromatic” substituent constants for structure–activity correlations. *J. Med. Chem.*, **16**, 1207–1216.
- Hansch, C. and Lien, E.J. (1968) Analysis of the structure–activity relationship in the adrenergic blocking activity of β -haloalkylamines. *Biochem. Pharmacol.*, **17**, 709–720.
- Hansch, C. and Lien, E.J. (1971) Structure–activity relationships in antifungal agents. A survey. *J. Med. Chem.*, **14**, 653–670.
- Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M. (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, **194**, 178–180.
- Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, F. and Streich, M. (1963) The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *J. Am. Chem. Soc.*, **85**, 2817–2824.
- Hansch, C., Quinlan, J.E. and Lawrence, G.L. (1968) The LFER between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.*, **33**, 347–350.
- Hansch, C., Rockwell, S.D., Leo, A. and Steller, E.E. (1977) Substituent constants for correlation analysis. *J. Med. Chem.*, **20**, 304–306.
- Hansch, C., Steinmetz, W.E., Leo, A.J., Mekapati, S. B., Kurup, A. and Hoekman, D. (2003) On the role of polarizability in chemical–biological interactions. *J. Chem. Inf. Comput. Sci.*, **43**, 120–125.
- Hansch, C., Telzer, B.R. and Zhang, L.T. (1995) Comparative QSAR in toxicology: examples from teratology and cancer chemotherapy of aniline mustards. *Crit. Rev. Toxicol.*, **25**, 67–89.
- Hansch, C., Unger, S.H. and Forsythe, A.B. (1973) Strategy in drug design. Cluster analysis as an aid in the selection of substituents. *J. Med. Chem.*, **16**, 1217–1222.
- Hansch, C. and Yoshimoto, M. (1974) Structure–activity relationships in immunochemistry. 2. Inhibition of complement by benzamides. *J. Med. Chem.*, **17**, 1160–1167.
- Hansch, C. and Zhang, L.T. (1992) QSAR of HIV inhibitors. *Bioorg. Med. Chem. Lett.*, **2**, 1165–1169.
- Hansen, B.G., Paya-Perez, A.B., Rahman, M. and Larsen, B.R. (1999) QSARs for K_{OW} and K_{OC} of PCB congeners: a critical examination of data, assumptions and statistical approaches. *Chemosphere*, **39**, 2209–2228.
- Hansen, P. and Mélot, H. (2003) Variable neighborhood search for extremal graphs. 6. Analyzing bounds for the connectivity index. *J. Chem. Inf. Comput. Sci.*, **43**, 1–14.
- Hansen, P. and Vukicević, D. (2007) Comparing the Zagreb indices. *Croat. Chem. Acta*, **80**, 165–168.
- Hansen, P. and Zheng, M. (1994) Bonds fixed by fixing bonds. *J. Chem. Inf. Comput. Sci.*, **34**, 297–304.
- Hansen, P.J. and Jurs, P.C. (1987) Prediction of olefin boiling points from molecular structure. *Anal. Chem.*, **59**, 2322–2327.
- Hansen, P.J. and Jurs, P.C. (1988a) Chemical applications of graph theory. Part I. Fundamentals and topological indices. *J. Chem. Educ.*, **65**, 574–580.
- Hansen, P.J. and Jurs, P.C. (1988b) Chemical applications of graph theory. Part II. Isomer enumeration. *J. Chem. Educ.*, **65**, 661–664.
- Hanser, T., Jauffret, P. and Kaufmann, G. (1996) A new algorithm for exhaustive ring perception in a molecular graph. *J. Chem. Inf. Comput. Sci.*, **36**, 1146–1152.
- Hansson, G. and Ahnoff, M. (1994) Chromatographic separation of amide diastereomers correlation with molecular descriptors. *J. Chromat.*, **666**, 505–517.
- Harada, A., Hanzawa, M., Saito, J. and Hashimoto, K. (1992) Quantitative analysis of structure–toxicity relationships of substituted anilines by use of BALB/3T3 cells. *Environ. Toxicol. Chem.*, **11**, 973–980.
- Haranczyk, M. and Holliday, J.D. (2008) Comparison of similarity coefficients for clustering and compound selection. *J. Chem. Inf. Model.*, **48**, 498–508.

- Harary, F. (1959) Status and contrastatus. *Sociometry*, **22**, 23–43.
- Harary, F. (1964) Combinatorial problems in graphical enumeration, in *Applied Combinatorial Mathematics* (ed. E.F. Beckenbach), John Wiley & Sons, Inc., New York, pp. 185–220.
- Harary, F. (1969a) *Graph Theory*, Addison-Wesley, Reading, MA.
- Harary, F. (1969b) *Proof Techniques in Graph Theory*, Academic Press, San Diego, CA.
- Harary, F., King, C., Mowshowitz, A. and Read, R.C. (1971) Cospectral graphs and digraphs. *Bull. London Math. Soc.*, **3**, 321–328.
- Harju, M., Andersson, P.L., Haglund, P. and Tysklind, M. (2002) Multivariate physico-chemical characterisation and quantitative structure–property relationship modelling of polybrominated diphenyl ethers. *Chemosphere*, **47**, 375–384.
- Harper, G., Bradshaw, J., Gittins, J.C., Green, D.V.S. and Leach, A.R. (2001) Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.*, **41**, 1295–1300.
- Harrington, E.C., Jr (1965) The desirability function. *Ind. Qual. Control*, **21**, 494–498.
- Harris, N.V., Smith, C. and Bowden, K. (1992) Antifolate and antibacterial activities of 6-substituted 2,4-diaminoquinazolines. *Eur. J. Med. Chem.*, **27**, 7–18.
- Harrison, A.G., Kebarle, P. and Lossing, F.P. (1961) Free radicals by mass spectrometry. XXI. The ionization potentials of some *meta* and *para* substituted benzyl radicals. *J. Am. Chem. Soc.*, **83**, 777–780.
- Hartley, R.V.L. (1928) Transmission of information. *Bell Syst. Tech. J.*, **7**, 535–563.
- Hasegawa, K., Arakawa, M. and Funatsu, K. (1999) 3D-QSAR study of insecticidal neonicotinoid compounds based on 3-way partial least squares model. *Chemom. Intell. Lab. Syst.*, **47**, 33–40.
- Hasegawa, K., Arakawa, M. and Funatsu, K. (2003) Simultaneous determination of bioactive conformations and alignment rules by multi-way PLS modeling. *Comp. Biol. Chem.*, **27**, 211.
- Hasegawa, K., Deushi, T., Yaegashi, O., Miyashita, Y. and Sasaki, S. (1995) Artificial neural network studies in quantitative structure–activity relationships of antifungal azoxy compounds. *Eur. J. Med. Chem.*, **30**, 569–574.
- Hasegawa, K. and Funatsu, K. (1998) GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. *J. Mol. Struct. (Theochem)*, **425**, 255–262.
- Hasegawa, K., Kimura, T. and Funatsu, K. (1999) GA strategy for variable selection in QSAR studies: application of GA-based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors. *J. Chem. Inf. Comput. Sci.*, **39**, 112–120.
- Hasegawa, K., Kimura, T., Miyashita, Y. and Funatsu, K. (1996) Nonlinear partial least squares modeling of phenyl alkylamines with the monoamine oxidase inhibitory activities. *J. Chem. Inf. Comput. Sci.*, **36**, 1025–1029.
- Hasegawa, K., Matsuoka, S., Arakawa, M. and Funatsu, K. (2002) New molecular surface-based 3D-QSAR method using Kohonen neural network and 3-way PLS. *Computers Chem.*, **26**, 583–589.
- Hasegawa, K., Matsuoka, S., Arakawa, M. and Funatsu, K. (2003) Multi-way PLS modeling of structure–activity data by incorporating electrostatic and lipophilic potentials on molecular surface. *Comp. Biol. Chem.*, **27**, 381–386.
- Hasegawa, K., Miyashita, Y. and Funatsu, K. (1997) GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.*, **37**, 329–334.
- Hasegawa, K., Morikami, K., Shiratori, Y., Ohtsuka, T., Aoki, Y. and Shimma, N. (2003) 3D-QSAR study of antifungal *N*-myristoyltransferase inhibitors by comparative molecular surface analysis. *Chemom. Intell. Lab. Syst.*, **69**, 51–59.
- Hasegawa, K., Shigyou, H. and Sonoki, H. (1995) Free–Wilson discriminant analysis of antiarrhythmic phenyl-pyridines using PLS. *Quant. Struct. -Act. Relat.*, **14**, 344–347.
- Hasegawa, K., Yokoo, N., Watanabe, K., Hirata, M., Miyashita, Y. and Sasaki, S. (1996) Multivariate Free–Wilson analysis of alpha chymotrypsin inhibitors using Pls. *Chemom. Intell. Lab. Syst.*, **33**, 63–69.
- Hassan, M., Bielawski, J.P., Hempel, J.C. and Waldman, M. (1996) Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Div.*, **2**, 64–74.
- Hassan, M., Brown, R.D., Varma-O'Brien, S. and Rogers, D. (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol. Div.*, **10**, 283–299.
- Hasse, H. (1952) *Über die klassenzahl abelscher Zahlkörper*, Akademie Verlag, Berlin, Germany.
- Hatch, F.T., Colvin, M.E. and Seidl, E.T. (1996) Structural and quantum chemical factors affecting mutagenic potency of aminoimidazo azaarenes. *Envir. Mol. Mutag.*, **27**, 314–330.
- Hatrik, S. and Zahradnik, P. (1996) Neural network approach to the prediction of the toxicity of

- benzothiazolium salts from molecular structure. *J. Chem. Inf. Comput. Sci.*, **36**, 992–995.
- Havelec, P. and Sevcik, J.G. (1996) Extended additivity model of parameter log(L16). *J. Phys. Chem. Ref. Data*, **25**, 1483–1439.
- Hawkins, D.M. (2004) The problem of overfitting. *J. Chem. Inf. Comput. Sci.*, **44**, 1–12.
- Hawkins, D.M., Basak, S.C. and Mills, D. (2003) Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.*, **43**, 579–586.
- Hawkins, D.M., Basak, S.C. and Mills, D. (2004) QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Envir. Toxicol. Pharmacol.*, **16**, 37–44.
- Hawkins, D.M., Basak, S.C. and Shi, X. (2001) QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.*, **41**, 663–670.
- Hays, S.J., Rice, M.J., Ortwine, D.F., Johnson, G., Schwarz, R.D., Boyd, D.K., Copeland, L.F., Vartanian, M.G. and Boxer, P.A. (1994) Substituted 2-benzothiazolamine as sodium flux inhibitors quantitative structure–activity relationships and anticonvulsant activity. *J. Pharm. Sci.*, **83**, 1425–1432.
- He, L. and Jurs, P.C. (2005) Assessing the reliability of a QSAR model's predictions. *J. Mol. Graph. Model.*, **23**, 503–523.
- He, L., Jurs, P.C., Custer, L.L., Durham, S.K. and Pearl, G.M. (2003) Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers. *Chem. Res. Toxicol.*, **16**, 1567–1580.
- He, L., Jurs, P.C., Kreatsoulas, C., Custer, L.L., Durham, S.K. and Pearl, G.M. (2005) Probabilistic neural network multiple classifier system for predicting the genotoxicity of quinolone and quinoline derivatives. *Chem. Res. Toxicol.*, **18**, 428–440.
- He, P.-A. and Wang, J. (2002) Characteristic sequences for DNA primary sequence. *J. Chem. Inf. Comput. Sci.*, **42**, 1080–1085.
- Head, R.D., Smythe, M.L., Oprea, T.I., Waller, C.L., Green, S.M. and Marshall, G.R. (1996) VALIDATE: a new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.*, **118**, 3959–3969.
- Headley, A.D., Starnes, S.D., Cheung, E.T. and Malone, P.L. (1995) Solvation effects on the relative basicity of propylamines. *J. Phys. Org. Chem.*, **8**, 26–30.
- Headley, A.D., Starnes, S.D., Wilson, L.Y. and Famini, G.R. (1994) Analysis of solute/solvent interactions for the acidity of acetic acids by theoretical descriptors. *J. Org. Chem.*, **59**, 8040–8046.
- Héberger, K. and Andrade, J.M. (2004) Procrustes rotation and pair-wise correlation: a parametric and a non-parametric method for variable selection. *Croat. Chem. Acta*, **77**, 117–125.
- Héberger, K. and Borosy, A.P. (1999) Comparison of chemometric methods for prediction of rate constants and activation energies of radical addition reactions. *J. Chemom.*, **13**, 473–489.
- Héberger, K. and Rajkó, R. (1997) Discrimination of statistically equivalent variables in quantitative structure–activity relationships, in *Quantitative Structure–Activity Relationships in Environmental Sciences VII* (eds F. Chen and G. Schüürmann), Society of Environmental Toxicology and Chemistry (SETAC), Pensacola, FL, pp. 425–433.
- Héberger, K. and Rajkó, R. (2002a) Generalization of pair correlation method (PCM) for non-parametric variable selection. *J. Chemom.*, **16**, 436–443.
- Héberger, K. and Rajkó, R. (2002b) Variable selection using pair-correlation method. *Environmental applications. SAR & QSAR Environ. Res.*, **13**, 541–554.
- Hefferlin, R.A. and Matus, M.T. (2001) Molecular similarity for small species: refining the isoelectronic index. *J. Chem. Inf. Comput. Sci.*, **41**, 484–494.
- Heiden, W., Moeckel, G. and Brickmann, K. (1993) A new approach to analysis and display lipophilicity/hydrophilicity mapped on molecular surface. *J. Comput. Aid. Mol. Des.*, **7**, 503–514.
- Heimstad, E.S. and Andersson, P.L. (2002) Docking and QSAR studies of an indirect estrogenic effect of hydroxylated PCBs. *Quant. Struct. -Act. Relat.*, **21**, 257–266.
- Heinzen, V.E.F., Filho, V.C. and Yunes, R.A. (1999) Correlation of activity of 2-(X-benzyloxy)-4,6-dimethoxyacetophenones with topological indices and with the Hansch equation. *Il Farmaco*, **54**, 125–129.
- Heinzen, V.E.F., Soares, M.F. and Yunes, R.A. (1999) Semi-empirical topological method for the prediction of the chromatographic retention of *cis*- and *trans*-alkene isomers and alkanes. *J. Chromat.*, **849**, 495–506.
- Heinzen, V.E.F. and Yunes, R.A. (1996) Using topological indices in the prediction of gas chromatographic retention indices of linear alkylbenzene isomers. *J. Chromat.*, **719**, 462–467.
- Helguera Morales, A., Cabrera Pérez, M.A., Combes, R.D. and Pérez González, M. (2006) Quantitative structure–activity relationship for the computational prediction of nitrocompounds carcinogenicity. *Toxicology*, **220**, 51–62.

- Helguera Morales, A., Duchowicz, P.R., Cabrera Pérez, M.A., Castro, E.A., Dias Soeiro Cordeiro, M. N. and Pérez González, M. (2006) Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. *Chemom. Intell. Lab. Syst.*, **81**, 180–187.
- Helguera Morales, A., Perez, M.A.C. and Pérez González, M. (2006) A radial distribution-function approach for predicting rodent carcinogenicity. *J. Mol. Model.*, **12**, 769–780.
- Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S. (1987a) Peptide quantitative structure–activity relationships, a multivariate approach. *J. Med. Chem.*, **30**, 1126–1135.
- Hellberg, S., Sjöström, M., Wikström, C. and Wold, S. (1987b) Peptide QSAR with SIMCA and PLS. *Pharmacochemistry Library*, **10**, 255–262.
- Hellberg, S., Sjöström, M. and Wold, S. (1986) The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure–activity relationship. *Acta Chem. Scand.*, **40**, 135–140.
- Hemken, H.G. and Lehmann, P.A. (1992) The use of computerized molecular structure scanning and principal component analysis to calculate molecular descriptors for QSAR. *Quant. Struct. - Act. Relat.*, **11**, 332–338.
- Hemmateenejad, B. (2004) Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based PCR. *J. Chemom.*, **18**, 475–485.
- Hemmateenejad, B. (2005) Correlation ranking procedure for factor selection in PC-ANN modeling and application to ADMETox evaluation. *Chemom. Intell. Lab. Syst.*, **75**, 231–245.
- Hemmateenejad, B., Akhond, M., Miri, R. and Shamsipur, M. (2003) Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1,4-dihydropyridines (nifedipine analogous). *J. Chem. Inf. Comput. Sci.*, **43**, 1328–1334.
- Hemmateenejad, B., Miri, R., Akhond, M. and Shamsipur, M. (2002) QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods. *Chemom. Intell. Lab. Syst.*, **64**, 91–99.
- Hemmateenejad, B., Miri, R., Jafarpour, M., Tabarzad, M. and Foroumadi, A. (2006) Multiple linear regression and principal component analysis-based prediction of the anti-tuberculosis activity of some 2-aryl-1,3,4-thiadiazole derivatives. *QSAR Comb. Sci.*, **25**, 56–66.
- Hemmateenejad, B., Miri, R., Tabarzad, M., Jafarpour, M. and Zand, F. (2004) Molecular modeling and QSAR analysis of the anticonvulsant activity of some *N*-phenyl-*N*⁰-(4-pyridinyl)-urea derivatives. *J. Mol. Struct. (Theochem)*, **684**, 43–49.
- Hemmateenejad, B., Safarpour, M.A., Miri, R. and Nesari, N. (2005) Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs. *J. Chem. Inf. Model.*, **45**, 190–199.
- Hemmateenejad, B., Safarpour, M.A. and Taghavi, F. (2003) Application of *ab initio* theory for the prediction of acidity constants of some 1-hydroxy-9,10-anthraquinone derivatives using genetic neural network. *J. Mol. Struct. (Theochem)*, **635**, 183–190.
- Hemmer, M. (2003) Expert systems, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1281–1299.
- Hemmer, M.C., Steinhauer, V. and Gasteiger, J. (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrat. Spect.*, **19**, 151–164.
- Hendrickson, J.B., Huang, P. and Toczko, A.G. (1987) Molecular complexity: a simplified formula adapted to individual atoms. *J. Chem. Inf. Comput. Sci.*, **27**, 63–67.
- Hendrickson, J.B. and Toczko, A.G. (1983) Unique numbering and cataloging of molecular structures. *J. Chem. Inf. Comput. Sci.*, **23**, 171–177.
- Hendriks, M.M.W.B., de Boer, J.H., Smilde, A.K. and Doornbos, D.A. (1992) Multicriteria decision making. *Chemom. Intell. Lab. Syst.*, **16**, 175–191.
- Henrie, I.I.R.N., Plummer, M.J., Smith, S.E., Yeager, W.H. and Witkowski, D.A. (1993) Discovery and optimization of a PSI electron-accepting 1,2,4-benzotriazine herbicide. *Quant. Struct. - Act. Relat.*, **12**, 27–37.
- Henry, D.R. and Block, J.H. (1979) Classification of drugs by discriminant analysis using fragment molecular connectivity values. *J. Med. Chem.*, **22**, 465–472.
- Henry, D.R. and Block, J.H. (1980a) Pattern recognition of steroids using fragment molecular connectivity. *J. Pharm. Sci.*, **69**, 1030–1034.
- Henry, D.R. and Block, J.H. (1980b) Steroid classification by discriminant analysis using fragment molecular connectivity. *Eur. J. Med. Chem.*, **15**, 133–138.
- Henry, D.R., Jurs, P.C. and Denny, W.A. (1982) Structure–antitumor activity relationships of 9-

- anilinoacridines using pattern recognition. *J. Math. Chem.*, **25**, 899–908.
- Henze, H.R. and Blair, C.M. (1931a) The number of isomeric alcohols of the methanol series. *J. Am. Chem. Soc.*, **53**, 3042–3046.
- Henze, H.R. and Blair, C.M. (1931b) The number of isomeric hydrocarbons of the methane series. *J. Am. Chem. Soc.*, **53**, 3077–3085.
- Henze, H.R. and Blair, C.M. (1933) The number of structurally isomeric hydrocarbons of the ethylene series. *J. Am. Chem. Soc.*, **55**, 680–686.
- Henze, H.R. and Blair, C.M. (1934) The number of structural isomers of the more important types of aliphatic compounds. *J. Am. Chem. Soc.*, **56**, 157.
- Herdan, J., Balaban, A.T., Stoica, G., Simon, Z., Mracec, M. and Niculescuduvaz, I. (1991) Compounds with potential cancer preventing activity. 1. Synthesis, physico-chemical properties and quantum chemical indexes of some phenolic and aminophenolic antioxidants. *Rev. Roum. Chim.*, **36**, 1147–1160.
- Heritage, T.W., Ferguson, A.M., Turner, D.B. and Willett, P. (1998) EVA: a novel theoretical descriptor for QSAR studies, in *3D QSAR in Drug Design*, Vol. 2 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 381–398.
- Hermann, A. and Zinn, P. (1995) List operations on chemical graphs. 6. Comparative study of combinatorial topological indexes of the Hosoya type. *J. Chem. Inf. Comput. Sci.*, **35**, 551–560.
- Hermann, R.B. (1972) Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.*, **76**, 2754–2759.
- Hermann, R.B. (1997) Modeling hydrophobic solvation of nonspherical systems: comparison of use of molecular surface area with accessible surface area. *J. Comput. Chem.*, **18**, 115–125.
- Hermens, J.L.M. and Verhaar, H.J.M. (1995) QSARs in environmental toxicology and chemistry: recent developments. *ACS Symp. Ser.*, **606**, 130–140.
- Herndon, W.C. (1973a) Enumeration of resonance structures. *Tetrahedron*, **29**, 3–12.
- Herndon, W.C. (1973b) Resonance energies of aromatic hydrocarbons. A quantitative test of resonance theory. *J. Am. Chem. Soc.*, **95**, 2404–2406.
- Herndon, W.C. (1974a) Isospectral molecules. *Tetrahedron Lett.*, **8**, 671–674.
- Herndon, W.C. (1974b) Resonance theory and the enumeration of Kekulé structures. *J. Chem. Educ.*, **51**, 10–15.
- Herndon, W.C. (1974c) The characteristic polynomial does not uniquely determine molecular topology. *J. Chem. Doc.*, **14**, 150–151.
- Herndon, W.C. (1988) Graph codes and a definition of graph similarity. *Comp. Math. Applic.*, **15**, 303–309.
- Herndon, W.C. and Bertz, S.H. (1987) Linear notations and molecular graph similarity. *J. Comput. Chem.*, **8**, 367–374.
- Herndon, W.C. and Ellzey, M.L., Jr (1974) Resonance theory. V. Resonance energies of benzenoid and nonbenzenoid π systems. *J. Am. Chem. Soc.*, **96**, 6631–6642.
- Herndon, W.C. and Ellzey, M.L., Jr (1975) Isospectral graphs and molecules. *Tetrahedron*, **31**, 99–107.
- Herndon, W.C. and Szentpály, L.V. (1986) Theoretical model of activation of carcinogenic polycyclic benzenoid aromatic hydrocarbons. Possible new classes of carcinogenic aromatic hydrocarbons. *J. Mol. Struct. (Theochem)*, **148**, 141–152.
- Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E. and Schuffenhauer, A. (2004a) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.*, **44**, 1177–1185.
- Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E. and Schuffenhauer, A. (2004b) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.*, **2**, 3256–3266.
- Hess, B.A., Jr and Schaad, L.J. (1971a) Hückel molecular orbital π resonance energies. A new approach. *J. Am. Chem. Soc.*, **93**, 305–310.
- Hess, B.A., Jr and Schaad, L.J. (1971b) Hückel molecular orbital π resonance energies. The benzenoid hydrocarbons. *J. Am. Chem. Soc.*, **93**, 2413–2416.
- Hess, B.A., Jr and Schaad, L.J. (1973) Hückel molecular orbital π -resonance energies. Heterocycles containing divalent sulfur. *J. Am. Chem. Soc.*, **95**, 3907–3912.
- Hess, B.A., Jr, Schaad, L.J. and Holyoke, C.W., Jr (1972) On the aromaticity of heterocycles containing the amine nitrogen or the ether oxygen. *Tetrahedron*, **28**, 3657–3667.
- Hess, B.A., Jr, Schaad, L.J. and Holyoke, C.W., Jr (1975) The aromaticity of heterocycles containing the imine hydrogen. *Tetrahedron*, **31**, 295–298.
- Hetnarski, B. and O'Brien, R.D. (1973) Charge transfer in cholinesterase inhibition. Role of the conjugation between carbamyl and aryl groups of aromatic carbamates. *Biochemistry*, **12**, 3883–3887.

- Hetnarski, B. and O'Brien, R.D. (1975) The charge-transfer constant. A new substituent constant for structure-activity relationships. *J. Med. Chem.*, **18**, 29-33.
- Hewitt, M., Cronin, M.T.D., Madden, J.C., Rowe, P.H., Johnson, C., Obi, A. and Enoch, S.J. (2007) Consensus QSAR models: do the benefits outweigh the complexity? *J. Chem. Inf. Model.*, **47**, 1460-1468.
- Hicks, M.G. and Jochum, C. (1990) Substructure search systems for large chemical databases. *Anal. Chim. Acta*, **235**, 87-92.
- Higuchi, T. and Davis, S.S. (1970) Thermodynamic analysis of structure-activity relationships of drugs: prediction of optimal structure. *J. Pharm. Sci.*, **59**, 1376-1383.
- Hilal, S.H., Karichoff, S.W. and Carreira, L.A. (1995) A rigorous test for SPARC's chemical reactivity models: estimation of more than 4300 ionization pK_a s. *Quant. Struct.-Act. Relat.*, **14**, 348-355.
- Hildebrand, J.H. and Scott, R.L. (1950) *Solubility of Nonelectrolytes*, Reinhold Publishing Corporation, New York.
- Hill, T.L. (1948) Steric effects. I. van der Waals potential energy curves. *J. Chim. Phys.*, **16**, 399-404.
- Hine, J. (1962) *Physical Organic Chemistry*, McGraw-Hill, Inc., New York.
- Hine, J. and Mookerjee, P.K. (1975) The intrinsic hydrophilic character of organic compounds. Correlations in terms of structural contributions. *J. Org. Chem.*, **40**, 292-298.
- Hinze, J. and Jaffé, H.H. (1962) Electronegativity. I. Orbital electronegativity of neutral atoms. *J. Am. Chem. Soc.*, **84**, 540-546.
- Hinze, J. and Jaffé, H.H. (1963a) Electronegativity. III. Orbital electronegativities and electron affinities of transition metals. *Can. J. Chem.*, **41**, 1315-1328.
- Hinze, J. and Jaffé, H.H. (1963b) Electronegativity. IV. Orbital electronegativities of the neutral atoms of the periods three A and four A and of positive ions of periods one and two. *J. Phys. Chem.*, **67**, 1501-1505.
- Hinze, J. and Welz, U. (1996) Broad smiles, in *Software Development in Chemistry*, Vol. 10 (ed. J. Gasteiger), Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main, Germany, pp. 59-65.
- Hinze, J., Whitehead, M.A. and Jaffé, H.H. (1963) Electronegativity. II. Bond and orbital electronegativities. *J. Am. Chem. Soc.*, **85**, 148-154.
- Hiob, R. and Karelson, M. (2000) Quantitative relationship between rate constants of the gas-phase homolysis of C-X bonds and molecular descriptors. *J. Chem. Inf. Comput. Sci.*, **40**, 1062-1071.
- Hiob, R. and Karelson, M. (2002) Quantitative relationship between rate constants of the gas-phase homolysis of N-N, O-O and N-O bonds and molecular descriptors. *Internet Electron. J. Mol. Des.*, **1**, 193-202.
- Hirashima, A., Kuwano, E. and Eto, M. (2003) Comparative receptor surface analysis of octopaminergic antagonists for the locust neuronal octopamine receptor. *Comp. Biol. Chem.*, **27**, 531-540.
- Hirono, S., Nakagome, I., Hirano, H., Yoshii, F. and Moriguchi, I. (1994) Noncongeneric structure pharmacokinetic property correlation studies using fuzzy adaptive least squares volume of distribution. *Biol. Pharm. Bull.*, **17**, 686-690.
- Hirono, S., Qian, L. and Moriguchi, I. (1991) High correlation between hydrophobic free energy and molecular surface area characterized by electrostatic potential. *Chem. Pharm. Bull.*, **39**, 3106-3109.
- Hirons, L., Holliday, J.D., Jelfs, S.P., Willett, P. and Gedeck, P. (2005) Use of the R-group descriptor for alignment-free QSAR. *QSAR Comb. Sci.*, **24**, 611-619.
- Hirst, J.D. (1996) Nonlinear quantitative structure-activity relationship for the inhibition of dihydrofolate reductase by pyrimidines. *J. Med. Chem.*, **39**, 3526-3532.
- Hocart, S.J., Reddy, V., Murphy, W.A. and Coy, D.H. (1995) Three-dimensional quantitative structure-activity relationships of somatostatin analogs. 1. Comparative molecular field analysis of growth hormone release inhibiting potencies. *J. Med. Chem.*, **38**, 1974-1989.
- Hocking, R.R. (1976) The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1-49.
- Hodes, L. (1976) Selection of descriptors according to discrimination and redundancy. Application to chemical structure searching. *J. Chem. Inf. Comput. Sci.*, **16**, 88-93.
- Hodes, L. (1981a) Computer-aided selection of compounds for antitumor screening: validation of a statistical-heuristic method. *J. Chem. Inf. Comput. Sci.*, **21**, 128-132.
- Hodes, L. (1981b) Selection of molecular fragment features for structure-activity studies in antitumor screening. *J. Chem. Inf. Comput. Sci.*, **21**, 132-136.
- Hodes, L., Hazard, G.F., Geran, R.I. and Richman, S. (1977) A statistical-heuristic method for automated selection of drugs for screening. *J. Med. Chem.*, **20**, 469-475.

- Hodgkin, E.E. and Richards, W.G. (1987) Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **14**, 105–110.
- Hodj Mohammadi, M.R., Ebrahimi, P. and Pourmorad, F. (2004) Quantitative structure–retention relationships (QSRR) of some CNS agents studied on DB-5 and DB-17 phases in gas chromatography. *QSAR Comb. Sci.*, **23**, 295–302.
- Hoefnagel, A.J., Oosterbeek, W. and Wepster, B.M. (1984) Substituent effects. 10. Critique of the “improved evaluation of field and resonance effects” proposed by Swain *et al.* *J. Org. Chem.*, **49**, 1993–1997.
- Hoerl, A.E. and Kennard, R.W. (1970) Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55–67.
- Hoffman, B., Cho, S.J., Zheng, W., Wyrick, S.D., Nichols, D.E., Mailman, R.B. and Tropsha, A. (1999) Quantitative structure–activity relationship modeling of dopamine D-1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.*, **42**, 3217–3226.
- Holder, A.J., Ye, L., Eick, J.D. and Chappelow, C.C. (2006a) A quantum-mechanical QSAR model to predict the refractive index of polymer matrices. *QSAR Comb. Sci.*, **25**, 905–911.
- Holder, A.J., Ye, L., Eick, J.D. and Chappelow, C.C. (2006b) An application of QM-QSAR to predict and rationalize the refractive index of a wide variety of simple organic/organosilicon molecules. *QSAR Comb. Sci.*, **25**, 342–349.
- Holder, A.J., Yourtee, D.M., White, D.A., Glaros, A.G. and Smith, R. (2003) Chain melting temperature estimation for phosphatidylcholines by quantum mechanically derived quantitative structure–property relationships. *J. Comput. Aid. Mol. Des.*, **17**, 223–230.
- Holik, M. and Halamek, J. (2002) Transformation of a Free–Wilson matrix into Fourier coefficients. *Quant. Struct. -Act. Relat.*, **20**, 422–428.
- Holland, J. (1975) *Adaptation in Artificial and Natural Systems*, University of Michigan Press, Ann Arbor, MI.
- Hollas, B. (2002) Correlation properties of the autocorrelation descriptor for molecules. *MATCH Commun. Math. Comput. Chem.*, **45**, 27–33.
- Hollas, B. (2003) Correlations in distance-based descriptors. *MATCH Commun. Math. Comput. Chem.*, **47**, 79–86.
- Hollas, B. (2005a) Asymptotically independent topological indices on random trees. *J. Math. Chem.*, **38**, 379–387.
- Hollas, B. (2005b) On the variance of topological indices that depend on the degree of a vertex. *MATCH Commun. Math. Comput. Chem.*, **54**, 341–350.
- Hollas, B. (2005c) The covariance of topological indices that depend on the degree of a vertex. *MATCH Commun. Math. Comput. Chem.*, **54**, 177–187.
- Hollas, B. (2006) An analysis of the redundancy of graph invariants used in chemoinformatics. *Disc. Appl. Math.*, **154**, 2484–2498.
- Hollas, B., Gutman, I. and Trinajstić, N. (2005) On reducing correlations between topological indices. *Croat. Chem. Acta*, **78**, 489–492.
- Holliday, J.D., Jelfs, S.P. and Willett, P. (2003) Calculation of intersubstituent similarity using R-group descriptors. *J. Chem. Inf. Comput. Sci.*, **43**, 406–411.
- Holliday, J.D., Ranade, S.S. and Willett, P. (1995) A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct. -Act. Relat.*, **14**, 501–506.
- Holliday, J.D., Salim, N., Whittle, M. and Willett, P. (2003) Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **43**, 819–828.
- Holmes, E., Nicholls, A.W., Lindon, J.C., Connor, S.C., Connelly, J.C., Haselden, J.N., Damment, S.J. P., Spraul, M., Neidig, P. and Nicholson, J.K. (2000) Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem. Res. Toxicol.*, **13**, 471–478.
- Holmes, E., Nicholson, J.K. and Tranter, G. (2001) Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. *Chem. Res. Toxicol.*, **14**, 182–191.
- Höltje, H.-D. (1975) Theoretische untersuchungen zu struktur–wirkungsbeziehungen bei monoachinoxidase-Hemmern der cyclopropylamin-reihe. *Arch. Pharm. (Weinheim Ger.)*, **308**, 438–444.
- Höltje, H.-D. (1976) Theoretische untersuchungen zu struktur–wirkungsbeziehungen von antihypertensiv wirkenden benzothiadiazin-1, 1-dioxiden. *Arch. Pharm. (Weinheim Ger.)*, **309**, 480–485.
- Höltje, H.-D. (1982) Theoretische untersuchungen zu struktur–wirkungsbeziehungen von ringsubstituierten verapamil-derivaten. *Arch. Pharm. (Weinheim Ger.)*, **315**, 317–323.
- Höltje, H.-D., Anzali, S., Dall, N. and Höltje, M. (1993) Binding site models, in *3D QSAR in Drug Design: Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, Germany, pp. 320–354.

- Höltje, H.-D., Baranowski, P., Spengler, J.P. and Schunack, W. (1985) Ein bindungsstellenmodell für H₂-antagonisten vom 4-pyrimidinon-typ. *Arch. Pharm. (Weinheim Ger.)*, **318**, 542–548.
- Höltje, H.-D. and Kier, L.B. (1974) A theoretical approach to structure–activity relationships of chloramphenicol and congeners. *J. Med. Chem.*, **17**, 814–819.
- Höltje, H.-D. and Tintelnot, M. (1984) Theoretical investigations on interactions between pharmacological molecules and receptor models. V. Construction of a model for the ribosomal binding site of chloramphenicol. *Quant. Struct. -Act. Relat.*, **3**, 6–9.
- Höltje, H.-D. and Vogelgesang, L. (1979) Theoretische untersuchung zur hemmung der noradrenalin-rückresorption durch phenylethylaminanaloge verbindungen. *Arch. Pharm. (Weinheim Ger.)*, **312**, 578–586.
- Holtz, H.D. and Stock, L.M. (1964) Dissociation constants for 4-substituted bicyclo[2,2,2]octane-1-carboxylic acids. Empirical and theoretical analysis. *J. Am. Chem. Soc.*, **86**, 5188–5194.
- Holzgrabe, U. and Hopfinger, A.J. (1996) Conformational analysis, molecular shape comparison, and pharmacophore identification of different allosteric modulators of muscarinic receptors. *J. Chem. Inf. Comput. Sci.*, **36**, 1018–1024.
- Hong, H., Wang, L.-S. and Han, S. (1996) Prediction adsorption coefficients (K_{oc}) for aromatic compounds by HPLC retention factors (K'). *Chemosphere*, **32**, 343–351.
- Hong, X. and Hopfinger, A.J. (2003) 3D-pharmacophores of flavonoid binding at the benzodiazepine GABAA receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.*, **43**, 324–336.
- Honório, K.M. and da Silva, A.B.F. (2002) A theoretical study on the influence of the frontier orbitals HOMO and LUMO and the size of C4 and C2 substituents in the psychoactivity of cannabinoid compounds. *J. Mol. Struct. (Theochem)*, **578**, 111–117.
- Hoover, K.R., Acree, W.E., Jr and Abraham, M.H. (2005) Chemical toxicity correlations for several fish species based on the Abraham solvation parameter model. *Chem. Res. Toxicol.*, **18**, 1497–1505.
- Hopfinger, A.J. (1980) A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.*, **102**, 7196–7206.
- Hopfinger, A.J. (1981) Inhibition of dihydrofolate reductase: structure–activity correlations of 2, 4-diamino-5-benzylpyrimidines based upon molecular shape analysis. *J. Med. Chem.*, **24**, 818–822.
- Hopfinger, A.J. (1983) Theory and application of molecular potential energy fields in molecular shape analysis: a quantitative structure–activity relationship study of 2,4-diamino-5-benzylpyrimidines as dihydrofolate reductase inhibitors. *J. Med. Chem.*, **26**, 990–996.
- Hopfinger, A.J. (1984) A QSAR study of the Ames mutagenicity of 1-(X-phenyl)-3,3-dialkyltriazines using molecular potential energy fields and molecular shape analysis. *Quant. Struct. -Act. Relat.*, **3**, 1–5.
- Hopfinger, A.J. and Battershell, R.D. (1976) Application of SCAP to drug design. 1. Prediction of octanol–water partition coefficients using solvent-dependent conformational analyses. *J. Med. Chem.*, **19**, 569–573.
- Hopfinger, A.J. and Burke, B.J. (1990) Molecular shape analysis: a formalism to quantitatively establish spatial molecular similarity, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 173–209.
- Hopfinger, A.J., Compadre, R.L.L., Koehler, M.G., Emery, S. and Seydel, J.K. (1987) An extended QSAR analysis of some 4-aminodiphenylsulfone antibacterial agents using molecular modeling and LFE-relationships. *Quant. Struct. -Act. Relat.*, **6**, 111–117.
- Hopfinger, A.J. and Patel, H.C. (1996) Application of genetic algorithms to the general QSAR problem and to guiding molecular diversity experiments, in *Genetic Algorithms in Molecular Modeling: Principles of QSAR and Drug Design*, Vol. 1 (ed. J. Devillers), Academic Press, London, UK.
- Hopfinger, A.J. and Potenzzone, R., Jr (1982) Ames test and antitumor activity of 1-(x-phenyl)-3,3-dialkyltriazines, quantitative structure-analysis studies based upon molecular shape analysis. *Mol. Pharm.*, **21**, 187–195.
- Hopfinger, A.J., Wang, S., Tokarski, J.S., Jin, B., Albuquerque, M., Madhav, P.J., and Duraiswami, C. (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.*, **119**, 10509–10524.
- Hopkinson, A.C. (1969) Unimolecular and bimolecular mechanisms in the acid-catalyzed hydrolysis of methyl esters of aliphatic monocarboxylic acids in aqueous sulphuric acid. *J. Chem. Soc., B*, 861–863.
- Hormann, R.E., Dinan, L. and Whiting, P. (2003) Superimposition evaluation of ecdysteroid agonist

- chemotypes through multidimensional QSAR. *J. Comput. Aid. Mol. Des.*, **17**, 135–153.
- Horvat, D., Graovac, A., Plavšić, D., Trinajstić, N. and Strunje, M. (1992) On the intercorrelation of topological indices in benzenoid hydrocarbons. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **26**, 401–408.
- Horvath, A.L. (1988) Estimate properties of organic compounds: simple polynomial equations relate the properties of organic compounds to their chemical structure. *Chem. Eng.*, **95**, 155–158.
- Horvath, A.L. (1992) *Molecular Design*, Elsevier, Amsterdam, The Netherlands, p. 1490.
- Horvath, D., Bonachera, F., Solov'ev, V., Gaudin, C. and Varnek, A. (2007) Stochastic versus stepwise strategies for quantitative structure–activity relationship generations: how much effort may the mining for successful QSAR models take? *J. Chem. Inf. Model.*, **47**, 927–939.
- Horvath, D. and Mao, B. (2003) Neighborhood behavior. Fuzzy molecular descriptors and their influence on the relationship between structural similarity and property similarity. *QSAR Comb. Sci.*, **22**, 498–509.
- Howell, D.C., Howson, W., Higginbottom, M., Naylor, D., Ratcliffe, G.S. and Williams, S. (1995) Quantitative structure–activity relationships (QSARs) of N-terminus fragments of NK1 tachykinin antagonists: a comparison of classical QSARs and three-dimensional QSARs from similarity matrices. *J. Med. Chem.*, **38**, 4454–4462.
- Horwitz, J.P., Massova, I., Wiese, T.E., Besler, B.H. and Corbett, T.H. (1994) Comparative molecular field analysis of the antitumor activity of 9H-thioxanthen-9-one derivatives against pancreatic ductal carcinoma 03. *J. Med. Chem.*, **37**, 781–786.
- Horwitz, J.P., Massova, I., Wiese, T.E., Wozniak, A.J., Corbett, T.H., Seboltleopold, J.S., Capps, D.B. and Leopold, W.R. (1993) Comparative molecular field analysis of *in vitro* growth inhibition of L1210 and HCT-8 cells by some pyrazoloacridines. *J. Med. Chem.*, **36**, 3511–3516.
- Höskuldsson, A. (1988) PLS regression methods. *J. Chemom.*, **2**, 211–228.
- Höskuldsson, A. (2001) Variable and subset selection in PLS regression. *Chemom. Intell. Lab. Syst.*, **55**, 23–28.
- Hosoya, H. (1971) Topological index. a newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jap.*, **44**, 2332–2339.
- Hosoya, H. (1972a) Graphical enumeration of the coefficients of the secular polynomials of the Hückel molecular orbitals. *Theor. Chim. Acta*, **25**, 215–222.
- Hosoya, H. (1972b) Topological index and thermodynamics properties. I. Empirical rules on the boiling point of saturated hydrocarbons. *Bull. Chem. Soc. Jap.*, **45**, 3415–3421.
- Hosoya, H. (1972c) Topological index as a sorting device for coding chemical structures. *J. Chem. Doc.*, **12**, 181–183.
- Hosoya, H. (1973) Topological index and Fibonacci numbers with relation to chemistry. *Fibonacci Quarterly*, **11**, 255–266.
- Hosoya, H. (1986) Topological index as a common tool for quantum chemistry, statistical mechanics, and graph theory, in *Mathematics and Computational Concepts in Chemistry* (ed. N. Trinajstić), Ellis Horwood, Chichester, UK, pp. 110–123.
- Hosoya, H. (1988) On some counting polynomials in chemistry. *Disc. Appl. Math.*, **19**, 239–257.
- Hosoya, H. (1990) Some recent advances in counting polynomials in chemical graph theory, in *Computational Chemical Graph Theory* (ed. D.H. Rouvray), Nova Science Publishers, New York, pp. 105–126.
- Hosoya, H. (1991) Factorization and recursion of the matching and characteristic polynomials of periodic polymer networks. *J. Math. Chem.*, **7**, 289–305.
- Hosoya, H. (1994) Topological twin graphs. Smallest pair of isospectral polyhedral graphs with eight vertices. *J. Chem. Inf. Comput. Sci.*, **34**, 428–431.
- Hosoya, H. (1999) Mathematical foundation of the organic electron theory: how do π -electron flow in conjugated molecules. *J. Mol. Struct. (Theochem)*, **461–462**, 473–482.
- Hosoya, H. (2002) Chemical meaning of octane number analyzed by topological indices. *Croat. Chem. Acta*, **75**, 433–445.
- Hosoya, H. (2003) From how to why. Graph-theoretical verification of quantum mechanical aspects of π -electron behaviors in conjugated systems. *Bull. Chem. Soc. Jap.*, **76**, 2233–2252.
- Hosoya, H. (2007) Important mathematical structures of the topological index Z for tree graphs. *J. Chem. Inf. Model.*, **47**, 744–750.
- Hosoya, H., Gotoh, M., Murakami, M. and Ikeda, S. (1999) Topological index and thermodynamic properties. 5. How can we explain the topological dependency of thermodynamic properties of alkanes with the topology of graphs? *J. Chem. Inf. Comput. Sci.*, **39**, 192–196.
- Hosoya, H., Gutman, I. and Nikolić, J. (1992) Topological indices of unbranched catacondensed

- benzenoid hydrocarbons. *Bull. Chem. Soc. Jap.*, **65**, 2011–2015.
- Hosoya, H., Hosoi, K. and Gutman, I. (1975) A topological index for the total π -electron energy. Proof of a generalized Hückel rule for an arbitrary network. *Theor. Chim. Acta*, **38**, 37–47.
- Hosoya, H. and Murakami, M. (1975) Topological index as applied to π -electronic systems. II. Topological bond order. *Bull. Chem. Soc. Jap.*, **48**, 3512–3517.
- Hosoya, H., Murakami, M. and Gotoh, M. (1973) Distance polynomial and characterization of a graph. *Natl. Sci. Rept. Ochanomizu Univ.*, **24**, 27–34.
- Hosoya, H. and Ohkami, N. (1983) Operator technique for obtaining the recursion formulas of characteristic and matching polynomials as applied to polyhex graphs. *J. Comput. Chem.*, **4**, 585–593.
- Hou, T.-J., Li, Z.-M., Li, Z., Liu, J. and Xu, X. (2000) Three-dimensional quantitative structure–activity relationship analysis of the new potent sulfonylureas using comparative molecular similarity indices analysis. *J. Chem. Inf. Comput. Sci.*, **40**, 1002–1009.
- Hou, T.-J., Wang, J. and Xu, X. (1999) Applications of genetic algorithms on the structure–activity correlation study of a group of nonnucleoside HIV-1 inhibitors. *Chemom. Intell. Lab. Syst.*, **45**, 303–310.
- Hou, T.-J., Xia, K., Zhang, W. and Xu, X. (2004) ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.*, **44**, 266–275.
- Hou, T.-J. and Xu, X. (2002) ADME evaluation in drug discovery. 1. Applications of genetic algorithms on the prediction of blood–brain partitioning of a large set drugs. *J. Mol. Model.*, **8**, 337–349.
- Hou, T.-J. and Xu, X. (2003) ADME evaluation in drug discovery 2. Prediction of partition coefficient by atom-additive approach based on atom-weighted solvent accessible surface areas. *J. Chem. Inf. Comput. Sci.*, **43**, 1058–1067.
- Hou, T.-J. and Xu, X. (2003) ADME evaluation in drug discovery 3. Modeling blood–brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comput. Sci.*, **43**, 2137–2152.
- Hou, T.-J., Zhang, W., Xia, K., Qiao, X.-B., and Xu, X. (2004) ADME evaluation in drug discovery 5. Correlation of Caco-2 permeation with simple molecular properties. *J. Chem. Inf. Comput. Sci.*, **44**, 1585–1600.
- Hou, T.-J., Zhu, L., Chen, L., and Xu, X. (2003) Mapping the binding site of a large set of quinazoline type EGF-R inhibitors using molecular field analyses and molecular docking studies. *J. Chem. Inf. Comput. Sci.*, **43**, 273–287.
- Howard, S.T. and Krygowski, T.M. (1997) Benzenoid hydrocarbon aromaticity in terms of charge density descriptors. *Can. J. Chem.*, **75**, 1174–1181.
- Howard, S.T., Krygowski, T.M., Ciesielski, A. and Wisiorowski, M. (1998) Angular group-induced alternation. II. The magnitude and the nature of the effect and its application to polynuclear benzenoid systems. *Tetrahedron*, **54**, 3533–3548.
- HQSAR: A New, Highly Predictive QSAR Technique, Ver. 1.0, Tripos Technical Notes, Tripos Associates, Inc., 1699 S Hanley Road, Suite 303, St. Louis, MO, <http://www.tripos.com/products/hqsar.html>.
- Hristozov, D., Da Costa, F.B. and Gasteiger, J. (2007) Sesquiterpene lactones-based classification of the family Asteraceae using neural networks and k -nearest neighbors. *J. Chem. Inf. Model.*, **47**, 9–19.
- Hu, C.-Y. and Xu, L. (1994) On Hall and Kier's topological state and total topological index. *J. Chem. Inf. Comput. Sci.*, **34**, 1251–1258.
- Hu, C.-Y. and Xu, L. (1996) On highly discriminating molecular topological index. *J. Chem. Inf. Comput. Sci.*, **36**, 82–90.
- Hu, C.-Y. and Xu, L. (1997) Developing molecular identification numbers by an all-paths method. *J. Chem. Inf. Comput. Sci.*, **37**, 311–315.
- Hu, M.K. (1962) Visual pattern recognition by moment invariants. *IRE Trans. Info. Theory*, **8**, 179–187.
- Hu, Q.-H., Wang, X.-J. and Brusseau, M.L. (1995) Quantitative structure–activity relationships for evaluating the influence of sorbate structure on sorption of organic compounds by soil. *Environ. Toxicol. Chem.*, **14**, 1133–1140.
- Hu, Q.-N., Liang, Y.-Z. and Fang, K.-T. (2003) The matrix expression, topological index and atomic attribute of molecular topological structure. *Journal of Data Science*, **1**, 361–389.
- Hu, Q.-N., Liang, Y.-Z., Peng, X.-L., Yin, H. and Fang, K.-T. (2004) Structural interpretation of a topological index. 1. External factor variable connectivity index (EFVCI). *J. Chem. Inf. Comput. Sci.*, **44**, 437–446.
- Hu, Q.-N., Liang, Y.-Z., Wang, Y.-L., Xu, C.-J., Zeng, Z.-D., Fang, K.-T., Peng, X.-L. and Hong, Y. (2003) External factor variable connectivity index. *J. Chem. Inf. Comput. Sci.*, **43**, 773–778.
- Hu, Y., Li, X., Shi, Y., Xu, T. and Gutman, I. (2005) On molecular graphs with smallest and greatest zeroth-order general Randić index. *MATCH Commun. Math. Comput. Chem.*, **54**, 425–434.

- Huang, H., Wang, X., Ou, W., Zhao, J., Shao, Y and Wang, L.-S. (2003) Acute toxicity of benzene derivatives to the tadpoles (*Rana japonica*) and QSAR analyses. *Chemosphere*, **53**, 963–970.
- Huang, M.-J. and Bodor, N. (1994) Quantitative structure–inhibitory activity relationships of substituted phenols on *Bacillus subtilis* spore germination. *Int. J. Quant. Chem.*, **52**, 181–185.
- Huang, Q.-G., Kong, L. and Wang, L.-S. (1996) Applications of frontier molecular orbital energies in QSAR studies. *Bull. Environ. Contam. Toxicol.*, **56**, 758–765.
- Huang, Q.-G., Song, W.-L. and Wang, L.-S. (1997) Quantitative relationship between the physiochemical characteristics as well as genotoxicity of organic pollutants and molecular autocorrelation topological descriptors. *Chemosphere*, **35**, 2849–2855.
- Huang, Q.-G., Wang, L.-S. and Han, S. (1995) The genotoxicity of substituted nitrobenzenes and the quantitative structure–activity relationship studies. *Chemosphere*, **30**, 915–923.
- Hübel, S., Rösner, T. and Franke, R. (1980) The evaluation of topological pharmacophores by heuristic approaches. *Pharmazie*, **35**, 424–433.
- Hückel, E. (1930) Zur quantentheorie der doppelbindung. *Z. Phys. (German)*, **60**, 423–456.
- Hückel, E. (1931a) Quantentheoretische beiträge zum benzolproblem. I. Die elektronenkonfiguration des benzols und verwandter verbindungen. *Z. Phys. (German)*, **70**, 204–286.
- Hückel, E. (1931b) Quantentheoretische beiträge zum benzolproblem. II. Quantentheorie der induzierten Polaritäten. *Z. Phys. (German)*, **72**, 310–337.
- Hückel, E. (1932) Quantentheoretische beiträge zum benzolproblem. III. Quantentheoretische beiträge zum problem der aromatischen und ungesättigten verbindungen. *Z. Phys. (German)*, **76**, 628–648.
- Hückel, E. (1933) Quantentheoretische beiträge zum benzolproblem. IV. Die freien radikale der organischen chemie. *Z. Phys. (German)*, **83**, 632.
- Huggins, M. (1956) Densities and optical properties of organic compounds in the liquid state. VI. The refractive indices of paraffin hydrocarbons and some of their derivatives. *Bull. Chem. Soc. Jap.*, **29**, 336–339.
- Huheey, J.E. (1965) The electronegativity of groups. *J. Phys. Chem.*, **69**, 3284–3291.
- Huheey, J.E. (1966) The electronegativity of multiply-bonded groups. *J. Phys. Chem.*, **70**, 2086–2092.
- Huibers, P.D.T. and Katritzky, A.R. (1998) Correlation of the aqueous solubility of hydrocarbons and halogenated hydrocarbons with molecular structure. *J. Chem. Inf. Comput. Sci.*, **38**, 283–292.
- Huibers, P.D.T., Lobanov, V.S., Katritzky, A.R., Shah, D.O. and Karelson, M. (1996) Prediction of critical micelle concentration using a quantitative structure–property relationship approach. 1. Nonionic surfactants. *Langmuir*, **12**, 1462–1470.
- Huibers, P.D.T., Lobanov, V.S., Katritzky, A.R., Shah, D.O. and Karelson, M. (1997) Prediction of critical micelle concentration using a quantitative structure–property relationship approach. *J. Colloid Interf. Sci.*, **187**, 113–120.
- Hull, R.D., Singh, S.B., Nachbar, R.B., Sheridan, R.P., Kearsley, S.K. and Fluder, E.M. (2001) Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.*, **44**, 1177–1184.
- Hunt, P.A. (1999) QSAR using 2D descriptors and TRIPOS' SIMCA. *J. Comput. Aid. Mol. Des.*, **13**, 453–467.
- Hunter LaFemina, D. and Jurs, P.C. (1985) A numerical index for characterizing data set separation. *J. Chem. Inf. Comput. Sci.*, **25**, 386–388.
- Hutter, M.C. (2003) Prediction of blood–brain barrier permeation using quantum chemically derived information. *J. Comput. Aid. Mol. Des.*, **17**, 415–433.
- Hutter, M.C. (2007) Separating drugs from nondrugs: a statistical approach using atom pair distributions. *J. Chem. Inf. Model.*, **47**, 186–194.
- Huuskonen, J.J. (2000) Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.*, **40**, 773–777.
- Huuskonen, J.J. (2001) QSAR modeling with the electrotopological state: TIBO derivatives. *J. Chem. Inf. Comput. Sci.*, **41**, 425–429.
- Huuskonen, J.J. (2003) QSAR modeling with the electrotopological state indices: predicting the toxicity of organic chemicals. *Chemosphere*, **50**, 949–953.
- Huuskonen, J.J., Livingstone, D.J. and Tetko, I.V. (2000) Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.*, **40**, 947–955.
- Huuskonen, J.J., Rantanen, J. and Livingstone, D.J. (2000) Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.*, **35**, 1081–1088.
- Huuskonen, J.J., Salo, M. and Taskinen, J. (1998) Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.*, **38**, 450–456.

- Hyde, R.M. and Livingstone, D.J. (1988) Perspectives in QSAR: computer chemistry and pattern recognition. *J. Comput. Aid. Mol. Des.*, **2**, 145–155.
- Iczkowski, R.P. and Margrave, J.L. (1961) Electronegativity. *J. Am. Chem. Soc.*, **83**, 3547–3551.
- Idoux, J.P., Hwang, P.T.R. and Hancock, C.K. (1973) Study of the alkaline hydrolysis and nuclear magnetic resonance spectra of some thiol esters. *J. Org. Chem.*, **38**, 4239–4243.
- Idoux, J.P., Scandrett, J.M. and Sikorski, J.A. (1977) Conformational influence of nonacyl groups on acyl group properties in *N*-monosubstituted amides and in other carboxylic acid derivatives: a 7-position proximity effect. *J. Am. Chem. Soc.*, **99**, 4577–4583.
- Ignatz-Hoover, F., Petrukhin, R., Karelson, M. and Katritzky, A.R. (2001) QSRR correlation of free-radical polymerization chain-transfer constants for styrene. *J. Chem. Inf. Comput. Sci.*, **41**, 295–299.
- Ihlenfeldt, W.D. and Gasteiger, J. (1994) Hash codes for the identification and classification of molecular structure elements. *J. Comput. Chem.*, **15**, 793–813.
- Ihlenfeldt, W.D., Takahashi, Y., Abe, H. and Sasaki, S. (1994) Computation and management of chemical properties in CACTVS: an extensible network approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.*, **34**, 109–116.
- Ijjaali, I., Petitet, F., Dubus, E., Barberan, O. and Michel, A. (2007) Assessing potency of c-Jun N-terminal kinase 3 (JNK3) inhibitors using 2D molecular descriptors and binary QSAR methodology. *Bioorg. Med. Chem.*, **15**, 4256–4264.
- Ikemoto, Y., Motoba, K., Suzuki, T. and Uchida, M. (1992) Quantitative structure–activity relationships of nonspecific and specific toxicants in several organism species. *Environ. Toxicol. Chem.*, **11**, 931–939.
- Immirzi, A. and Perini, B. (1977) Prediction of density in organic crystals. *Acta Cryst.*, **33**, 216–218.
- Imre, G., Veress, G., Volford, A. and Farkas, Ö. (2003) Molecules from the Minkowski space: an approach to building 3D molecular structures. *J. Mol. Struct. (Theochem)*, **666–667**, 51–59.
- Inamoto, N. and Masuda, S. (1977) Substituent effects on C-13 chemical shifts in aliphatic and aromatic series. Proposal of new inductive substituent parameter (ι ; iota) and the application. *Tetrahedron Lett.*, **18**, 3287–3290.
- Inamoto, N. and Masuda, S. (1982) Revised method for calculation of group electronegativities. *Chem. Lett.*, 1003–1007.
- Inamoto, N., Masuda, S., Tori, K. and Yoshimura, Y. (1978) Effects of fixed substituents upon substituent chemical shifts of the C-1 atom in *m*- and *p*-disubstituted benzenes. Correlation with inductive substituent parameter (ι). *Tetrahedron Lett.*, **19**, 4547–4550.
- Isaeva, G.A., Dmitriev, A.V. and Isaev, P.P. (2001) QSAR relationships for the anesthetic activity of acetanilides analyzed by regression and quantum–chemical methods. *Pharm. Chem. J.*, **35**, 348–350.
- Ishihama, Y. and Asakawa, N. (1999) Characterization of lipophilicity scales using vectors from solvation energy descriptors. *J. Pharm. Sci.*, **88**, 1305–1312.
- Ishihama, Y., Oda, N. and Asakawa, N. (1996) Hydrophobicity of cationic solutes measured by electrokinetic chromatography with cationic microemulsions. *Anal. Chem.*, **68**, 4281–4284.
- ISIS/Draw 2.1, MDL Information Systems, 14600 Catalina Street, San Leandro, CA.
- Isogai, Y. and Itoh, T. (1984) Fractal analysis of tertiary structure. *J. Phys. Soc. Japan*, **53**, 2162.
- Itskowitz, P. and Berkowitz, M.L. (1997) Chemical potential equalization principle: direct approach from density functional theory. *J. Phys. Chem. A*, **101**, 5687–5691.
- IUPAC Recommendations (1997) Glossary of terms in computational drug design. *Prot. Struct. Funct. Gen.*, **69**, 1137–1152.
- IUPAC Recommendations (1998) Glossary of terms used in medicinal chemistry. *Prot. Struct. Funct. Gen.*, **70**, 1129–1143.
- Ivanciuc, O. (1988a) Chemical graph polynomials. Part 1. The polynomial description of generalized chemical graphs. *Rev. Roum. Chim.*, **33**, 709–717.
- Ivanciuc, O. (1988b) Topological and empirical models. 1. The prediction of the Gibbs energies of formation for alkanes. *Rev. Roum. Chim.*, **33**, 839–845.
- Ivanciuc, O. (1989) Design on topological indices. 1. Definition of a vertex topological index in the case of 4-trees. *Rev. Roum. Chim.*, **34**, 1361–1368.
- Ivanciuc, O. (1992) Chemical graph polynomials. Part 2. The propagation diagram algorithm for the computation of the characteristic polynomial of molecular graphs. *Rev. Roum. Chim.*, **37**, 1341–1345.
- Ivanciuc, O. (1993) Chemical graph polynomials. Part 3. The Laplacian polynomial of molecular graphs. *Rev. Roum. Chim.*, **38**, 1499–1508.
- Ivanciuc, O. (1995) Artificial neural networks applications. Part 1. Estimation of the total π -electron energy of benzenoid hydrocarbons. *Rev. Roum. Chim.*, **40**, 1093–1101.
- Ivanciuc, O. (1996) Artificial neural networks applications. 2. Using theoretical descriptors of molecular structure in quantitative structure–

- activity relationships analysis of the inhibition of dihydrofolate reductase. *Rev. Roum. Chim.*, **41**, 645–652.
- Ivanciuc, O. (1997) Artificial neural networks applications. Part 3. A quantitative structure–activity relationship for the actinidin hydrolysis of substituted-phenyl hippurates. *Rev. Roum. Chim.*, **42**, 325–332.
- Ivanciuc, O. (1998a) Artificial neural networks applications. Part 4. Quantitative structure–activity relationships for the estimation of the relative toxicity of phenols for *Tetrahymena*. *Rev. Roum. Chim.*, **43**, 255–260.
- Ivanciuc, O. (1998b) Artificial neural networks applications. Part 7. Estimation of bioconcentration factors in fish using solvatochromic parameters. *Rev. Roum. Chim.*, **43**, 347–354.
- Ivanciuc, O. (1998c) Artificial neural networks applications. Part 9. MolNet prediction of alkane boiling points. *Rev. Roum. Chim.*, **43**, 885–894.
- Ivanciuc, O. (1998d) Chemical graph polynomials. Part 4. Non-isomorphic graphs with identical acyclic polynomials. *Rev. Roum. Chim.*, **43**, 1173–1179.
- Ivanciuc, O. (1998e) Design of topological indices. Part 9. A new recurrence relationship for the Hosoya $Z(\text{Ch})$ index of a molecular graph. *Rev. Roum. Chim.*, **43**, 481–484.
- Ivanciuc, O. (1999a) A new recurrence relationship for the computation of the number of Kekulé structures of benzenoid hydrocarbons. *Rev. Roum. Chim.*, **44**, 91–93.
- Ivanciuc, O. (1999b) Artificial neural networks. Part 11. MolNet prediction of alkane densities. *Rev. Roum. Chim.*, **44**, 619–631.
- Ivanciuc, O. (1999c) Design of topological indices. Part 11. Distance-valency matrices and derived molecular graph descriptors. *Rev. Roum. Chim.*, **44**, 519–528.
- Ivanciuc, O. (1999d) Molecular graph descriptors used in neural network models, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 697–777.
- Ivanciuc, O. (2000a) Design of topological indices. Part 12. Parameters for vertex- and edge-weighted molecular graphs. *Rev. Roum. Chim.*, **45**, 289–301.
- Ivanciuc, O. (2000b) Design of topological indices. Part 13. Structural descriptors computed from the Szeged molecular matrices. *Rev. Roum. Chim.*, **45**, 475–493.
- Ivanciuc, O. (2000c) Design of topological indices. Part 14. Distance-valency matrices and structural descriptors for vertex- and edge-weighted molecular graphs. *Rev. Roum. Chim.*, **45**, 587–596.
- Ivanciuc, O. (2000d) Design of topological indices. Part 15. The Szeged index of vertex- and edge-weighted molecular graphs. *Rev. Roum. Chim.*, **45**, 895–903.
- Ivanciuc, O. (2000e) Design of topological indices. Part 16. Matrix power operators for molecular graphs. *Rev. Roum. Chim.*, **45**, 1027–1044.
- Ivanciuc, O. (2000f) Design of topological indices. Part 17. The Szeged operator as a source of new structural descriptors. *Rev. Roum. Chim.*, **45**, 1105–1114.
- Ivanciuc, O. (2000g) Molecular structure encoding into artificial neural networks topology. *Roum. Chem. Quart. Rev.*, **8**, 197–220.
- Ivanciuc, O. (2000h) QSAR and QSPR molecular descriptors computed from the resistance distance and electrical conductance matrices. *ACH - Models Chem.*, **137**, 607–631.
- Ivanciuc, O. (2000i) QSAR comparative study of Wiener descriptors for weighted molecular graphs. *J. Chem. Inf. Comput. Sci.*, **40**, 1412–1422.
- Ivanciuc, O. (2001a) 3D QSAR models, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 233–280.
- Ivanciuc, O. (2001b) Design of topological indices. Part 18. Modeling the physical properties of alkanes with molecular graph descriptors derived from the Hosoya operator. *Rev. Roum. Chim.*, **46**, 129–141.
- Ivanciuc, O. (2001c) Design of topological indices. Part 19. Computation of vertex and molecular graph structural descriptors with operators. *Rev. Roum. Chim.*, **46**, 243–253.
- Ivanciuc, O. (2001d) Design of topological indices. Part 22. Structural descriptors computed from truncated molecular matrices. *Rev. Roum. Chim.*, **46**, 411–420.
- Ivanciuc, O. (2001e) Design of topological indices. Part 23. Structural descriptors derived from the distance-path matrix of vertex- and edge-weighted molecular graphs. *Rev. Roum. Chim.*, **46**, 543–552.
- Ivanciuc, O. (2001f) Design of topological indices. Part 25. Burden molecular matrices and derived structural descriptors for glycine antagonists QSAR models. *Rev. Roum. Chim.*, **46**, 1047–1066.
- Ivanciuc, O. (2001g) Design of topological indices. Part 26. Structural descriptors computed from the Laplacian matrix of weighted molecular graphs: modeling the aqueous solubility of aliphatic alcohols. *Rev. Roum. Chim.*, **46**, 1331–1347.
- Ivanciuc, O. (2001h) New neural networks for structure–property models, in *QSPR/QSAR*

- Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 213–231.
- Ivanciuc, O. (2002a) Building-block computation of the Ivanciuc–Balaban indices for the virtual screening of combinatorial libraries. *Internet Electron. J. Mol. Des.*, **1**, 1–9.
- Ivanciuc, O. (2002b) Design of topological indices. Part 27. Szeged matrix for vertex- and edge-weighted molecular graphs as a source of structural descriptors for QSAR models. *Rev. Roum. Chim.*, **47**, 479–492.
- Ivanciuc, O. (2002c) Design of topological indices. Part 28. Distance complement matrix and related structural descriptors for QSAR and QSPR models. *Rev. Roum. Chim.*, **47**, 577–594.
- Ivanciuc, O. (2002d) Design of topological indices. Part 29. QSAR and QSPR structural descriptors from the resistance distance matrix. *Rev. Roum. Chim.*, **47**, 675–686.
- Ivanciuc, O. (2002e) Structure–odor relationships for pyrazines with support vector machines. *Internet Electron. J. Mol. Des.*, **1**, 269–284.
- Ivanciuc, O. (2002f) Support vector machine classification of the carcinogenic activity of polycyclic aromatic hydrocarbons. *Internet Electron. J. Mol. Des.*, **1**, 203–218.
- Ivanciuc, O. (2002g) Support vector machine identification of the aquatic toxicity mechanism of organic compounds. *Internet Electron. J. Mol. Des.*, **1**, 157–172.
- Ivanciuc, O. (2003a) Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. *Internet Electron. J. Mol. Des.*, **2**, 195–208.
- Ivanciuc, O. (2003b) Canonical numbering and constitutional symmetry, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 139–160.
- Ivanciuc, O. (2003c) Graph theory in chemistry, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 103–138.
- Ivanciuc, O. (2003d) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.*, **31**, 359–362.
- Ivanciuc, O. (2003e) Support vector machines classification of black and green teas based on their metal content. *Internet Electron. J. Mol. Des.*, **2**, 348–357.
- Ivanciuc, O. (2003f) Topological indices, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 981–1003.
- Ivanciuc, O. (2004a) Similarity matrices quantitative structure–activity relationships for anticonvulsant phenylacetanilides. *Internet Electron. J. Mol. Des.*, **3**, 426–442.
- Ivanciuc, O. (2004b) Support vector machines prediction of the mechanism of toxic action from hydrophobicity and experimental toxicity against *Pimephales promelas* and *Tetrahymena pyriformis*. *Internet Electron. J. Mol. Des.*, **3**, 802–821.
- Ivanciuc, O. (2005) Support vector regression quantitative structure–activity relationships (QSAR) for benzodiazepine receptor ligands. *Internet Electron. J. Mol. Des.*, **4**, 181–193.
- Ivanciuc, O. (2008) Electrotopological state indices, in *Molecular Drug Properties*, Vol. 37 (ed. R. Mannhold), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 85–109.
- Ivanciuc, O., Babic, D. and Balaban, A.T. (1999) Correlation between strain energies of proper fullerenes and their topological invariants. Part II. Fullerenes with isolated pentagons. *Fullerene Sci. Technol.*, **7**, 1–15.
- Ivanciuc, O. and Balaban, A.T. (1992a) Nonisomorphic graphs with identical atomic counts of self-returning walks: isocodal graphs. *J. Math. Chem.*, **11**, 155–167.
- Ivanciuc, O. and Balaban, A.T. (1992b) Recurrence relationships for the computation of Kekulé structures. *J. Math. Chem.*, **11**, 169–177.
- Ivanciuc, O. and Balaban, A.T. (1994a) Design of topological indices. Part 5. Precision and error in computing graph theoretic invariants for molecules containing heteroatoms and multiple bonds. *MATCH Commun. Math. Comput. Chem.*, **30**, 117–139.
- Ivanciuc, O. and Balaban, A.T. (1994b) Design of topological indices. Part 8. Path matrices and derived molecular graph invariants. *MATCH Commun. Math. Comput. Chem.*, **30**, 141–152.
- Ivanciuc, O. and Balaban, A.T. (1996a) Characterization of chemical structures by the atomic counts of self-returning walks: on the construction of isocodal graphs. *Croat. Chem. Acta*, **69**, 63–74.
- Ivanciuc, O. and Balaban, A.T. (1996b) Design of topological indices. Part 3. New identification numbers for chemical structures: MINID and MINSID. *Croat. Chem. Acta*, **69**, 9–16.
- Ivanciuc, O. and Balaban, A.T. (1996c) Design of topological indices. Part 6. A new topological parameter for the steric effect of alkyl substituents. *Croat. Chem. Acta*, **69**, 75–83.
- Ivanciuc, O. and Balaban, A.T. (1999a) Design of topological indices. Part 20. Molecular structure

- descriptors computed with information on distance operators. *Rev. Roum. Chim.*, **44**, 479–489.
- Ivanciuc, O. and Balaban, A.T. (1999b) Design of topological indices. Part 21. Molecular graph operators for the computation of geometric structural descriptors. *Rev. Roum. Chim.*, **44**, 539–547.
- Ivanciuc, O. and Balaban, A.T. (1999c) The graph description of chemical structures, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 59–167.
- Ivanciuc, O., Balaban, T.-S. and Balaban, A.T. (1993a) Chemical graphs with degenerate topological indices based on information on distances. *J. Math. Chem.*, **14**, 21–33.
- Ivanciuc, O., Balaban, T.-S. and Balaban, A.T. (1993b) Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. *J. Math. Chem.*, **12**, 309–318.
- Ivanciuc, O., Balaban, T.-S., Filip, P. and Balaban, A.T. (1992) Design of topological indices. Part 7. Analytical formulae for local vertex invariants of linear and monocyclic molecular graphs. *MATCH Commun. Math. Comput. Chem.*, **28**, 151–164.
- Ivanciuc, O. and Devillers, J. (1999) Algorithms and software for the computation of topological indices and structure–property models, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 779–804.
- Ivanciuc, O., Diudea, M.V. and Khadikar, P.V. (1998) New topological matrices and their polynomials. *Indian J. Chem.*, **37**, 574–585.
- Ivanciuc, O. and Ivanciuc, T. (1999) Matrices and structural descriptors computed from molecular graphs distances, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 221–277.
- Ivanciuc, O., Ivanciuc, T. and Balaban, A.T. (1998a) Design of topological indices. Part 10. Parameters based on electronegativity and covalent radius for the computation of molecular graph descriptors for heteroatom-containing molecules. *J. Chem. Inf. Comput. Sci.*, **38**, 395–401.
- Ivanciuc, O., Ivanciuc, T. and Balaban, A.T. (1998b) Quantitative structure–property relationship study of normal boiling points for halogen-/oxygen-/sulfur-containing organic compounds using the CODESSA program. *Tetrahedron*, **54**, 9129–9142.
- Ivanciuc, O., Ivanciuc, T. and Balaban, A.T. (1999a) Vertex- and edge-weighted molecular graphs and derived structural descriptors, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 169–220.
- Ivanciuc, O., Ivanciuc, T. and Balaban, A.T. (2000a) The complementary distance matrix, a new molecular graph metric. *ACH - Models Chem.*, **137**, 57–82.
- Ivanciuc, O., Ivanciuc, T. and Balaban, A.T. (2002a) QSAR models for the dermal penetration of polycyclic aromatic hydrocarbons. *Internet Electron. J. Mol. Des.*, **1**, 559–571.
- Ivanciuc, O., Ivanciuc, T. and Balaban, A.T. (2002b) Quantitative structure–property relationship evaluation of structural descriptors derived from the distance and reverse Wiener matrices. *Internet Electron. J. Mol. Des.*, **1**, 467–487.
- Ivanciuc, O., Ivanciuc, T. and Balaban, A.T. (2002c) Quantitative structure–property relationships for the normal boiling temperatures of acyclic carbonyl compounds. *Internet Electron. J. Mol. Des.*, **1**, 252–268.
- Ivanciuc, O., Ivanciuc, T. and Cabrol-Bass, D. (2000b) 3D quantitative structure–activity relationships with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists. *Analysis*, **28**, 637–642.
- Ivanciuc, O., Ivanciuc, T. and Cabrol-Bass, D. (2001a) Comparative receptor surface analysis (CoRSA) model for calcium channel antagonists. *SAR & QSAR Environ. Res.*, **12**, 93–111.
- Ivanciuc, O., Ivanciuc, T. and Cabrol-Bass, D. (2002d) QSAR for dihydrofolate reductase inhibitors with molecular graph structural descriptors. *J. Mol. Struct. (Theochem)*, **582**, 39–51.
- Ivanciuc, O., Ivanciuc, T., Cabrol-Bass, D. and Balaban, A.T. (2000c) Comparison of weighting schemes for molecular graph descriptors: application in quantitative structure–retention relationship models for alkylphenols in gas–liquid chromatography. *J. Chem. Inf. Comput. Sci.*, **40**, 732–743.
- Ivanciuc, O., Ivanciuc, T., Cabrol-Bass, D. and Balaban, A.T. (2000d) Evaluation in quantitative structure–property relationship models of structural descriptors derived from information-theory operators. *J. Chem. Inf. Comput. Sci.*, **40**, 631–643.
- Ivanciuc, O., Ivanciuc, T., Cabrol-Bass, D. and Balaban, A.T. (2000e) Investigation of alkane branching (and resulting partial ordering) by

- topological indices. *MATCH Commun. Math. Comput. Chem.*, **42**, 155–180.
- Ivanciuc, O., Ivanciuc, T., Cabrol-Bass, D. and Balaban, A.T. (2002e) Optimum structural descriptors derived from the Ivanciuc–Balaban operator. *Internet Electron. J. Mol. Des.*, **1**, 319–331.
- Ivanciuc, O., Ivanciuc, T. and Diudea, M.V. (1997) Molecular graph matrices and derived structural descriptors. *SAR & QSAR Environ. Res.*, **7**, 63–87.
- Ivanciuc, O., Ivanciuc, T. and Diudea, M.V. (1999b) Polynomials and spectra of molecular graphs. *Roum. Chem. Quart. Rev.*, **7**, 41–67.
- Ivanciuc, O., Ivanciuc, T., Filip, P.A. and Cabrol-Bass, D. (1999c) Estimation of the liquid viscosity of organic compounds with a quantitative structure–property model. *J. Chem. Inf. Comput. Sci.*, **39**, 515–524.
- Ivanciuc, O., Ivanciuc, T. and Klein, D.J. (2001b) Intrinsic graph distances compared to Euclidean distances for correspondent graph embedding. *MATCH Commun. Math. Comput. Chem.*, **44**, 251–278.
- Ivanciuc, O., Ivanciuc, T. and Klein, D.J. (2001c) Quantitative structure–property relationships generated with optimizable even/odd Wiener polynomial descriptors. *SAR & QSAR Environ. Res.*, **12**, 1–16.
- Ivanciuc, O., Ivanciuc, T., Klein, D.J., Seitz, W.A. and Balaban, A.T. (2001d) Quantitative structure–retention relationships for gas chromatographic retention indices of alkylbenzenes with molecular graph descriptors. *SAR & QSAR Environ. Res.*, **11**, 419–452.
- Ivanciuc, O., Ivanciuc, T., Klein, D.J., Seitz, W.A. and Balaban, A.T. (2001e) Wiener index extension by counting even/odd graph distances. *J. Chem. Inf. Comput. Sci.*, **41**, 536–549.
- Ivanciuc, O. and Klein, D.J. (2002a) Building-block computation of Wiener-type indices for the virtual screening of combinatorial libraries. *Croat. Chem. Acta*, **75**, 577–601.
- Ivanciuc, O. and Klein, D.J. (2002b) Computing Wiener-type indices for virtual combinatorial libraries generated from heteroatom-containing building blocks. *J. Chem. Inf. Comput. Sci.*, **42**, 8–22.
- Ivanciuc, O., Laidboeur, T. and Cabrol-Bass, D. (1997) Degeneracy of topologic distance descriptors for cubic molecular graphs: examples of small fullerenes. *J. Chem. Inf. Comput. Sci.*, **37**, 485–488.
- Ivanciuc, O., Mathura, V.S., Midoro-Horiuti, T., Braun, W., Goldblum, R.M. and Schein, C.H. (2003) Detecting potential IgE-reactive sites on food proteins using a sequence and structure database, SDAP-food. *J. Agr. Food Chem.*, **51**, 4830–4837.
- Ivanciuc, O., Oezguen, N., Mathura, V.S., Schein, C. H. and Braun, W. (2004) Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr. Med. Chem.*, **11**, 583–593.
- Ivanciuc, O., Rabine, J.-P. and Cabrol-Bass, D. (1997) ¹³C NMR chemical shift sum prediction for alkanes using neural networks. *Computers Chem.*, **21**, 437–443.
- Ivanciuc, O., Schein, C.H. and Braun, W. (2002) Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics*, **18**, 1358–1364.
- Ivanciuc, O., Taraviras, S.L. and Cabrol-Bass, D. (2000) Quasi-orthogonal basis sets of molecular graph descriptors as a chemical diversity measure. *J. Chem. Inf. Comput. Sci.*, **40**, 126–134.
- Ivanciuc, T. and Ivanciuc, O. (2002) Quantitative structure–retention relationship study of gas chromatographic retention indices for halogenated compounds. *Internet Electron. J. Mol. Des.*, **1**, 94–107.
- Ivanciuc, T., Ivanciuc, O. and Klein, D.J. (2005) Posetic quantitative superstructure/activity relationships (QSSARs) for chlorobenzenes. *J. Chem. Inf. Model.*, **45**, 870–879.
- Ivanov, J., Karabunarliev, S. and Mekenyan, O. (1994) 3DGEN: a system for exhaustive 3D molecular design proceeding from molecular topology. *J. Chem. Inf. Comput. Sci.*, **34**, 234–243.
- Ivanov, J. and Schüürmann, G. (1999) Simple algorithms for determining the molecular symmetry. *J. Chem. Inf. Comput. Sci.*, **39**, 728–737.
- Ivanova, A.A., Palyulin, V.A., Zefirov, A.N. and Zefirov, N.S. (2004) Fragment descriptors in QSPR: application to heat capacity calculation. *Russ. J. Org. Chem.*, **40**, 644–649.
- Ivanusević, M., Nikolić, S. and Trinajstić, N. (1991) A QSAR study of antidotal activity of H-oximes. *Rev. Roum. Chim.*, **36**, 389–398.
- Iwase, K., Komatsu, K., Hirono, S., Nakagawa, S. and Moriguchi, I. (1985) Estimation of hydrophobicity based on the solvent-accessible surface area of molecules. *Chem. Pharm. Bull.*, **33**, 2114–2121.
- Iyer, M. and Hopfinger, A.J. (2007) Treating chemical diversity in QSAR analysis: modeling diverse HIV-1 integrase inhibitors using 4D fingerprints. *J. Chem. Inf. Model.*, **47**, 1945–1960.
- Iyer, M., Mishra, R., Han, Y. and Hopfinger, A.J. (2002) Predicting blood–brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharm. Res.*, **19**, 1611–1621.
- Iyer, M., Tseng, Y.J., Senese, C.L., Liu, J. and Hopfinger, A.J. (2007) Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. *Mol. Pharm.*, **4**, 218–231.

- Iyer, M., Zheng, T., Hopfinger, A.J. and Tseng, Y.J. (2007) QSAR analyses of skin penetration enhancers. *J. Chem. Inf. Model.*, **47**, 1130–1149.
- Izrailev, S. and Agrafiotis, D.K. (2001a) A new method for building regression tree models for QSAR based on artificial ant colony systems. *J. Chem. Inf. Comput. Sci.*, **41**, 176–180.
- Izrailev, S. and Agrafiotis, D.K. (2001b) Variable selection for QSAR by artificial ant colony systems. *SAR & QSAR Environ. Res.*, **13**, 417–423.
- Izrailev, S. and Agrafiotis, D.K. (2004) A method for quantifying and visualizing the diversity of QSAR models. *J. Mol. Graph. Model.*, **22**, 275–284.
- Jackel, H. and Nendza, M. (1994) Reactive substructures in the prediction of aquatic toxicity data. *Aquat. Toxicol.*, **29**, 305–314.
- Jackson, J.E. (1991) *A User's Guide to Principal Components*, John Wiley & Sons, Inc., New York, p. 570.
- Jacobs, M.N. (2004) *In silico* tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology*, **205**, 43–53.
- Jaén-Oltra, J., Salabert-Salvador, M.T., García-March, F.J., Pérez-Giménez, F. and Tomás-Vert, F. (2000) Artificial neural network applied to prediction of fluoroquinolone antibacterial activity by topological methods. *J. Med. Chem.*, **43**, 1143–1148.
- Jaffé, H.H. (1953) A reexamination of the Hammett equation. *Chem. Rev.*, **53**, 191–261.
- Jafvert, C.T., Chu, W. and Vanhoof, P.L. (1995) A quantitative structure–activity relationship for solubilization of nonpolar compounds by nonionic surfactant micelles. *ACS Symp. Ser.*, **594**, 24–37.
- Jäger, R., Kast, S.M. and Brickmann, J. (2003) Parametrization strategy for the MolFESD concept: quantitative surface representation of local hydrophobicity. *J. Chem. Inf. Comput. Sci.*, **43**, 237–247.
- Jaguar, Schrödinger, LLC, New York.
- Jain, A., Yang, G. and Yalkowsky, S.H. (2004a) Estimation of melting points of organic compounds. *Ind. Eng. Chem. Res.*, **43**, 7618–7621.
- Jain, A., Yang, G. and Yalkowsky, S.H. (2004b) Estimation of total entropy of melting of organic compounds. *Ind. Eng. Chem. Res.*, **43**, 4376–4379.
- Jain, A.N. (2000) Morphological similarity: a 3D molecular similarity method correlated with protein–ligand recognition. *J. Comput. Aid. Mol. Des.*, **14**, 199–213.
- Jain, A.N., Dietterich, T.G., Lathrop, R.H., Chapman, D., Critchlow, R.E., Bauer, B.E., Webster, T.A. and Lozano-Perez, T. (1994) Compass: a shape-based machine learning tool for drug design. *J. Comput. Aid. Mol. Des.*, **8**, 635–652.
- Jain, A.N., Harris, N.L. and Park, J.Y. (1995) Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5-HT_{1A} receptor. *J. Med. Chem.*, **38**, 1295–1308.
- Jain, A.N., Koile, K. and Chapman, D. (1994) Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.*, **37**, 2315–2327.
- Jain, N. and Yalkowsky, S.H. (2001) Estimation of the aqueous solubility. I. Application to organic nonelectrolytes. *J. Pharm. Sci.*, **90**, 234–252.
- Jaiswal, M. and Khadikar, P.V. (2004) QSAR study on tadpole narcosis using PI index: a case of heterogeneous set of compounds. *Bioorg. Med. Chem.*, **12**, 1731–1736.
- Jalali-Heravi, M. and Fatemi, M.H. (2001) Artificial neural network modeling of Kovats retention indices for noncyclic and monocyclic terpenes. *J. Chromat.*, **915**, 177–183.
- Jalali-Heravi, M. and Garkani-Nejad, Z. (2002a) Prediction of electrophoretic mobilities of alkyl- and alkenylpyridines in capillary electrophoresis using artificial neural networks. *J. Chromat.*, **971**, 207–215.
- Jalali-Heravi, M. and Garkani-Nejad, Z. (2002b) Prediction of relative response factors for flame ionization and photoionization detection using self-training artificial neural networks. *J. Chromat.*, **950**, 183–194.
- Jalali-Heravi, M. and Garkani-Nejad, Z. (2002c) Use of self-training artificial neural networks in modeling of gas chromatographic relative retention times of a variety of organic compounds. *J. Chromat.*, **945**, 173–184.
- Jalali-Heravi, M. and Konouz, E. (2005) Use of quantitative structure–activity relationships in prediction of cmc of nonionic surfactants. *Quant. Struct. -Act. Relat.*, **19**, 135–141.
- Jalali-Heravi, M. and Kyani, A. (2004) Use of computer-assisted methods for the modeling of the retention time of a variety of volatile organic compounds: a PCA-MLR-ANN approach. *J. Chem. Inf. Comput. Sci.*, **44**, 1328–1335.
- Jalali-Heravi, M., Noroozian, E. and Mousavi, M. (2004) Prediction of relative response factors of electron-capture detection for some polychlorinated biphenyls using chemometrics. *J. Chromat.*, **1023**, 247–254.
- Jalali-Heravi, M. and Parastar, F. (1999) Computer modeling of the rate of glycine conjugation of some benzoic acid derivatives: a QSAR study. *Quant. Struct. -Act. Relat.*, **18**, 134–138.
- Jalbout, A.F. and Li, X. (2003) Anti-HIV-1 inhibitors of various molecules using principles of connectivity. *J. Mol. Struct. (Theochem)*, **663**, 19–23.

- Jalbout, A.F. and Li, X. (2003) Bond order weighted Wiener numbers. *J. Mol. Struct. (Theochem)*, **663**, 9–14.
- Jalbout, A.F. and Li, X. (2003) Topological index-quantum chemical bond order relations. *J. Mol. Struct. (Theochem)*, **638**, 1–4.
- Jalbout, A.F., Zhou, Z.Y., Li, X., Solimannejad, M. and Ma, Y. (2003) Molecular connectivity relationships to electrostatic potential derived parameters. *J. Mol. Struct. (Theochem)*, **664–665**, 15–19.
- Jamois, E.A., Hassan, M. and Waldman, M. (2000) Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.*, **40**, 63–70.
- Janežič, D., Lučić, B., Miličević, A., Nikolić, J., Trinajstić, N. and Vukicević, D. (2007) Hosoya matrices as the numerical realization of graphical matrices and derived structural descriptors. *Croat. Chem. Acta*, **80**, 271–276.
- Janežič, D., Lučić, B., Nikolić, S., Miličević, A. and Trinajstić, N. (2006) Boiling points of alcohols – a comparative QSPR study. *Internet Electron. J. Mol. Des.*, **5**, 192–200.
- Janežič, D., Miličević, A., Nikolić, S. and Trinajstić, N. (2007) *Graph Theoretical Matrices in Chemistry*. University of Kragujevac, Kragujevac, Serbia, 205.
- Janini, G.M., Johnston, K. and Zielinski, W.L., Jr (1975) Use of a nematic liquid crystal for gas–liquid chromatographic separation of polyaromatic hydrocarbons. *Anal. Chem.*, **47**, 670–674.
- Jäntschi, L. (2004a) MDF – a new QSPR/QSAR molecular descriptors family. *Leonardo Journal of Sciences*, **4**, 68–85.
- Jäntschi, L. (2004b) Water activated carbon organics adsorption structure–property relationships. *Leonardo Journal of Sciences*, **5**, 63–73.
- Jäntschi, L. (2005) Molecular descriptors family on structure–activity relationships. I. Review of the methodology. *Leonardo Electron. J. Pract. Technol.*, **6**, 76–98.
- Jäntschi, L. and Bolboacă, S. (2005) Molecular descriptors family on structure–activity relationships. 4. Molar refraction of cyclic organophosphorus compounds. *Leonardo Electron. J. Pract. Technol.*, **7**, 55–102.
- Jäntschi, L. and Bolboacă, S. (2006) Molecular descriptors family on structure–activity relationships. 5. Antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivatives. *Leonardo Journal of Sciences*, **8**, 77–88.
- Jäntschi, L. and Bolboacă, S. (2007) Results from the use of molecular descriptors family on structure–property/activity relationships. *Int. J. Mol. Sci.*, **8**, 189–203.
- Jäntschi, L., Katona, G. and Diudea, M.V. (2000) Modeling molecular properties by Cluj indices. *MATCH Commun. Math. Comput. Chem.*, **41**, 151–188.
- Japertas, P., Didziapetris, R. and Petrauskas, A. (2002) Fragmental methods in the design of new compounds. Applications of the advanced algorithm builder. *Quant. Struct. -Act. Relat.*, **21**, 23–37.
- Japertas, P., Didziapetris, R. and Petrauskas, A. (2003) Fragmental methods in the analysis of biological activities of diverse compound sets. *Mini Rev. Med. Chem.*, **3**, 797–808.
- Jaworska, J.S., Comber, M., Auer, C. and Van Leeuwen, C.J. (2003) Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ. Health Persp.*, **111**, 1358–1360.
- Jaworska, J.S., Nikolova-Jeliazkova, N. and Aldenberg, T. (2004) Review of methods for applicability domain estimation. Report, The European Commission – Joint Research Centre, Ispra, Italy.
- Jaworska, J.S., Nikolova-Jeliazkova, N. and Aldenberg, T. (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *ATLA*, **33**, 445–459.
- Jaworska, J.S. and Schultz, T.W. (1993) Quantitative relationships of structure–activity and volume fraction for selected nonpolar and polar narcotic chemicals. *SAR & QSAR Environ. Res.*, **1**, 3–19.
- Jefford, C.W., Grigorov, M., Weber, J., Lüthi, H.P. and Tronchet, J.M.J. (2000) Correlating the molecular electrostatic potentials of some organic peroxides with their antimalarial activities. *J. Chem. Inf. Comput. Sci.*, **40**, 354–357.
- Jeffrey, H.J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res.*, **18**, 2163–2170.
- Jelcic, Z. (2004) Solvent molecular descriptors on poly (D,L-lactide-co-glycolide) particle size in emulsification-diffusion process. *Coll. Surf. A Physico-chem. Eng. Aspects*, **242**, 159–166.
- Jelfs, S.P., Ertl, P. and Selzer, P. (2007) Estimation of pK_a for druglike compounds using semiempirical and information-based descriptors. *J. Chem. Inf. Model.*, **47**, 450–459.
- Jenkins, H.D.B., Kelly, E.J. and Samuel, C.J. (1994) A novel computational approach to the estimation of steric parameters application to the Menshutkin reaction. *Tetrahedron Lett.*, **35**, 6543–6546.
- Jenkins, H.D.B., Samuel, C.J. and Stafford, J.E. (1995) A novel computational approach to the estimation of steric parameters. II. Extension to thiazoles. *Tetrahedron Lett.*, **36**, 6159–6162.

- Jensen, B.F., Refsgaard, H.H.F., Bro, R. and Brockhoff, P.B. (2005) Classification of membrane permeability of drug candidates: a methodological investigation. *QSAR Comb. Sci.*, **24**, 449–457.
- Jensen, B.F., Sørensen, M.D., Kissmeyer, A.-M., Björklund, F., Sonne, K., Engelsen, S.B. and Nørgaard, L. (2003) Prediction of *in vitro* metabolic stability of calcitriol analogs by QSAR. *J. Comput. Aid. Mol. Des.*, **17**, 849–859.
- Jensen, B.F., Vind, C., Padkjær, S.B., Brockhoff, P.B. and Refsgaard, H.H.F. (2007) *In silico* prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted *k*-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.*, **50**, 501–511.
- Jerman-Blazic Dzonova, B. and Trinajstić, N. (1982) Computer-aided enumeration and generation of the Kekulé structures in conjugated hydrocarbons. *Computers Chem.*, **6**, 121–132.
- Jerman-Blazic, B., Fabic-Petrac, I. and Randić, M. (1989) Evaluation of the molecular similarity and property prediction for QSAR purposes. *Chemom. Intell. Lab. Syst.*, **6**, 49–63.
- Jetter, K., Depczynski, U., Molt, K. and Niemöller, A. (2000) Principles and applications of wavelet transformation to chemometrics. *Anal. Chim. Acta*, **420**, 169–180.
- Jewell, N.E., Turner, D.B., Willett, P. and Sexton, G.J. (2001) Automatic generation of alignments for 3D QSAR analyses. *J. Mol. Graph. Model.*, **20**, 111–121.
- Jezierska, A., Vračko, M. and Basak, S.C. (2004) Counter-propagation artificial neural network as a tool for the independent variable selection: structure–mutagenicity study of aromatic amines. *Mol. Div.*, **8**, 371–377.
- Jiang, C., Li, Y., Tian, Q. and You, T.-P. (2003) QSAR study of catalytic asymmetric reactions with topological indices. *J. Chem. Inf. Comput. Sci.*, **43**, 1876–1881.
- Jiang, H., Chen, K., Wang, H.W., Tang, Y., Chen, J.Z. and Ji, R.Y. (1994) 3D QSAR study on ether and ester analogs of artemisinin with comparative molecular field analysis. *Acta Pharmacol. Sin.*, **15**, 481.
- Jiang, S., Liang, H. and Bai, F. (2006) New structural parameters and permanents of adjacency matrices of fullerenes. *MATCH Commun. Math. Comput. Chem.*, **56**, 131–139.
- Jiang, Y.-R., Liu, J.-Y., Hu, Y.-H. and Fujita, T. (2003) Novel topological index for research on structure–property relationships of complex organic compounds. *J. Comput. Chem.*, **24**, 842–849.
- Jiang, Y., Qian, X. and Shao, Y. (1995) The evaluation of moments for benzenoid hydrocarbons. *Theor. Chim. Acta*, **90**, 135–144.
- Jiang, Y., Tang, A. and Hoffmann, R.D. (1984) Evaluation of moments and their application in Hückel molecular orbital theory. *Theor. Chim. Acta*, **66**, 183–192.
- Jiang, Y. and Zhang, H. (1989) Stability and reactivity based on moment analysis. *Theor. Chim. Acta*, **75**, 279–297.
- Jiang, Y. and Zhang, H. (1990) Aromaticities and reactivities based on energy partitioning. *Prot. Struct. Funct. Gen.*, **62**, 451–456.
- Jiang, Y. and Zhu, H. (1994) Evaluation of level pattern indices. *J. Chem. Inf. Comput. Sci.*, **34**, 377–380.
- Jiang, Y., Zhu, H., Zhang, H. and Gutman, I. (1989) Moment expansion of Hückel molecular energies. *Chem. Phys. Lett.*, **159**, 159–164.
- Jiskra, J., Claessens, H.A., Cramers, C.A. and Kaliszan, A. (2002) Quantitative structure–retention relationships in comparative studies of behavior of stationary phases under high-performance liquid chromatography and capillary electrochromatography conditions. *J. Chromat.*, **977**, 193–206.
- Joao, H.C., Devreese, K., Pauwels, R., Declercq, E., Henson, G.W. and Bridger, G.J. (1995) Quantitative structural activity relationship study of bis-tetraazacyclic compounds. A novel series of HIV-1 and HIV-2 inhibitors. *J. Med. Chem.*, **38**, 3865–3873.
- Jochum, C. and Gasteiger, J. (1977) Canonical numbering and constitutional symmetry. *J. Chem. Inf. Comput. Sci.*, **17**, 113–117.
- Jochum, C., Hicks, M.G. and Sunkel, J. (eds) (1988) *Physical Property Prediction in Organic Chemistry*, Springer-Verlag, Berlin, Germany, p. 554.
- John, P.E. and Diudea, M.V. (2004) The second path matrix of the graph and its characteristic polynomial. *Carpathian J. Math.*, **2**, 235–239.
- John, P.E., Khadikar, P.V. and Singh, J. (2007) A method of computing the PI index of benzenoid hydrocarbons using orthogonal cuts. *J. Math. Chem.*, **42**, 37–45.
- John, P.E., Mallion, R.B. and Gutman, I. (1998) An algorithm for counting spanning trees in labeled molecular graphs homeomorphic to cata-condensed systems. *J. Chem. Inf. Comput. Sci.*, **38**, 108–112.
- Johnson, C.D. (1973) *The Hammett Equation*, Cambridge University Press, Cambridge, UK.
- Johnson, J.L.H. and Yalkowsky, S.H. (2005) Two new parameters for predicting the entropy of melting:

- eccentricity (ϵ) and spirality (μ). *Ind. Eng. Chem. Res.*, **44**, 7559–7566.
- Johnson, M.A. (1989) A review and examination of mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.*, **3**, 117–145.
- Johnson, M.A., Basak, S.C. and Maggiora, G.M. (1998) Characterization of molecular similarity methods for property prediction. *Math. Comput. Modelling*, **11**, 630–635.
- Johnson, M.A., Gifford, E. and Tsai, C.-C. (1990) Similarity concepts in modeling chemical transformation pathways, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 289–320.
- Johnson, M.A. and Maggiora, G.M. (eds) (1990) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, Inc., New York, p. 393.
- Johnson, R.A. and Wichern, D.W. (1992) *Applied Multivariate Analysis*, Prentice-Hall, Englewood Cliffs, NJ, p. 642.
- Johnson, S.R. and Jurs, P.C. (1999) Prediction of the clearing temperatures of a series of liquid crystals from molecular structure. *Chem. Mater.*, **11**, 1007–1023.
- Jórárt, B., Martinek, T.A. and Márki, Á. (2005) The 3D structure of the binding pocket of the human oxytocin receptor for benzoxazine antagonists, determined by molecular docking, scoring functions and 3D-QSAR methods. *J. Comput. Aid. Mol. Des.*, **19**, 341–356.
- Jolles, G. and Woolridge, K.R.H. (eds) (1984) *Drug Design: Fact or Fantasy?* Academic Press, London, UK.
- Jolliffe, I.T. (1972) Discarding variables in a principal component analysis. I. Artificial data. *Appl. Stat.*, **21**, 160–173.
- Jolliffe, I.T. (1973) Discarding variables in a principal component analysis. II. Real data. *Appl. Stat.*, **22**, 21–31.
- Jolliffe, I.T. (1986) *Principal Component Analysis*, Springer-Verlag, New York, p. 272.
- Jonathan, P., McCarthy, W.V. and Roberts, A.M. (1996) Discriminant analysis with singular covariance matrices. A method incorporating cross-validation and efficient randomized permutation tests. *J. Chemom.*, **10**, 189–213.
- Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M. and Wold, S. (1989) Multivariate parametrization of 55 coded and non-coded amino acids. *Quant. Struct. -Act. Relat.*, **8**, 204–209.
- Jonsson, J., Norberg, T. and Carlsson, L. (1993) Quantitative sequence-activity models (QSAM) – tools for sequence design. *Nucleic Acids Res.*, **20**, 733–739.
- Jordan, S.N., Leach, A.R. and Bradshaw, J. (1995) The application of neural networks in conformational analysis. 1. Prediction of minimum and maximum interatomic distances. *J. Chem. Inf. Comput. Sci.*, **35**, 640–650.
- Jørgensen, F.S., Jensen, L.H., Capion, D. and Christensen, I.T. (2001) Prediction of blood-brain barrier penetration, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 281–285.
- Jørgensen, P., Olsen, J. and Helgaker, T. (2000) *Molecular Electronic-Structure Theory*, John Wiley & Sons, Ltd, Chichester, UK, p. 938.
- Jørgensen, W.L. and Duffy, E.M. (2000) Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.*, **10**, 1155–1158.
- Jørgensen, W.L. and Duffy, E.M. (2002) Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.*, **54**, 355–366.
- Joshi, R.K., Meister, Th., Scapozza, L. and Ha, T.-K. (1993) Development of new molecular descriptors using conformational energies from quantum calculations and their application in QSAR analysis, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 362–363.
- Joshi, R.K., Meister, Th., Scapozza, L. and Ha, T.-K. (1994) A new quantum chemical approach in QSAR-analysis. Parametrisation of conformational energies into molecular descriptors Jmn (steric) and Jsn (electronic). *Arzneim. Forsch. (German)*, **44**, 779–790.
- Jouan-Rimbaud, D., Massart, D.L. and de Noord, O.E. (1996) Random correlation in variable selection for multivariate calibration with a genetic algorithm. *Chemom. Intell. Lab. Syst.*, **35**, 213–220.
- Jouan-Rimbaud, D., Walczak, B., Poppi, R.J., de Noord, O.E. and Massart, D.L. (1997) Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration. *Anal. Chem.*, **69**, 4317–4323.
- Joubert, L., Guillemoles, J.-F. and Adamo, C. (2003) A theoretical investigation of the dye-redox mediator interaction in dye-sensitized photovoltaic cells. *Chem. Phys. Lett.*, **371**, 378–385.
- Jover, J., Bosque, R. and Sales, J. (2004) Determination of lithium cation basicity from molecular structure. *J. Chem. Inf. Comput. Sci.*, **44**, 1727–1736.
- Joyce, S.J., Osguthorpe, D.J., Padgett, J.A. and Price, G.J. (1995) Neural-network prediction of glass-transition temperatures from monomer

- structure. *J. Chem. Soc. Faraday Trans.*, **91**, 2491–2496.
- Judson, P.N. (1992a) QSAR and expert systems in the prediction of biological activity. *Pestic. Sci.*, **36**, 155–160.
- Judson, P.N. (1992b) Structural similarity searching using descriptors developed for structure–activity relationship studies. *J. Chem. Inf. Comput. Sci.*, **32**, 657–663.
- Judson, R. (1996) Genetic algorithms and their use in chemistry, in *Reviews in Computational Chemistry*, Vol. 10 (eds K.B. Lipkowitz and D. Boyd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1–73.
- Jug, K. (1983) A bond order approach to ring current and aromaticity. *J. Org. Chem.*, **48**, 1344–1348.
- Jug, K. (1984) Bond order as a tool for molecular structure and reactivity. *Croat. Chem. Acta*, **57**, 941–953.
- Julg, A. and François, Ph. (1967) Recherches sur la géométrie de quelques hydrocarbures non-alternants: son influence sur les énergies de transition, une nouvelle définition de l'aromaticité. *Theor. Chim. Acta*, **8**, 249–259.
- Junghans, M. and Pretsch, E. (1997) Estimation of partition coefficients of organic compounds. Local database modeling with uniform-length structure descriptors. *Fresen. J. Anal. Chem.*, **359**, 88–92.
- Jurić, A., Gagro, M., Nikolić, S. and Trinajstić, N. (1992) Molecular topological index: an application in the QSAR study of toxicity of alcohols. *J. Math. Chem.*, **11**, 179–186.
- Jurić, A., Nikolić, S. and Trinajstić, N. (1997) Topological resonance energies of thienopyrimidines. *Croat. Chem. Acta*, **70**, 841–846.
- Jurs, P.C. (2003) Quantitative structure–property relationships, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1314–1335.
- Jurs, P.C., Chou, J.T. and Yuan, M. (1979) Computer-assisted structure–activity studies of chemical carcinogens. A heterogeneous data set. *J. Med. Chem.*, **22**, 476–483.
- Jurs, P.C., Dixon, J.S. and Egolf, L.M. (1995) Representations of molecules, in *Chemometrics Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 15–38.
- Jurs, P.C., Hasan, M.N., Hansen, P.J. and Rohrbaugh, R.H. (1988) Prediction of physico-chemical properties of organic compounds from molecular structure, in *Physical Property Prediction in Organic Chemistry* (eds C. Jochum, M.G. Hicks and J. Sunkel), Springer-Verlag, Berlin, Germany, pp. 209–233.
- Jurs, P.C., Hasan, M.N., Henry, D.R., Stouch, T.R. and Whalen-Pedersen, E.K. (1983) Computer-assisted studies of molecular structure and carcinogenic activity. *Fund. Appl. Toxicol.*, **3**, 343–349.
- Jurs, P.C. and Lawson, R.G. (1991) Analysis of chemical structure–biological activity relationships using clustering methods. *Chemom. Intell. Lab. Syst.*, **10**, 81–83.
- Jurs, P.C., Stouch, T.R., Czerwinski, M. and Narvaez, J.N. (1985) Computer-assisted studies of molecular structure–biological activity relationships. *J. Chem. Inf. Comput. Sci.*, **25**, 296–308.
- Juvan, M. and Mohar, B. (1995) Bond contributions to the Wiener index. *J. Chem. Inf. Comput. Sci.*, **35**, 217–219.
- Kabankin, A.S. and Gabrielyan, L.I. (2005) Relationship between structure and hepatoprotector activity of adamantane derivatives. Part 2. Application of autocorrelative, substructural and 3D molecular descriptors. *Pharm. Chem. J.*, **39**, 135–139.
- Kabankin, A.S. and Gabrielyan, L.I. (2006) Features of the principal component analysis for the classification of biologically active compounds in terms of molecular descriptors. *Pharm. Chem. J.*, **40**, 307–311.
- Kabankin, A.S. and Kurlyandskii, B.A. (2001) Discriminant analysis of the relationship between topological molecular structure and carcinogenicity of aromatic amines. *Pharm. Chem. J.*, **35**, 257–259.
- Kabankin, A.S., Radkevich, L.A., Gabrielyan, L.I., Zhestkov, V.P., Ostapchuk, N.V. and Pyn'ko, N.E. (2005) Relationship between structure and hepatoprotector activity of indole derivatives. *Pharm. Chem. J.*, **39**, 191–196.
- Kahn, I., Fara, D., Karelson, M. and Maran, U. (2005) QSPR treatment of the soil sorption coefficients of organic pollutants. *J. Chem. Inf. Model.*, **45**, 94–105.
- Kahn, I., Sild, S. and Maran, U. (2007) Modeling the toxicity of chemicals to *Tetrahymena pyriformis* using heuristic multilinear regression and heuristic back-propagation neural networks. *J. Chem. Inf. Model.*, **47**, 2271–2279.
- Kaiser, K.L.E. (2003) The use of neural networks in QSARs for acute aquatic toxicological endpoints. *J. Mol. Struct. (Theochem)*, **622**, 85–95.
- Kaiser, K.L.E. and Niculescu, S.P. (1999) Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*Pimephales promelas*): a study based on 865 compounds. *Chemosphere*, **38**, 3237–3245.

- Kaliszan, A., Nasal, A. and Turowski, M. (1996) Quantitative structure–retention relationships in the examination of the topography of the binding site of antihistamine drugs on α_1 -acid glycoprotein. *J. Chromat.*, **722**, 25–32.
- Kaliszan, A., van Straten, M.A., Markuszewski, M., Cramers, C.A. and Claessens, H.A. (1999) Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure–retention relationships. *J. Chromat.*, **855**, 455–486.
- Kaliszan, R. (1977) Correlation between the retention indices and the connectivity indices of alcohols and methyl esters with complex cyclic structure. *Chromatographia*, **10**, 529.
- Kaliszan, R. (1979) The relationship between the connectivity indices and the thermodynamic parameters describing the interaction of fat acid methyl esters with polar and nonpolar stationary phases. *Chromatographia*, **12**, 171–174.
- Kaliszan, R. (1981) Chromatography in studies of quantitative structure–activity relationships. *J. Chromat.*, **220**, 71–83.
- Kaliszan, R. (1986) Quantitative relationship between molecular structure and chromatographic retention. Implication in physical, analytical, and medicinal chemistry. *CRC Crit. Rev. Anal. Chem.*, **16**, 323–383.
- Kaliszan, R. (1987) *Quantitative Structure–Chromatographic Retention Relationships*, John Wiley & Sons, Inc., New York, p. 304.
- Kaliszan, R. (1992) Quantitative structure–retention relationships. *Anal. Chem.*, **64**, 619A–631.
- Kaliszan, R. (1993) Quantitative structure retention relationships applied to reversed phase high performance liquid chromatography. *J. Chromat.*, **656**, 417–435.
- Kaliszan, R. (2007) QSRR: quantitative structure–(chromatographic) retention relationships. *Chem. Rev.*, **107**, 3212–3246.
- Kaliszan, R. and Foks, H. (1977) The relationship between R_M values and the connectivity indices for pyrazine carbothioamide derivatives. *Chromatographia*, **10**, 346–349.
- Kaliszan, R., Kaliszan, A., Noctor, T.A., Purcell, W.P. and Wainer, I.W. (1992) Mechanism of retention of benzodiazepines in affinity, reversed phase and adsorption high performance liquid chromatography in view of quantitative structure–retention relationships. *J. Chromat.*, **609**, 69–81.
- Kaliszan, R., Kaliszan, A. and Wainer, I.W. (1993) Deactivated hydrocarbonaceous silica and immobilized artificial membrane stationary phases in high performance liquid chromatographic determination of hydrophobicities of organic bases relationship to log P and Clog P . *J. Pharm. Biomed. Anal.*, **11**, 505–511.
- Kaliszan, R. and Lamparczyk, H. (1978) A relationship between the connectivity indices and retention indices of polycyclic aromatic hydrocarbons. *J. Chromatogr. Sci.*, **16**, 246–251.
- Kaliszan, R., Lamparczyk, H. and Radecki, A. (1979) A relationship between repression of dimethylnitrosamine-demethylase by polycyclic aromatic hydrocarbons and their shape. *Biochem. Pharmacol.*, **28**, 123–125.
- Kaliszan, R., Nasal, A. and Bucinski, A. (1994) Chromatographic hydrophobicity parameter determined on an immobilized artificial membrane column relationships to standard measures of hydrophobicity and bioactivity. *Eur. J. Med. Chem.*, **29**, 163–170.
- Kaliszan, R., Noctor, T.A. and Wainer, I.W. (1992) Quantitative structure–enantioselective retention relationships for the chromatography of 1,4-benzodiazepines on a human serum albumin based HPLC chiral stationary phase: an approach to the computational prediction of retention and enantioselectivity. *Chromatographia*, **33**, 546–550.
- Kaliszan, R., Osmialowski, K., Tomellini, S.A., Hsu, S.-H., Fazio, S.D. and Hartwick, R.A. (1985) Non-empirical descriptors of sub-molecular polarity and dispersive interactions in reversed-phase HPLC. *Chromatographia*, **20**, 705–708.
- Kalivas, J.H. (1995) *Adaption of Simulated Annealing to Chemical Optimization Problems*, Elsevier, Amsterdam, The Netherlands, p. 473.
- Kalivas, J.H., Forrester, J.B. and Seipel, H.A. (2004) QSAR modeling based on the bias/variance compromise: a harmonious and parsimonious approach. *J. Comput. Aid. Mol. Des.*, **18**, 537–547.
- Kamenska, V., Mekenyan, O., Sterev, A. and Nedjalkova, Z. (1996) Application of the dynamic quantitative structure–activity relationship method for modeling antibacterial activity of quinolone derivatives. *Arzneim. Forsch. (German)*, **46**, 423–428.
- Kaminski, J.J. (1994) Computer assisted drug design and selection. *Adv. Drug Deliv. Rev.*, **14**, 331–337.
- Kamlet, M.J., Abboud, J.-L.M., Abraham, M.H. and Taft, R.W. (1983) Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, π^* , α , and β and some methods for simplifying the generalized solvatochromic equation. *J. Org. Chem.*, **48**, 2877–2887.

- Kamlet, M.J., Abboud, J.-L.M. and Taft, R.W. (1977) The solvatochromic comparison method. 6. The π^* scale of solvent polarities. *J. Am. Chem. Soc.*, **99**, 6027–6038.
- Kamlet, M.J., Abboud, J.-L.M. and Taft, R.W. (1981) An examination of linear solvation energy relationships. *Prog. Phys. Org. Chem.*, **13**, 485–630.
- Kamlet, M.J., Abraham, M.H., Doherty, R.M. and Taft, R.W. (1984) Solubility properties in polymers and biological media. 4. Correlations of octanol/water partition coefficients with solvatochromic parameters. *J. Am. Chem. Soc.*, **106**, 464–466.
- Kamlet, M.J., Carr, P.W., Taft, R.W. and Abraham, M. H. (1981) Linear solvation energy relationships. 13. Relationships between the Hildebrand solubility parameter, δ_H , and the solvatochromic parameter π^* . *J. Am. Chem. Soc.*, **103**, 6062–6066.
- Kamlet, M.J., Doherty, P.J., Taft, R.W., Abraham, M. H., Veith, G.D. and Abraham, D.J. (1987a) Solubility properties in polymers and biological media. 8. An analysis of the factors that influence toxicities of organic nonelectrolytes to the golden orfe fish (*Leuciscus idus melanotus*). *Environ. Sci. Technol.*, **21**, 149–155.
- Kamlet, M.J., Doherty, P.J., Veith, G.D., Taft, R.W. and Abraham, M.H. (1986a) Solubility properties in polymers and biological media. 7. An analysis of toxicant properties that influence inhibition of bioluminescence in *Photobacterium phosphoreum* (the Microtox test). *Environ. Sci. Technol.*, **20**, 690–695.
- Kamlet, M.J., Doherty, R.M., Abboud, J.-L.M., Abraham, M.H. and Taft, R.W. (1986b) Linear solvation energy relationships. 36. Molecular properties governing solubilities of organic nonelectrolytes in water. *J. Pharm. Sci.*, **75**, 338–349.
- Kamlet, M.J., Doherty, R.M., Abboud, J.-L.M., Abraham, M.H. and Taft, R.W. (1986c) Solubility: a new look. *Chemotherapy*, **16**, 566–576.
- Kamlet, M.J., Doherty, R.M., Abraham, M.H., Marcus, Y. and Taft, R.W. (1987b) Linear solvation energy relationships. 41. Important differences between aqueous solubility relationships for aliphatic and aromatic solutes. *J. Phys. Chem.*, **91**, 1996–2004.
- Kamlet, M.J., Doherty, R.M., Abraham, M.H., Marcus, Y. and Taft, R.W. (1988a) Linear solvation energy relationships. 46. An improved equation for correlation and prediction of octanol/water partition coefficients of organic nonelectrolytes (including strong hydrogen bond donor solutes). *J. Phys. Chem.*, **92**, 5244–5255.
- Kamlet, M.J., Doherty, R.M., Abraham, M.H. and Taft, R.W. (1988b) Solubility properties in biological media. 12. Regarding the mechanism of nonspecific toxicity or narcosis by organic nonelectrolytes. *Quant. Struct. -Act. Relat.*, **7**, 71–81.
- Kamlet, M.J., Doherty, R.M., Carr, P.W., Mackay, D., Abraham, M.H. and Taft, R.W. (1988c) Linear solvation energy relationships. 44. Parameter estimation rules that allow accurate prediction of octanol/water partition coefficients and other solubility and toxicity properties of polychlorinated biphenyls and polycyclic aromatic hydrocarbons. *Environ. Sci. Technol.*, **22**, 503–509.
- Kamlet, M.J., Doherty, R.M., Fiserova-Bergerova, V., Carr, P.W., Abraham, M.H. and Taft, R.W. (1987c) Solubility properties in polymers and biological media. 9. Prediction of solubility and partition of organic nonelectrolytes in blood and tissues from solvatochromic parameters. *J. Pharm. Sci.*, **76**, 14–17.
- Kamlet, M.J., Jones, M.E., Taft, R.W. and Abboud, J.-L. M. (1979) Linear solvation energy relationships. Part 2. Correlations of electronic spectral data for aniline indicators with solvent π^* and β values. *J. Chem. Soc. Perkin Trans. 2*, 342–348.
- Kamlet, M.J. and Taft, R.W. (1979a) Linear solvation energy relationships. Part 1. Solvent polarity–polarizability effects on infrared spectra. *J. Chem. Soc. Perkin Trans. 2*, 337–341.
- Kamlet, M.J. and Taft, R.W. (1979b) Linear solvation energy relationships. Part 3. Some reinterpretations of solvent effects based on correlations with solvent π^* and α values. *J. Chem. Soc. Perkin Trans. 2*, 349–356.
- Kang, H., Choi, H. and Park, H. (2007) Prediction of molecular solvation free energy based on the optimization of atomic solvation parameters with genetic algorithm. *J. Chem. Inf. Model.*, **47**, 509–514.
- Kang, Y.K. and Jhon, M.S. (1982) Additivity of atomic static polarizabilities and dispersion coefficients. *Theor. Chim. Acta*, **61**, 41–48.
- Kantola, A., Villar, H.O. and Loew, G.H. (1991) Atom based parametrization for a conformationally dependent hydrophobic index. *J. Comput. Chem.*, **12**, 681–689.
- Karabunarliev, S., Mekenyan, O., Karcher, W., Russom, C.L. and Bradbury, S.P. (1996a) Quantum chemical descriptors for estimating the acute toxicity of electrophiles to the fathead minnow (*Pimephales promelas*). An analysis based on molecular mechanisms. *Quant. Struct. -Act. Relat.*, **15**, 302–310.
- Karabunarliev, S., Mekenyan, O., Karcher, W., Russom, C.L. and Bradbury, S.P. (1996b) Quantum chemical descriptors for estimating the acute

- toxicity of substituted benzenes to the guppy (*Poecilia reticulata*) and fathead minnow (*Pimephales promelas*). *Quant. Struct. -Act. Relat.*, **15**, 311–320.
- Karabunarliev, S., Nikolova, N., Nikolova, N. and Mekenyan, O. (2003) Rule interpreter: a chemical language for structure-based screening. *J. Mol. Struct. (Theochem)*, **622**, 53–62.
- Karakoc, E., Sahinalp, S.C. and Cherkasov, A.R. (2006) Comparative QSAR and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. *J. Chem. Inf. Model.*, **46**, 2167–2182.
- Karcher, W. and Devillers, J. (eds) (1990a) *Practical Application of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, Kluwer Academic Publishers for the European Communities, Dordrecht, The Netherlands, p. 475.
- Karcher, W. and Devillers, J. (1990b) SAR and QSAR in environmental chemistry and toxicology: scientific tool or wishful thinking? in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 1–12.
- Karcher, W. and Karabunarliev, S. (1996) The use of computer based structure–activity relationships in the risk assessment of industrial chemicals. *J. Chem. Inf. Comput. Sci.*, **36**, 672–677.
- Karelson, M. (2000) *Molecular Descriptors in QSAR/QSPR*, Wiley–Interscience, New York, p. 430.
- Karelson, M. (2001) Electronic and electrical effects of solvents, in *Handbook of Solvents* (ed. G. Wypych), ChemTec Publishing, Toronto, Canada, pp. 639–682.
- Karelson, M., Lobanov, V.S. and Katritzky, A.R. (1996) Quantum–chemical descriptors in QSAR/QSPR studies. *Chem. Rev.*, **96**, 1027–1043.
- Karelson, M. and Perkson, A. (1999) QSPR prediction of densities of organic liquids. *Computers Chem.*, **23**, 49–59.
- Karmarkar, S., Agrawal, V.K., Mathur, K.C. and Khadikar, P.V. (2002) QSAR studies on the toxicity of insecticides. *Bulg. Chem. Ind.*, **73**, 99–103.
- Karolak-Wojciechowska, J., Mrozek, A., Czyłkowski, R., Tekiner-Gulbas, B., Aki-Sener, E. and Yalçın, I. (2007) Five-membered heterocycles. Part IV. Impact of heteroatom on benzazole aromaticity. *J. Mol. Struct.*, **839**, 125–131.
- Kasai, K., Umeyama, H. and Tomonaga, A. (1988) The study of partition coefficients. The prediction of log *P* value based on molecular structure. *Bull. Chem. Soc. Jap.*, **61**, 2701–2706.
- Kastenholz, M.A., Pastor, M., Cruciani, G., Haaksma, E.E.J. and Fox, T. (2000) GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.*, **43**, 3033–3044.
- Kasum, D., Trinajstić, N. and Gutman, I. (1981) Chemical graph theory. III. On the permanent polynomial. *Croat. Chem. Acta*, **54**, 321–328.
- Kato, Y., Inoue, A., Yamada, M., Tomioka, N. and Itai, A. (1992) Automatic superposition of drug molecules based on their common receptor site. *J. Comput. Aid. Mol. Des.*, **6**, 475–486.
- Kato, Y., Itai, A. and Iitaka, Y. (1987) A novel method for superimposing molecules and receptor mapping. *Tetrahedron*, **43**, 5229–5236.
- Katona, G. and Diudea, M.V. (2003) Correlating ability of Cluj type indices. *Studia Univ. Babes-Bolyai*, **48**, 41–76.
- Katritzky, A.R. and Fara, D.C. (2005) How chemical structure determines physical, chemical, and technological properties: an overview illustrating the potential of quantitative structure–property relationships for fuels science. *Energy & Fuels*, **19**, 922–935.
- Katritzky, A.R., Fara, D.C., Kuanar, M., Hur, E. and Karelson, M. (2005) The classification of solvents by combining classical QSPR methodology with principal component analysis. *J. Phys. Chem. A*, **109**, 10323–10341.
- Katritzky, A.R. and Gordeeva, E.V. (1993) Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.*, **33**, 835–857.
- Katritzky, A.R., Ignatchenko, E.S., Barcock, R.A., Lobanov, V.S. and Karelson, M. (1994) Prediction of gas chromatographic retention times and response factors using a general quantitative structure–property relationship treatment. *Anal. Chem.*, **66**, 1799–1807.
- Katritzky, A.R., Jain, R., Lomaka, A., Petrukhin, R., Karelson, M., Visser, A.E. and Rogers, R.D. (2002) Correlation of the melting points of potential ionic liquids (imidazolium bromides and benzimidazolium bromides) using the CODESSA program. *J. Chem. Inf. Comput. Sci.*, **42**, 225–231.
- Katritzky, A.R., Jain, R., Lomaka, A., Petrukhin, R., Maran, U. and Karelson, M. (2001) Perspective on the relationship between melting points and chemical structure. *Cryst. Growth Des.*, **1**, 261–265.
- Katritzky, A.R., Kuanar, M., Fara, D.C., Karelson, M., Acree, W.E., Jr, Solov'ev, V.P. and Varnek, A. (2005) QSAR modeling of blood:air and tissue:air partition coefficients using theoretical descriptors. *Bioorg. Med. Chem.*, **13**, 6450–6463.

- Katritzky, A.R., Lobanov, V.S. and Karelson, M. (1995) QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.*, **24**, 279–287.
- Katritzky, A.R., Lobanov, V.S. and Karelson, M. (1998) Normal boiling points for organic compounds: correlation and prediction by a quantitative structure–property relationship. *J. Chem. Inf. Comput. Sci.*, **38**, 28–41.
- Katritzky, A.R., Lobanov, V.S., Karelson, M., Murugan, R., Grendze, M.P. and Toomey, J.E., Jr (1996) Comprehensive descriptors for structural and statistical analysis. 1. Correlations between structure and physical properties of substituted pyridines. *Rev. Roum. Chim.*, **41**, 851–867.
- Katritzky, A.R., Lomaka, A., Petrukhin, R., Jain, R., Karelson, M., Visser, A.E. and Rogers, R.D. (2002) QSPR correlation of the melting point for pyridinium bromides, potential ionic liquids. *J. Chem. Inf. Comput. Sci.*, **42**, 71–74.
- Katritzky, A.R., Maran, U., Karelson, M. and Lobanov, V.S. (1997) Prediction of melting points for the substituted benzenes: a QSPR approach. *J. Chem. Inf. Comput. Sci.*, **37**, 913–919.
- Katritzky, A.R., Maran, U., Lobanov, V.S. and Karelson, M. (2000) Structurally diverse quantitative structure–property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.*, **40**, 1–18.
- Katritzky, A.R., Mu, L. and Karelson, M. (1996a) A QSPR study of the solubility of gases and vapors in water. *J. Chem. Inf. Comput. Sci.*, **36**, 1162–1168.
- Katritzky, A.R., Mu, L. and Karelson, M. (1997) QSPR treatment of the unified nonspecific solvent polarity scale. *J. Chem. Inf. Comput. Sci.*, **37**, 756–761.
- Katritzky, A.R., Mu, L. and Karelson, M. (1998) Relationships of critical temperatures to calculated molecular properties. *J. Chem. Inf. Comput. Sci.*, **38**, 293–299.
- Katritzky, A.R., Mu, L., Lobanov, V.S. and Karelson, M. (1996b) Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.*, **100**, 10400–10407.
- Katritzky, A.R., Oliferenko, A.A., Oliferenko, P.A., Petrukhin, R., Tatham, D.B., Maran, U., Lomaka, A. and Acree, W.E., Jr (2003a) A general treatment of solubility. 1. The QSPR correlation of solvation free energies of single solutes in series of solvents. *J. Chem. Inf. Comput. Sci.*, **43**, 1794–1805.
- Katritzky, A.R., Oliferenko, A.A., Oliferenko, P.A., Petrukhin, R., Tatham, D.B., Maran, U., Lomaka, A. and Acree, W.E., Jr (2003b) A general treatment of solubility. 2. QSPR prediction of free energies of solvation of specified solutes in ranges of solvents. *J. Chem. Inf. Comput. Sci.*, **43**, 1806–1814.
- Katritzky, A.R., Pacureanu, L., Dobchev, D. and Karelson, M. (2007) QSPR study of critical micelle concentration of anionic surfactants using computational molecular descriptors. *J. Chem. Inf. Model.*, **47**, 782–793.
- Katritzky, A.R., Perumal, S. and Petrukhin, R. (2001a) A QSRR treatment of solvent effects on the decarboxylation of 6-nitrobenzisoxazole-3-carboxylates employing molecular descriptors. *J. Org. Chem.*, **66**, 4036–4040.
- Katritzky, A.R., Perumal, S., Petrukhin, R. and Kleinpeter, E. (2001b) CODESSA-based theoretical QSPR model for hydantoin HPLC-RT lipophilicities. *J. Chem. Inf. Comput. Sci.*, **41**, 569–574.
- Katritzky, A.R., Petrukhin, R., Jain, R. and Karelson, M. (2001a) QSPR analysis of flash points. *J. Chem. Inf. Comput. Sci.*, **41**, 1521–1530.
- Katritzky, A.R., Petrukhin, R., Perumal, S., Karelson, M., Prakash, I. and Desai, N. (2002) A QSPR study of sweetness potency using the CODESSA program. *Croat. Chem. Acta*, **75**, 475–502.
- Katritzky, A.R., Petrukhin, R., Tatham, D.B., Basak, S. C., Benfenati, E., Karelson, M. and Maran, U. (2001b) Interpretation of quantitative structure–property and –activity relationships. *J. Chem. Inf. Comput. Sci.*, **41**, 679–685.
- Katritzky, A.R., Sild, S. and Karelson, M. (1998a) Correlation and prediction of the refractive indices of polymers by QSPR. *J. Chem. Inf. Comput. Sci.*, **38**, 1171–1176.
- Katritzky, A.R., Sild, S. and Karelson, M. (1998b) General quantitative structure–property relationship treatment of the refractive index of organic compounds. *J. Chem. Inf. Comput. Sci.*, **38**, 840–844.
- Katritzky, A.R., Sild, S., Lobanov, V.S. and Karelson, M. (1998c) Quantitative structure–property relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers. *J. Chem. Inf. Comput. Sci.*, **38**, 300–304.
- Katritzky, A.R., Tamm, T., Wang, Y. and Karelson, M. (1999a) A unified treatment of solvent properties. *J. Chem. Inf. Comput. Sci.*, **39**, 692–698.
- Katritzky, A.R., Tamm, T., Wang, Y., Sild, S. and Karelson, M. (1999b) QSPR treatment of solvent scales. *J. Chem. Inf. Comput. Sci.*, **39**, 684–691.
- Katritzky, A.R. and Tatham, D.B. (2001) Correlation of the solubilities of gases and vapors in methanol and ethanol with their molecular structures. *J. Chem. Inf. Comput. Sci.*, **41**, 358–363.

- Katritzky, A.R., Tatham, D.B. and Maran, U. (2001) Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure–toxicity relationships. *J. Chem. Inf. Comput. Sci.*, **41**, 1162–1176.
- Katritzky, A.R., Wang, Y., Sild, S., Tamm, T. and Karelson, M. (1998) QSPR studies on vapor pressure, aqueous solubility, and prediction of water–air partition coefficients. *J. Chem. Inf. Comput. Sci.*, **38**, 720–725.
- Kauffman, G.W. and Jurs, P.C. (2000) Prediction of inhibition of the sodium ion–proton antiporter by benzoylguanidine derivatives from molecular structure. *J. Chem. Inf. Comput. Sci.*, **40**, 753–761.
- Kauffman, G.W. and Jurs, P.C. (2001a) Prediction of surface tension, viscosity, and thermal conductivity for common organic solvents using quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.*, **41**, 408–418.
- Kauffman, G.W. and Jurs, P.C. (2001b) QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J. Chem. Inf. Comput. Sci.*, **41**, 1553–1560.
- Kauvar, L.M., Higgins, D.L., Villar, H.O., Sportsman, J.R., Engqvist-Goldstein, Å., Bukar, R., Bauer, K.E., Dilley, H. and Rocke, D.M. (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.*, **2**, 107–118.
- Kawashima, Y., Sato, M., Yamamoto, S., Shimazaki, Y., Chiba, Y., Satake, M., Iwata, C. and Hatayama, K. (1995) Structure–activity relationship study of TXA₂ (2) receptor antagonists 4-(2-(4-substituted phenylsulfonylamino)ethylthio)phenoxyacetic acids and related compounds. *Chem. Pharm. Bull.*, **43**, 1132–1136.
- Kawashima, Y., Yamada, Y., Asaka, T., Misawa, Y., Kashimura, M., Morimoto, S., Ono, T., Nagate, T., Hatayama, K., Hirono, S. and Moriguchi, I. (1994) Structure–activity relationship study of 6-O-methylerythromycin-9-O-substituted oxime derivatives. *Chem. Pharm. Bull.*, **42**, 1088–1095.
- Kazius, J., McGuire, R. and Bursi, R. (2005) Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, **48**, 312–320.
- Kaznessis, Y.N., Snow, M.E. and Blankley, C.J. (2001) Prediction of blood–brain partitioning using Monte Carlo simulations of molecules in water. *J. Comput. Aid. Mol. Des.*, **15**, 697–708.
- Kearsley, S.K., Sallamack, S., Fluder, E.M., Andose, J. D., Mosley, R.T. and Sheridan, R.P. (1996) Chemical similarity using physico-chemical property descriptors. *J. Chem. Inf. Comput. Sci.*, **36**, 118–127.
- Kearsley, S.K. and Smith, G.M. (1990) An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.*, **3**, 615–633.
- Keinan, S. and Avnir, D. (1998) Quantitative chirality in structure–activity correlations. Shape recognition by trypsin, by the D2 dopamine receptor, and by cholinesterases. *J. Am. Chem. Soc.*, **120**, 6152–6159.
- Kekulé, A. (1865) Sur la constitution des substances aromatiques. *Bull. Soc. Chim. Fran. (French)*, **3**, 98–110.
- Kekulé, A. (1866a) Lehrbuch der organischen chemie erlangen, Germany.
- Kekulé, A. (1866b) Untersuchungen über aromatische Verbindungen. *Liebigs Ann. Chem.*, **137**, 129–136.
- Kelder, J., Grootenhuis, P.D.J., Bayada, D.M., Delbressinc, L.P.C. and Ploemen, J.P. (1999) Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.*, **16**, 1514–1519.
- Kelkar, M.A., Pednekar, D.V., Pimple, S.R. and Akamanchi, K.G. (2004) 3D QSAR studies of inhibitors of cholesterol ester transfer protein (CETP) by CoMFA. CoMSIA and GFA methodologies. *Med. Chem. Res.*, **13**, 590–604.
- Keller, T.H., Pichota, A. and Yin, Z. (2006) A practical view of ‘druggability’. *Drug Discov. Today*, **10**, 357–361.
- Kellogg, G.E. (1997) Finding optimum field models for 3-D CoMFA. *Med. Chem. Res.*, **7**, 417–427.
- Kellogg, G.E. and Abraham, D.J. (1992) KEY, LOCK, and LOCKSMITH: complementary hydrophobic map predictions of drug structure from a known receptor–receptor structure of known drugs. *J. Mol. Graph.*, **10**, 212–217.
- Kellogg, G.E. and Abraham, D.J. (2000) Hydrophobicity: is log $P_{o/w}$ more than the sum of its parts? *Eur. J. Med. Chem.*, **35**, 651–661.
- Kellogg, G.E., Joshi, G.S. and Abraham, D.J. (1992) New tools for modeling and understanding hydrophobicity and hydrophobic interactions. *Med. Chem. Res.*, **1**, 444–453.
- Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H. (1996) E-state fields: applications to 3D QSAR. *J. Comput. Aid. Mol. Des.*, **10**, 513–520.
- Kellogg, G.E., Sernus, S.F. and Abraham, D.J. (1991) HINT: a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aid. Mol. Des.*, **5**, 545–552.
- Kelly, D.P., Spillane, W.J. and Newell, J. (2005) Development of structure–taste relationships for monosubstituted phenylsulfamate sweeteners

- using classification and regression tree (CART) analysis. *J. Agr. Food Chem.*, **53**, 6750–6758.
- Kelvin, L. (1904) Baltimore lectures on molecular dynamics and the wave theory of light, in *Baltimore Lectures*, C.J. Clay and Sons, London, UK, pp. 618–619.
- Kemsley, E.K. (1998) A genetic algorithm (GA) approach to the calculation of canonical variates (CVs). *TRAC*, **17**, 24–34.
- Kerber, A., Laue, R., Meringer, M. and Rücker, C. (2004) Molecules *in silico*: the generation of structural formulae and its applications. *Journal of Combinatorial Chemistry Japan*, **3**, 85–96.
- Keseru, G. and Molnár, L. (2002) METAPRINT: a metabolic fingerprint. Application to cassette design for high-throughput ADME screening. *J. Chem. Inf. Comput. Sci.*, **42**, 437–444.
- Ketelaar, J.A.A. (1958) *Chemical Constitution. An Introduction to the Theory of the Chemical Bond*, Elsevier, Amsterdam, The Netherlands, p. 448.
- Kettaneh-Wold, N., MacGregor, J., Dayal, B. and Wold, S. (1994) Multivariate design of process experiments (M-DOPE). *Chemom. Intell. Lab. Syst.*, **23**, 39–50.
- Kezele, N., Klasinc, L., von Knop, J., Ivaniš, S. and Nikolić, S. (2002) Computing the variable vertex-connectivity index. *Croat. Chem. Acta*, **75**, 651–661.
- Khadikar, P.V., Agrawal, V.K. and Karmarkar, S. (2002) Prediction of lipophilicity of polyacenes using quantitative structure–activity relationships. *Bioorg. Med. Chem.*, **10**, 3499–3507.
- Khadikar, P.V., Clare, B.W., Balaban, A.T., Supuran, C.T., Agarwal, V.K., Singh, J., Joshi, A.K. and Lekhwani, M. (2007) QSAR modeling of carbonic anhydrase-I, -II and -IV inhibitory activities: relative correlation potential of six topological indices. *Rev. Roum. Chim.*, **51**, 703–717.
- Khadikar, P.V., Deshpande, N.V., Kale, P.P., Dobrynin, A.A., Gutman, I. and Dömötör, G. (1995) The Szeged index and an analogy with the Wiener index. *J. Chem. Inf. Comput. Sci.*, **35**, 547–550.
- Khadikar, P.V., Deshpande, N.V., Kale, P.P. and Gutman, I. (1994) Spectral moments of polyacenes. *J. Chem. Inf. Comput. Sci.*, **34**, 1181–1183.
- Khadikar, P.V., Diudea, M.V., Singh, J., John, P.E., Shrivastava, A., Singh, S., Karmarkar, S., Lakhwani, M. and Thakur, P. (2006) Use of PI index in computer-aided designing of bioactive compounds. *Curr. Bioact. Comp.*, **2**, 19–56.
- Khadikar, P.V., Joshi, S., Bajaj, A.V. and Mandloi, D. (2004) Correlations between the benzene character of acenes or helicenes and simple molecular descriptors. *Bioorg. Med. Chem. Lett.*, **14**, 1187–1191.
- Khadikar, P.V., Kale, P.P., Deshpande, N.V., Karmarkar, S. and Agrawal, V.K. (2001) Novel PI indices of hexagonal chains. *J. Math. Chem.*, **29**, 143–150.
- Khadikar, P.V. and Karmarkar, S. (2001) A novel PI index and its applications to QSPR/QSAR studies. *J. Chem. Inf. Comput. Sci.*, **41**, 934–949.
- Khadikar, P.V., Karmarkar, S., Singh, S. and Shrivastava, A. (2002) Use of the PI index in predicting toxicity of nitrobenzene derivatives. *Bioorg. Med. Chem.*, **10**, 3163–3170.
- Khadikar, P.V., Lukovits, I., Agrawal, V.K., Shrivastava, S., Jaiswal, M., Gutman, I., Karmarkar, S. and Shrivastava, A. (2003) Equalized electronegativity and topological indices: application for modeling toxicity of nitrobenzene derivatives. *Indian J. Chem.*, **42**, 1436–1441.
- Khadikar, P.V., Phadnis, A. and Shrivastava, A. (2002) QSAR study on toxicity to aqueous organisms using the PI index. *Bioorg. Med. Chem.*, **10**, 1181–1188.
- Khadikar, P.V., Sharma, V., Karmarkar, S. and Supuran, C.T. (2005a) Novel use of chemical shift in NMR as molecular descriptor: a first report on modeling carbonic anhydrase inhibitory activity and related parameters. *Bioorg. Med. Chem. Lett.*, **15**, 931–936.
- Khadikar, P.V., Sharma, V. and Varma, R.G. (2005b) Novel estimation of lipophilicity using ^{13}C NMR chemical shifts as molecular descriptor. *Bioorg. Med. Chem. Lett.*, **15**, 421–425.
- Khadikar, P.V., Singh, S., Mandloi, D., Joshi, S. and Bajaj, A.V. (2003) QSAR study on bioconcentration factor (BCF) of polyhalogenated biphenyls using the PI index. *Bioorg. Med. Chem.*, **11**, 5045–5050.
- Khadikar, P.V., Singh, S. and Shrivastava, A. (2002) Novel estimation of lipophilic behavior of polychlorinated biphenyl. *Bioorg. Med. Chem. Lett.*, **12**, 1125–1128.
- Khlebnikov, A.I., Akhmedzhanov, R.R., Naboka, O.I., Bakibaev, A.A., Tartynova, M.I., Novozheeva, T.P. and Saratikov, A.S. (2005) New cytochrome P-450 ligands based on urea derivatives. *Pharm. Chem. J.*, **39**, 18–21.
- Kiang, Y.-S. (1980) Determinant of the adjacency matrix and Kekule structures. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **14**, 541–547.
- Kiang, Y.-S. (2008) Calculation of the determinant of the adjacency matrix and the stability of conjugated molecules. *Int. J. Quant. Chem.*, **18**, 331–338.
- Kiang, Y.-S. and Tang, A. (1986) A graphical evaluation of characteristic polynomials of Hückel trees. *Int. J. Quant. Chem.*, **29**, 229–240.
- Kidera, A., Konisci, Y., Ooi, T. and Scheraga, H.A. (1985a) Relation between sequence similarity and

- structural similarity in proteins. Role of important properties of amino acids. *J. Prot. Chem.*, **4**, 265–297.
- Kidera, A., Konisci, Y., Ooi, T. and Scheraga, H.A. (1985b) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Prot. Chem.*, **4**, 23–55.
- Kier, L.B. (1971) *Molecular Orbital Theory in Drug Research*, Academic Press, New York.
- Kier, L.B. (1980a) Structural information from molecular connectivity $^4\chi_{pc}$ index. *J. Pharm. Sci.*, **69**, 1034–1039.
- Kier, L.B. (1980b) Use of molecular negentropy to encode structure governing biological activity. *J. Pharm. Sci.*, **69**, 807–810.
- Kier, L.B. (1985) A shape index from molecular graphs. *Quant. Struct. -Act. Relat.*, **4**, 109–116.
- Kier, L.B. (1986a) Distinguishing atom differences in a molecular graph shape index. *Quant. Struct. -Act. Relat.*, **5**, 7–12.
- Kier, L.B. (1986b) Indexes of molecular shape from chemical graphs. *Acta Pharm. Jugosl.*, **36**, 171–188.
- Kier, L.B. (1986c) Shape indexes of orders one and three from molecular graphs. *Quant. Struct. -Act. Relat.*, **5**, 1–7.
- Kier, L.B. (1987a) A structure based approach to molecular shape, in *QSAR in Drug Design and Toxicology* (eds D. Hadzi and B. Jerman-Blazic), Elsevier, Amsterdam, The Netherlands.
- Kier, L.B. (1987b) Inclusion of symmetry as a shape attribute in kappa index analysis. *Quant. Struct. -Act. Relat.*, **6**, 8–12.
- Kier, L.B. (1987c) Indexes of molecular shape from chemical graph. *Med. Res. Rev.*, **7**, 417–440.
- Kier, L.B. (1987d) The substituent steric effect index based on the molecular graph. *Quant. Struct. -Act. Relat.*, **6**, 117–122.
- Kier, L.B. (1989) An index of molecular flexibility from kappa shape attributes. *Quant. Struct. -Act. Relat.*, **8**, 221–224.
- Kier, L.B. (1990) Indexes of molecular shape from chemical graphs, in *Computational Chemical Graph Theory* (ed. D.H. Rouvray), Nova Science Publishers, New York, pp. 151–174.
- Kier, L.B. (1995) Atom-level descriptors for QSAR analyses, in *Chemometrics Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 39–47.
- Kier, L.B. (1997) Kappa shape indices for similarity analysis. *Med. Chem. Res.*, **7**, 394–406.
- Kier, L.B. (2006) My journey through structure: the structure of my journey. *Internet Electron. J. Mol. Des.*, **5**, 181–191.
- Kier, L.B., Cheng, C.-K. and Testa, B. (2003) Studies of ligand diffusion pathways over a protein surface. *J. Chem. Inf. Comput. Sci.*, **43**, 255–258.
- Kier, L.B., Di Paolo, T. and Hall, L.H. (1977) Structure–activity studies on odor molecules using molecular connectivity. *J. Theor. Biol.*, **67**, 585–595.
- Kier, L.B. and Glennon, R.A. (1978) Psychotomimetic phenalkylamines as serotonin antagonists: a SAR analysis. *Life Sci.*, **22**, 1589.
- Kier, L.B. and Hall, L.H. (1976a) *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, p. 257.
- Kier, L.B. and Hall, L.H. (1976b) Molecular connectivity. VII. Specific treatment of heteroatoms. *J. Pharm. Sci.*, **65**, 1806–1809.
- Kier, L.B. and Hall, L.H. (1977a) Structure–activity studies on hallucinogenic amphetamines using molecular connectivity. *J. Med. Chem.*, **20**, 1631–1636.
- Kier, L.B. and Hall, L.H. (1977b) The nature of structure–activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem.*, **12**, 307–312.
- Kier, L.B. and Hall, L.H. (1978) A molecular connectivity study of muscarinic receptor affinity of acetylcholine antagonists. *J. Pharm. Sci.*, **67**, 1408–1412.
- Kier, L.B. and Hall, L.H. (1979) Molecular connectivity analyses of structure influencing chromatographic retention indexes. *J. Pharm. Sci.*, **68**, 120–122.
- Kier, L.B. and Hall, L.H. (1981) Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.*, **70**, 583–589.
- Kier, L.B. and Hall, L.H. (1983a) Estimation of substituent group electronic influence from molecular connectivity delta values. *Quant. Struct. -Act. Relat.*, **2**, 163–167.
- Kier, L.B. and Hall, L.H. (1983b) General definition of valence delta-values for molecular connectivity. *J. Pharm. Sci.*, **72**, 1170–1173.
- Kier, L.B. and Hall, L.H. (1983c) Structural information and flexibility index from the molecular connectivity $^3\chi_p$ index. *Quant. Struct. -Act. Relat.*, **2**, 55–59.
- Kier, L.B. and Hall, L.H. (1986) *Molecular Connectivity in Structure–Activity Analysis*, Research Studies Press–John Wiley & Sons, Ltd, Chichester, UK, p. 262.
- Kier, L.B. and Hall, L.H. (1990a) An electrotopological-state index for atoms in molecules. *Pharm. Res.*, **7**, 801–807.
- Kier, L.B. and Hall, L.H. (1990b) The molecular connectivity of non-sigma electrons. *Rep. Mol. Theory*, **1**, 121–125.

- Kier, L.B. and Hall, L.H. (1991) A differential molecular connectivity index. *Quant. Struct. -Act. Relat.*, **10**, 134–140.
- Kier, L.B. and Hall, L.H. (1992a) An atom-centered index for drug QSAR models, in *Advances in Drug Design*, Vol. 22 (ed. B. Testa), Academic Press, New York.
- Kier, L.B. and Hall, L.H. (1992b) An index of atom electrotopological state, in *QSAR in Design of Bioactive compounds, A Telesymposium* (ed. A. Biaggi), Prous Science, Barcelona, Spain.
- Kier, L.B. and Hall, L.H. (1992c) Atom description in QSAR models: development and use of an atom level index. *Adv. Drug Res.*, **22**, 1–38.
- Kier, L.B. and Hall, L.H. (1993) The generation of molecular structures for a graph based QSAR equation. *Quant. Struct. -Act. Relat.*, **12**, 383–388.
- Kier, L.B. and Hall, L.H. (1995) A QSAR model of the OH radical reaction with CFCs. *SAR & QSAR Environ. Res.*, **3**, 97–100.
- Kier, L.B. and Hall, L.H. (1997a) Quantitative information analysis: the new center of gravity in medicinal chemistry. *Med. Chem. Res.*, **7**, 335–339.
- Kier, L.B. and Hall, L.H. (1997b) The E-state as an extended free valence. *J. Chem. Inf. Comput. Sci.*, **37**, 548–552.
- Kier, L.B. and Hall, L.H. (1999a) Molecular connectivity chi indices for database analysis and structure–property modeling, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 307–360.
- Kier, L.B. and Hall, L.H. (1999b) *Molecular Structure Description. The Electrotopological State*, Academic Press, London, UK, p. 246.
- Kier, L.B. and Hall, L.H. (1999c) The electrotopological state: structure modeling for QSAR and database analysis, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 491–562.
- Kier, L.B. and Hall, L.H. (1999d) The kappa indices for modeling molecular shape and flexibility, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 455–489.
- Kier, L.B. and Hall, L.H. (2000) Intermolecular accessibility: the meaning of molecular connectivity. *J. Chem. Inf. Comput. Sci.*, **40**, 792–795.
- Kier, L.B. and Hall, L.H. (2001a) Database organization and searching with e-state indices. *MATCH Commun. Math. Comput. Chem.*, **44**, 215–235.
- Kier, L.B. and Hall, L.H. (2001b) Molecular connectivity: intermolecular accessibility and encounter simulation. *J. Mol. Graph. Model.*, **20**, 76–83.
- Kier, L.B. and Hall, L.H. (2002) The meaning of molecular connectivity: a bimolecular accessibility model. *Croat. Chem. Acta*, **75**, 371–382.
- Kier, L.B., Hall, L.H. and Frazer, J.W. (1991) An index of electrotopological state for atoms in molecules. *J. Math. Chem.*, **7**, 229–241.
- Kier, L.B., Hall, L.H. and Frazer, J.W. (1993) Design of molecules from quantitative structure–activity relationship models. 1. Information transfer between path and vertex degree counts. *J. Chem. Inf. Comput. Sci.*, **33**, 143–147.
- Kier, L.B., Hall, L.H., Murray, W.J. and Randić, M. (1975) Molecular connectivity. I. Relationship to nonspecific local anesthesia. *J. Pharm. Sci.*, **64**, 1971–1974.
- Kier, L.B., Murray, W.J. and Hall, L.H. (1975) Molecular connectivity. 4. Relationships to biological activities. *J. Med. Chem.*, **18**, 1272–1274.
- Kier, L.B., Murray, W.J., Randić, M. and Hall, L.H. (1976a) Molecular connectivity concept applied to density. *J. Pharm. Sci.*, **65**, 1226–1230.
- Kier, L.B., Murray, W.J., Randić, M. and Hall, L.H. (1976b) Molecular connectivity. V. Connectivity series concept applied to density. *J. Pharm. Sci.*, **65**, 1226–1230.
- Kier, L.B., Simons, R.J. and Hall, L.H. (1978) Structure–activity studies on mutagenicity of nitrosoamines using molecular connectivity. *J. Pharm. Sci.*, **67**, 725–726.
- Kier, L.B. and Testa, B. (1995) Complexity and emergence in drug research. *Adv. Drug Res.*, **26**, 1–43.
- Kim, K.H. (1992) 3D quantitative structure–activity relationships description of electronic effects directly from 3D structures using a grid comparative molecular field analysis (CoMFA) approach. *Quant. Struct. -Act. Relat.*, **11**, 127–134.
- Kim, K.H. (1992) 3D quantitative structure–activity relationships investigation of steric effects with descriptors directly from 3D structures using a comparative molecular field analysis (CoMFA) approach. *Quant. Struct. -Act. Relat.*, **11**, 453–460.
- Kim, K.H. (1992c) 3D quantitative structure–activity relationships nonlinear dependence described directly from 3D structures using a comparative molecular field analysis (CoMFA) approach. *Quant. Struct. -Act. Relat.*, **11**, 309–317.
- Kim, K.H. (1993a) 3D quantitative structure–activity relationships: describing hydrophobic interactions

- directly from 3D structures using a comparative molecular field analysis (CoMFA) approach. *Quant. Struct. -Act. Relat.*, **12**, 232–238.
- Kim, K.H. (1993b) Comparison of classical and 3D QSAR, in *3D QSAR in Drug Design. Theory Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 619–642.
- Kim, K.H. (1993c) Nonlinear dependence in comparative molecular field analysis. *J. Comput. Aid. Mol. Des.*, **7**, 71–82.
- Kim, K.H. (1993d) Separation of electronic, hydrophobic, and steric effects in 3D quantitative structure–activity relationships with descriptors directly from 3D structures using a comparative molecular field analysis (CoMFA) approach. *Curr. Top. Med. Chem.*, **1**, 453–467.
- Kim, K.H. (1993e) Use of indicator variable in comparative molecular field analysis. *Med. Chem. Res.*, **3**, 257–267.
- Kim, K.H. (1993f) Use of the hydrogen-bond potential function in comparative molecular field analysis (CoMFA): an extension of CoMFA, in *Trends in QSAR and Molecular Modelling 92* (ed. C. G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 245–251.
- Kim, K.H. (1995a) *Comparative Molecular Field Analysis (CoMFA)* (ed. P.M. Dean), Chapman & Hall London, UK, pp. 291–331.
- Kim, K.H. (1995) Comparison of classical QSAR and comparative molecular field analysis: toward lateral validations, in *Classical and Three-Dimensional QSAR in Agrochemistry* (eds C. Hansch and T. Fujita), American Chemical Society, Washington, DC, pp. 302–317.
- Kim, K.H. (1995c) Description of the reversed-phase high-performance liquid chromatography (RP-HPLC) capacity factors and octanol–water partition coefficients of 2-pyrazine and 2-pyridine analogues directly from the three-dimensional structures using comparative molecular field analysis (CoMFA) approach. *Quant. Struct. -Act. Relat.*, **14**, 8–18.
- Kim, K.H. (1998) List of CoMFA references, 1993–1997, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 317–338.
- Kim, K.H. (2001) Thermodynamic aspects of hydrophobicity and biological QSAR. *J. Comput. Aid. Mol. Des.*, **15**, 367–380.
- Kim, K.H., Greco, G. and Novellino, E. (1998) A critical review of recent CoMFA applications, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 257–315.
- Kim, K.H., Greco, G., Novellino, E., Silipo, C. and Vittoria, A. (1993) Use of the hydrogen bond potential function in a comparative molecular field analysis (CoMFA) on a set of benzodiazepines. *J. Comput. Aid. Mol. Des.*, **7**, 263–280.
- Kim, K.H. and Kim, D.H. (1995) Description of hydrophobicity parameters of a mixed set from their three-dimensional structures. *Bioorg. Med. Chem.*, **3**, 1389–1396.
- Kim, K.H. and Martin, Y.C. (1991a) Direct prediction of dissociation constants (pK_a 's) of clonidine-like imidazolines, 2-substituted imidazoles, and 1-methyl-2-substituted-imidazoles from 3D structures using a comparative molecular field analysis (CoMFA) approach. *J. Med. Chem.*, **34**, 2056–2060.
- Kim, K.H. and Martin, Y.C. (1991b) Direct prediction of linear free energy substituent effects from 3D structures using comparative molecular field analysis. 1. Electronic effects of substituted benzoic acids. *J. Org. Chem.*, **56**, 2723–2729.
- Kim, K.H. and Martin, Y.C. (1991c) Evaluation of electrostatic and steric descriptors for 3D-QSAR: the H^+ and CH_3 probes using comparative molecular field analysis (CoMFA) and the modified partial least squares method, in *QSAR: Rational Approaches to the Design of Bioactive Compounds* (eds C. Silipo and A. Vittoria), Elsevier, Amsterdam, The Netherlands, pp. 151–154.
- Kim, Y.S., Kim, J.H., Kim, J.S., and No, K.T. (2002) Prediction of glass transition temperature (T_g) of some compounds in organic electroluminescent devices with their molecular properties. *J. Chem. Inf. Comput. Sci.*, **42**, 75–81.
- Kimura, T., Hasegawa, K. and Funatsu, K. (1998) GA strategy for variable selection in QSAR studies: GA-based region selection for CoMFA modeling. *J. Chem. Inf. Comput. Sci.*, **38**, 276–282.
- Kimura, T., Miyashita, Y., Funatsu, K. and Sasaki, S.I. (1996) Quantitative structure–activity relationships of the synthetic substrates for elastase enzyme using nonlinear partial least squares regression. *J. Chem. Inf. Comput. Sci.*, **36**, 185–189.
- King, J.W. (1989) A Z-weighted information content index. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **16**, 165–170.
- King, J.W. (1993) The inverse molecular transform index: a descriptor for molecular similarity analysis. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **20**, 139–145.
- King, J.W. (1994) Correlation of the partition coefficient with the molecular transform index in series of organophosphorus compounds. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **21**, 209–214.

- King, J.W. and Kassel, R.J. (1991) Dimensional response of the integrated molecular transform. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **18**, 289–297.
- King, J.W. and Kassel, R.J. (1992) Molecular transform quantization of enzyme surface probes. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **19**, 179–185.
- King, J.W., Kassel, R.J. and King, B.B. (1990) The integrated molecular transform as a correlation parameter. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **17**, 27–34.
- King, J.W. and Molnar, S.P. (1996) A unitary numerical descriptor of conformation. *J. Mol. Struct. (Theochem)*, **370**, 181–186.
- King, J.W. and Molnar, S.P. (1997) Correlation of organic diamagnetic susceptibility with structure via the integrated molecular transform. *Int. J. Quant. Chem.*, **64**, 635–645.
- King, J.W. and Molnar, S.P. (2000) Molecular structural index control in property-directed clustering and correlation. *Int. J. Quant. Chem.*, **80**, 1164–1171.
- King, R.B. (ed.) (1983) *Chemical Applications of Topology and Graph Theory*, Elsevier, Amsterdam, The Netherlands.
- King, R.B. (1991) Experimental tests of chirality algebra. *J. Math. Chem.*, **7**, 69–84.
- King, R.B. (2002) Riemann surfaces as descriptors for symmetrical negative curvature carbon and boron nitride structures. *Croat. Chem. Acta*, **75**, 447–473.
- King, R.D. and Srinivasan, A. (1996) Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environ. Health Persp.*, **104**, 1031–1040.
- King, R.D. and Srinivasan, A. (1997) The discovery of indicator variables for QSAR using inductive logic programming. *J. Comput. Aid. Mol. Des.*, **11**, 571–580.
- King, R.D., Srinivasan, A. and Dehaspe, L. (2001) Warmr: a data mining tool for chemical data. *J. Comput. Aid. Mol. Des.*, **15**, 173–181.
- Kiralj, R. and Ferreira, M.M.C. (2002) Predicting bond lengths in planar benzenoid polycyclic aromatic hydrocarbons: a chemometric approach. *J. Chem. Inf. Comput. Sci.*, **42**, 508–523.
- Kiralj, R. and Ferreira, M.M.C. (2003a) A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors I. The chemometric approach. *J. Mol. Graph. Model.*, **21**, 435–448.
- Kiralj, R. and Ferreira, M.M.C. (2003b) Molecular graphics-structural and molecular graphics descriptors in a QSAR study of 17- α -acetoxyprogesterones. *J. Braz. Chem. Soc.*, **14**, 20–26.
- Kiralj, R., Takahata, Y. and Ferreira, M.M.C. (2003) QSAR of progestogens: use of a priori and computed molecular descriptors and molecular graphics. *QSAR Comb. Sci.*, **22**, 430–448.
- Kirby, E.C. (1994) Sensitivity of topological indices to methyl group branching in octanes and azulenes, or what does a topological index? *J. Chem. Inf. Comput. Sci.*, **34**, 1030–1035.
- Kireev, D.B. (1995) ChemNet: a novel neural network based method for graph/property mapping. *J. Chem. Inf. Comput. Sci.*, **35**, 175–180.
- Kireev, D.B., Chrétien, J.R., Bernard, P. and Ros, F. (1998) Application of Kohonen neural networks in classification of biologically active compounds. *SAR & QSAR Environ. Res.*, **8**, 93–107.
- Kireev, D.B., Chrétien, J.R. and Raevsky, O.A. (1995) Molecular modeling and quantitative structure–activity studies of anti HIV-1 2-heteroarylquinoline-4-amines. *Eur. J. Med. Chem.*, **30**, 395–402.
- Kireev, D.B., Fetisov, V.I. and Zefirov, N.S. (1994) Approximate molecular electrostatic potential computations. Applications to quantitative structure–activity relationships. *J. Mol. Struct. (Theochem)*, **304**, 143–150.
- Kirkwood, J.J. and Westheimer, F.H. (1938) The electrostatic influence of substituents on the dissociation constants of organic acids. I. *J. Chim. Phys.*, **6**, 506–512.
- Kitaigorodsky, A.I. (1973) *Molecular Crystals and Molecules*, Academic Press, New York, 553 pp.
- Kitchen, D.B., Stahura, F.L. and Bajorath, J. (2004) Computational techniques for diversity analysis and compound classification. *Mini Rev. Med. Chem.*, **4**, 1029–1039.
- Klamt, A. (1996) Estimation of gas-phase hydroxyl radical rate constants of oxygenated compounds based on molecular orbital calculations. *Chemosphere*, **32**, 717–726.
- Klamt, A. and Eckert, F. (2001) COSMO-RS: a novel way from quantum chemistry to free energy, solubility, and general QSAR-descriptors for partitioning, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 195–205.
- Klamt, A., Eckert, F. and Hornig, M. (2001) COSMO-RS: a novel view to physiological solvation and partition questions. *J. Comput. Aid. Mol. Des.*, **15**, 355–365.
- Klamt, A., Jonas, V., Bürger, T. and Lohrenz, J.C.W. (1998) Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A*, **102**, 5074–5085.
- Klappa, S.A. and Long, G.R. (1992) Computer assisted determination of the biological activity of polychlorinated biphenyls using gas chromatographic retention indexes as molecular descriptors. *Anal. Chim. Acta*, **259**, 89–93.

- Klavžar, S. (2007) On the PI index: PI-partitions and Cartesian product graphs. *MATCH Commun. Math. Comput. Chem.*, **57**, 573–586.
- Klavžar, S. and Gutman, I. (1996) A comparison of the Schultz molecular topological index with the Wiener index. *J. Chem. Inf. Comput. Sci.*, **36**, 1001–1003.
- Klavžar, S. and Gutman, I. (2003) Relation between Wiener-type topological indices of benzenoid hydrocarbons. *Chem. Phys. Lett.*, **373**, 328–332.
- Klavžar, S., Gutman, I. and Rajapakse, A. (1997) Wiener numbers of pericondensed benzenoid hydrocarbons. *Croat. Chem. Acta*, **70**, 979–999.
- Klavžar, S., Rajapaxi, A. and Gutman, I. (1996) On the Szeged and the Wiener index of graphs. *Appl. Math. Lett.*, **67**, 45–49.
- Klavžar, S., Žigert, P. and Gutman, I. (2000) An algorithm for the calculation of the hyper-Wiener index of benzenoid hydrocarbons. *Computers Chem.*, **24**, 229–233.
- Klawun, C. and Wilkins, C.L. (1996a) Joint neural network interpretation of infrared and mass spectra. *J. Chem. Inf. Comput. Sci.*, **36**, 249–257.
- Klawun, C. and Wilkins, C.L. (1996b) Optimization of functional group prediction from infrared spectra using neural networks. *J. Chem. Inf. Comput. Sci.*, **36**, 69–81.
- Klebe, G. (1993) Structural alignment of molecules, in *3D QSAR in Drug Design. Theory Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 173–199.
- Klebe, G. (1998) Comparative molecular similarity indices analysis: CoMSIA, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 87–104.
- Klebe, G. and Abraham, U. (1993) On the prediction of binding properties of drug molecules by comparative molecular field analysis. *J. Med. Chem.*, **36**, 70–80.
- Klebe, G. and Abraham, U. (1999) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput. Aid. Mol. Des.*, **13**, 1–10.
- Klebe, G., Abraham, U. and Mietzner, T. (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.*, **37**, 4130–4146.
- Klebe, G., Mietzner, T. and Weber, F. (1994) Different approaches toward an automatic alignment of drug molecules: application to sterol mimics, thrombin and thermolysin inhibitors. *J. Comput. Aid. Mol. Des.*, **8**, 751–778.
- Klein, C.T., Kaiblinger, N. and Wolschann, P. (2002) Internally defined distances in 3D-quantitative structure–activity relationships. *J. Comput. Aid. Mol. Des.*, **16**, 79–93.
- Klein, C.T., Kaiser, D. and Ecker, G. (2004) Topological distance based 3D descriptors for use in QSAR and diversity analysis. *J. Chem. Inf. Comput. Sci.*, **44**, 200–209.
- Klein, C.T., Kaiser, D., Kopp, S., Chiba, P. and Ecker, G.F. (2002) Similarity based SAR (SIBAR) as tool for early ADME profiling. *J. Comput. Aid. Mol. Des.*, **16**, 785–793.
- Klein, D.J. (1986) Chemical graph-theoretic cluster expansion. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **69**, 701–712.
- Klein, D.J. (1995) Similarity and dissimilarity in posets. *J. Math. Chem.*, **18**, 321–348.
- Klein, D.J. (1997) Graph geometry, graph metrics, and Wiener. *MATCH Commun. Math. Comput. Chem.*, **35**, 7–27.
- Klein, D.J. (2002) Resistance–distance sum rules. *Croat. Chem. Acta*, **75**, 633–649.
- Klein, D.J. (2003a) Graph theoretically formulated electronic–structure theory. *Internet Electron. J. Mol. Des.*, **2**, 814–834.
- Klein, D.J. (2003b) Partitioning of Wiener-type indices, especially for trees. *Indian J. Chem.*, **42**, 1264–1269.
- Klein, D.J. and Babic, D. (1997) Partial orderings in chemistry. *J. Chem. Inf. Comput. Sci.*, **37**, 656–671.
- Klein, D.J. and Bytautas, L. (2000) Directed reaction graphs as posets. *MATCH Commun. Math. Comput. Chem.*, **42**, 261–290.
- Klein, D.J. and Gutman, I. (1999) Wiener-number-related sequences. *J. Chem. Inf. Comput. Sci.*, **39**, 534–536.
- Klein, D.J. and Ivanciuc, O. (2001) Graph cyclicity, excess conductance, and resistance deficit. *J. Math. Chem.*, **30**, 271–287.
- Klein, D.J., Lukovits, I. and Gutman, I. (1995) On the definition of the hyper-Wiener index for cycle-containing structures. *J. Chem. Inf. Comput. Sci.*, **35**, 50–52.
- Klein, D.J., Mihalić, Z., Plavšić, D. and Trinajstić, N. (1992) Molecular topological index: a relation with the Wiener index. *J. Chem. Inf. Comput. Sci.*, **32**, 304–305.
- Klein, D.J., Palacios, J.L., Randić, M. and Trinajstić, N. (2004) Random walks and chemical graph theory. *J. Chem. Inf. Comput. Sci.*, **44**, 1521–1525.
- Klein, D.J. and Randić, M. (1993) Resistance distance. *J. Math. Chem.*, **12**, 81–95.
- Klein, D.J., Randić, M., Babic, D., Lučić, B., Nikolić, S. and Trinajstić, N. (1997) Hierarchical

- orthogonalization of descriptors. *Int. J. Quant. Chem.*, **63**, 215–222.
- Klein, D.J., Schmalz, T.G. and Bytautas, L. (1999) Chemical sub-structural cluster expansions for molecular properties. *SAR & QSAR Environ. Res.*, **10**, 131–156.
- Klein, D.J. and Trinajstić, N. (1989) Foundations of conjugated-circuits models. *Prot. Struct. Funct. Gen.*, **61**, 2107–2115.
- Klein, D.J. and Zhu, H.Y. (1998) Distances and volumina for graphs. *J. Math. Chem.*, **23**, 179–195.
- Klir, G.J. and Folger, T.A. (1988) *Fuzzy Sets, Uncertainty, and Information*, Prentice-Hall, Englewood Cliffs, NJ, p. 356.
- Klocker, J., Wailzer, B., Buchbauer, G. and Wolschann, P. (2002a) Aroma quality differentiation of pyrazine derivatives using self-organizing molecular field analysis and artificial neural network. *J. Agr. Food Chem.*, **50**, 4069–4075.
- Klocker, J., Wailzer, B., Buchbauer, G. and Wolschann, P. (2002b) Bayesian neural networks for aroma classification. *J. Chem. Inf. Comput. Sci.*, **42**, 1443–1449.
- Klon, A.E. and Diller, D.J. (2007) Library fingerprints: a novel approach to the screening of virtual libraries. *J. Chem. Inf. Model.*, **47**, 1354–1365.
- Klon, A.E., Glick, M. and Davies, J.W. (2004) Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.*, **47**, 4356–4359.
- Klopman, G. (1984) Artificial intelligence approach to structure–activities studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.*, **106**, 7315–7321.
- Klopman, G. (1992) MULTICASE. 1. A hierarchical computer automated structure evaluation program. *Quant. Struct. -Act. Relat.*, **11**, 176–184.
- Klopman, G. (1998) The MultiCASE program. II. Baseline activity identification algorithm (BAIA). *J. Chem. Inf. Comput. Sci.*, **38**, 78–81.
- Klopman, G. and Buyukbingol, E. (1988) An artificial intelligence approach to the study of the structural moieties relevant to drug–receptor interactions in aldose reductase inhibitors. *Mol. Pharm.*, **34**, 852–862.
- Klopman, G. and Chakravarti, S.K. (2003) Structure–activity relationship study of a diverse set of estrogen receptor ligands (I) using MultiCASE expert system. *Chemosphere*, **51**, 445–459.
- Klopman, G. and Henderson, R.V. (1991) A graph theory-based “Expert System” methodology for structure–activity studies. *J. Math. Chem.*, **7**, 187–216.
- Klopman, G. and Iroff, L. (1981) Calculation of partition coefficients by the charge density method. *J. Comput. Chem.*, **2**, 157–160.
- Klopman, G., Li, J.Y., Wang, S. and Dimayuga, M. (1994) Computer automated log *P* calculations based on an extended group contribution approach. *J. Chem. Inf. Comput. Sci.*, **34**, 752–781.
- Klopman, G., Namboodiri, K. and Schochet, M. (1985) Simple method of computing the partition coefficient. *J. Comput. Chem.*, **6**, 28–38.
- Klopman, G. and Raychaudhury, C. (1988) A novel approach to the use of graph theory in structure–activity relationship studies. Application to the qualitative evaluation of mutagenicity in a series of nonfused ring aromatic compounds. *J. Comput. Chem.*, **9**, 232–243.
- Klopman, G. and Raychaudhury, C. (1990) Vertex indices of molecular graphs in structure–activity relationships: a study of the convulsant–anticonvulsant activity of barbiturates and the carcinogenicity of unsubstituted polycyclic aromatic hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **30**, 12–19.
- Klopman, G., Raychaudhury, C. and Henderson, R.V. (1988) A new approach to structure–activity using distance information content of graph vertices: a study with phenylalkylamines. *Math. Comput. Modelling*, **11**, 635–640.
- Klopman, G. and Rosenkranz, H.S. (1994) Approaches to SAR in carcinogenesis and mutagenesis prediction of carcinogenicity/mutagenicity using multi-case. *Mut. Res.*, **305**, 33–46.
- Klopman, G., Stefan, L.R. and Saiakhov, R.D. (2002) ADME evaluation: 2. A computer model for the prediction of intestinal absorption in humans. *Eur. J. Pharm. Sci.*, **17**, 253–263.
- Klopman, G. and Wang, S. (1991) A computer automated structure evaluation (CASE) approach to calculation of partition coefficient. *J. Comput. Chem.*, **12**, 1025–1032.
- Klopman, G., Wang, S., and Balthasar, D.M. (1992) Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.*, **32**, 474–482.
- Klopman, G. and Zhu, H. (2001) Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.*, **41**, 439–445.
- Klopman, G., Zhu, H., Ecker, G. and Chiba, P. (2003) MCASE study of the multidrug resistance reversal activity of propafenone analogs. *J. Comput. Aid. Mol. Des.*, **17**, 291–297.

- Knop, J.V. and Trinajstić, N. (1980) Chemical graph theory. II. On the graph theoretical polynomials of conjugated structures. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **14**, 503–520.
- Knotts, T.A., Wilding, W.V., Oscarson, J.L. and Rowley, R.L. (2001) Use of the DIPPR database for development of QSPR correlations: surface tension. *J. Chem. Eng. Data*, **46**, 1007–1012.
- Kobayashi, S., Shinohara, H., Tabata, K., Yamamoto, N. and Miyai, A. (2006) Stereo structure-controlled and electronic structure-controlled estrogen-like chemicals to design and develop non-estrogenic bisphenol A analogs based on chemical hardness concept. *Chem. Pharm. Bull.*, **54**, 1633–1638.
- Koch, R. (1982) Molecular connectivity and acute toxicity of environmental pollutants. *Chemosphere*, **11**, 925–931.
- Koch, W. and Holthausen, M.C. (2001) *A Chemist's Guide to Density Functional Theory*. Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 300.
- Koehler, M.G., Grigoras, S. and Dunn, W.J. III (1988) The relationship between chemical structure and the logarithm of the partition coefficient. *Quant. Struct. -Act. Relat.*, **7**, 150–159.
- Kohonen, T. (1989) *Self-Organization and Associative Memory*. Springer, Berlin, Germany, p. 312.
- Kohonen, T. (1990) The self-organizing map. *Proc. IEEE*, **78**, 1464–1480.
- Kompany-Zareh, M. (2003) A QSPR study of boiling point of saturated alcohols using genetic algorithm. *Acta Chim. Sloven.*, **50**, 259–273.
- Konovalov, D.A., Coomans, D., Deconinck, E. and Vander Heyden, Y. (2007) Benchmarking of QSAR models for blood–brain barrier permeation. *J. Chem. Inf. Model.*, **47**, 1648–1656.
- Konstantinova, E.V. (1996) The discrimination ability of some topological and information distance indices for graphs of unbranched hexagonal systems. *J. Chem. Inf. Comput. Sci.*, **36**, 54–57.
- Konstantinova, E.V. (2006) On some applications of information indices in chemical graph theory, in *General Theory of Information Transfer and Combinatorics* (eds R. Ahlswede, L. Baumer, N. Cai, et al.), Springer-Verlag, Berlin Heidelberg, Germany, pp. 831–852.
- Konstantinova, E.V. and Diudea, M.V. (2000) The Wiener polynomial derivatives and other topological indices in chemical research. *Croat. Chem. Acta*, **73**, 383–403.
- Konstantinova, E.V. and Paleev, A.A. (1990) Sensitivity of topological indices of polycyclic graphs. *Vychisl. Sistemy (Russian)*, **136**, 38–48.
- Konstantinova, E.V. and Skorobogatov, V.A. (1995) Molecular hypergraphs: the new representation of nonclassical molecular structures with polycentric delocalized bonds. *J. Chem. Inf. Comput. Sci.*, **35**, 472–478.
- Konstantinova, E.V., Skorobogatov, V.A. and Vidyuk, M.V. (2003) Application of information theory in chemical graph theory. *Indian J. Chem.*, **42**, 1227–1240.
- Konstantinova, E.V. and Vidyuk, M.V. (2003) Discriminating tests of information and topological indices. Animals and trees. *J. Chem. Inf. Comput. Sci.*, **43**, 1860–1871.
- Kopecky, J., Bocek, K. and Vlachová, D. (1965) Chemical structure and biological activity on *m*- and *p*-disubstituted derivatives of benzene. *Nature*, **207**, 981.
- Koppel, I.A. and Paju, A.I. (1974) *React. Sposobnost. Org. Soedin.*, **11**, 137–140.
- Körner, W. (1869) Fatti per servire alla determinazione del luogo chimico nelle sostanze aromatiche. *Giornale di Scienze Naturali ed Economiche*, **5**, 212–256.
- Körner, W. (1874) Studi sulla Isomeria delle Così Dette Sostanze Aromatiche a Sei Atomi di Carbonio. *Gazz. Chim. It.*, **4**, 242.
- Korolev, D., Balakin, K.V., Nikolsky, Y., Kirillov, E., Ivanenkov, Y.A., Savchuk, N.P., Ivashchenko, A.A. and Nokolskaya, T. (2003) Modeling of human cytochrome P450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.*, **46**, 3631–3643.
- Kosower, E.M. (1958a) The effect of solvent on spectra. I. A new empirical measure of solvent polarity: Z-values. *J. Am. Chem. Soc.*, **80**, 3253–3260.
- Kosower, E.M. (1958b) The effect of solvent on spectra. II. Correlation of spectral absorption data with Z-values. *J. Am. Chem. Soc.*, **80**, 3261–3267.
- Kotani, T. and Higashiura, K. (2002) Rapid evaluation of molecular shape similarity index using pairwise calculation of the nearest atomic distances. *J. Chem. Inf. Comput. Sci.*, **42**, 58–63.
- Kotnik, M., Oblak, M., Humljan, J., Gobec, S., Urleb, U. and Solmajer, T. (2004) Quantitative structure–activity relationships of *Streptococcus pneumoniae* MurD transition state analogue inhibitors. *QSAR Comb. Sci.*, **23**, 399–405.
- Kourounakis, A. and Bodor, N. (1995) Quantitative structure–activity relationships of catechol derivatives on nerve growth factor secretion in L-M cells. *Pharm. Res.*, **12**, 1199–1204.
- Kovalishyn, V.V., Tetko, I.V., Luik, A.I., Artemenko, A. G. and Kuz'min, V.E. (2001) A new algorithm for spatial learning of artificial neural networks based on lattice models of chemical structures for QSAR analysis. *Pharm. Chem. J.*, **35**, 78–84.

- Kovalishyn, V.V., Tetko, I.V., Luik, A.I., Kholodovych, V.V., Villa, A.E.P. and Livingstone, D.J. (1998) Neural network studies. 3. Variable selection in the cascade-correlation learning architecture. *J. Chem. Inf. Comput. Sci.*, **38**, 651–659.
- Kovatcheva, A., Buchbauer, G., Golbraikh, A. and Wolschann, P. (2003) QSAR modeling of α -campholenic derivatives with sandalwood odor. *J. Chem. Inf. Comput. Sci.*, **43**, 259–266.
- Kovatcheva, A., Golbraikh, A., Oloff, S., Feng, J., Zheng, W. and Tropsha, A. (2005) QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. *SAR & QSAR Environ. Res.*, **16**, 93–102.
- Kovatcheva, A., Golbraikh, A., Oloff, S., Xiao, Y.-D., Zheng, W., Wolschann, P., Buchbauer, G. and Tropsha, A. (2004) Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.*, **44**, 582–595.
- Kováts, E. (1968) Zu Fragen der Polarität. *Chimia*, **22**, 459.
- KOWWIN, Syracuse Research Corporation, North Syracuse, NY.
- Kraak, M.H.S., Wijnands, P., Govers, H.A.J., Admiraal, W.A. and de Voogt, P. (1997) Structural-based differences in ecotoxicity of benzoquinoline isomers to the zebra mussel (*Dreissena polymorpha*). *Environ. Toxicol. Chem.*, **16**, 2158–2163.
- Kraker, J.J., Hawkins, D.M., Basak, S.C., Natarajan, R. and Mills, D. (2007) Quantitative structure–activity relationship (QSAR) modeling of juvenile hormone activity: comparison of validation procedures. *Chemom. Intell. Lab. Syst.*, **87**, 33–42.
- Kränz, H., Vill, V. and Meyer, B. (1996) Prediction of material properties from chemical structures. The clearing temperature of nematic liquid crystal derived from their chemical structures by artificial neural networks. *J. Chem. Inf. Comput. Sci.*, **36**, 1173–1177.
- Krawczuk, A., Voelkel, A., Lulek, J., Urbaniak, R. and Szyrwinska, K. (2003) Use of topological indices of polychlorinated biphenyls in structure–retention relationships. *J. Chromat.*, **1018**, 63–71.
- Krenkel, G., Castro, E.A. and Toropov, A.A. (2001a) Improved molecular descriptors based on the optimization of correlation weights of local graph invariants. *Int. J. Mol. Sci.*, **2**, 57–65.
- Krenkel, G., Castro, E.A. and Toropov, A.A. (2001b) Improved molecular descriptors to calculate boiling points based on the optimization of correlation weights of local graph invariants. *J. Mol. Struct. (Theochem)*, **542**, 107–113.
- Krenkel, G., Castro, E.A. and Toropov, A.A. (2002) 3D and 4D molecular models derived from the ideal symmetry method: prediction of alkanes normal boiling points. *Chem. Phys. Lett.*, **355**, 517–528.
- Krieger, A.M. and Zhang, P. (2006) Generalized final prediction error criteria, in *Encyclopedia of Statistical Sciences* (eds S. Kotz, C.B. Read, N. Balakrishnan and B. Vidaković), John Wiley & Sons, Inc., New York, pp. 1–4.
- Kriegel, J.M., Arnhold, T., Beck, B. and Fox, T. (2005a) A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. *J. Comput. Aid. Mol. Des.*, **19**, 189–201.
- Kriegel, J.M., Arnhold, T., Beck, B. and Fox, T. (2005b) Prediction of human cytochrome P450 inhibition using support vector machines. *QSAR Comb. Sci.*, **24**, 491–502.
- Krivka, P., Jericevic, Z. and Trinajstić, N. (1985) On the computation of characteristic polynomial of a chemical graph. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **19**, 129–147.
- Kroemer, R.T., Ettmayer, P. and Hecht, P. (1995) 3D-quantitative structure–activity relationships of human immunodeficiency virus type-1 proteinase inhibitors: comparative molecular field analysis of 2-heterosubstituted statine derivatives – implications for the design of novel inhibitors. *J. Med. Chem.*, **38**, 4917–4928.
- Kroemer, R.T., Hecht, P. and Liedl, K.R. (1996) Different electrostatic descriptors in comparative molecular field analysis: a comparison of molecular electrostatic and Coulomb potentials. *J. Comput. Chem.*, **17**, 1296–1308.
- Kruja, E., Marks, J., Blair, A. and Waters, R. (2002) A short note on the history of graph drawing, in *Proceedings of International Symposium on Graph Drawing*, Springer-Verlag, Berlin, Germany, pp. 602–606.
- Kruszewski, J. and Krygowski, T.M. (1972) Definition of aromaticity basing on the harmonic oscillator model. *Tetrahedron Lett.*, **36**, 3839–3842.
- Krygowski, T.M. (1993) Crystallographic studies of inter- and intramolecular interactions reflected in aromatic character of π -electron systems. *J. Chem. Inf. Comput. Sci.*, **33**, 70–78.
- Krygowski, T.M., Anulewicz, R. and Kruszewski, J. (1983) Crystallographic studies and physico-chemical properties of π -electron compounds. III. Stabilization energy and the Kekulé structure contributions derived from experimental bond lengths. *Acta Cryst.*, **39**, 732–739.
- Krygowski, T.M., Anulewicz, R. and Wisiorowski, M. (1995) Derivation of the Kekulé structure contributions from experimental bond lengths for π -electron systems with NN and NO bonds.

- Extension of the HOSE model. *Pol. J. Chem.*, **69**, 1579–1584.
- Krygowski, T.M. and Ciesielski, A. (1995) Local aromatic character of C_{60} and C_{70} and their derivatives. *J. Chem. Inf. Comput. Sci.*, **35**, 1001–1003.
- Krygowski, T.M., Ciesielski, A., Bird, C.W. and Kotschy, A. (1995) Aromatic character of the benzene ring present in various topological environments in benzenoid hydrocarbons. Nonequivalence of indices of aromaticity. *J. Chem. Inf. Comput. Sci.*, **35**, 203–210.
- Krygowski, T.M. and Cyranski, M. (1996a) Separation of the energetic and geometric contributions to the aromaticity of π -electron carbocyclics. *Tetrahedron*, **52**, 1713–1722.
- Krygowski, T.M. and Cyranski, M. (1996b) Separation of the energetic and geometric contributions to the aromaticity. Part IV. A general model for the π -electron systems. *Tetrahedron*, **52**, 10255–10264.
- Krygowski, T.M., Cyranski, M., Ciesielski, A., Swirska, B. and Leszczynski, P. (1996) Separation of the energetic and geometric contributions to aromaticity. 2. Analysis of the aromatic character of benzene rings in their various topological environments in the benzenoid hydrocarbons. Crystal and molecular structure of coronene. *J. Chem. Inf. Comput. Sci.*, **36**, 1135–1141.
- Krygowski, T.M. and Cyranski, M.K. (2001) Structural aspects of aromaticity. *Chem. Rev.*, **101**, 1385–1419.
- Krygowski, T.M., Cyranski, M.K., Czarnocki, Z., Häfelinger, G. and Katritzky, A.R. (2000) Aromaticity: a theoretical concept of immense practical importance. *Tetrahedron*, **56**, 1783–1796.
- Krygowski, T.M., Ejsmont, K., Stepień, B.T., Cyranski, M.K., Poater, J. and Solà, M. (2004) Relation between the substituent effect and aromaticity. *J. Org. Chem.*, **69**, 6634–6640.
- Krygowski, T.M. and Stepień, B.T. (2005) Sigma- and pi-electron delocalization: focus on substituent effects. *Chem. Rev.*, **105**, 3482–3512.
- Krygowski, T.M. and Wieckowski, T. (1981) Analysis of the hydrogen-bridge in carboxylic acids in terms of stabilization energy derived from bond lengths. Non-Hammett properties of *p*-substituted benzoic acids in the crystalline state. *Croat. Chem. Acta*, **54**, 193–202.
- Krygowski, T.M., Wrona, P.K., Zielkowska, U. and Reichardt, C. (1985) Empirical parameters of the Lewis acidity and basicity for aqueous binary solvent mixtures. *Tetrahedron*, **41**, 4519–4527.
- Krzanowski, W.J. (1988) *Principles of Multivariate Analysis*, Oxford University Press, New York, p. 564.
- Krzyzaniak, J.F., Myrdal, P.B., Simamora, P. and Yalkowsky, S.H. (1995) Boiling point and melting point prediction for aliphatic, non-hydrogen-bonding compounds. *Ind. Eng. Chem. Res.*, **34**, 2530–2535.
- Kuanar, M., Kuanar, S.K. and Mishra, B.K. (1999a) Correlation of critical micelle concentration of nonionic surfactants with molecular descriptors. *Indian J. Chem.*, **38**, 113–118.
- Kuanar, M., Kuanar, S.K., Mishra, B.K. and Gutman, I. (1999b) Correlation of line graph parameters with physico-chemical properties of octane isomers. *Indian J. Chem.*, **38**, 525–528.
- Kubinyi, H. (1976) Quantitative structure–activity relationships. 2. A mixed approach. based on Hansch and Free–Wilson analysis. *J. Med. Chem.*, **19**, 587–600.
- Kubinyi, H. (1976b) Quantitative structure–activity relationships. IV. Non-linear dependence of biological activity on hydrophobic character: a new model. *Arzneim. Forsch. (German)*, **26**, 1991–1997.
- Kubinyi, H. (1977) Quantitative structure–activity relationships. 7. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J. Med. Chem.*, **20**, 625–629.
- Kubinyi, H. (1979) Lipophilicity and biological activity. Drug transport and drug distribution in model systems and biological systems. *Arzneim. Forsch. (German)*, **29**, 1067–1080.
- Kubinyi, H. (1988a) Current problems in quantitative structure–activity relationships, in *Physical Property Prediction in Organic Chemistry* (eds C. Jochum, M.G. Hicks and J. Sunkel), Springer-Verlag, Berlin, Germany, pp. 235–247.
- Kubinyi, H. (1988b) Free–Wilson analysis. Theory, applications and its relationship to Hansch analysis. *Quant. Struct. -Act. Relat.*, **7**, 121–133.
- Kubinyi, H. (1990) The Free–Wilson method and its relationship to the extrathermodynamic approach, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 589–643.
- Kubinyi, H. (ed.) (1993a) *3D QSAR in Drug Design. Theory, Methods, and Applications*, ESCOM, Leiden, The Netherlands, p. 760.
- Kubinyi, H. (1993b) *QSAR: Hansch Analysis and Related Approaches*, VCH Publishers, Weinheim, Germany, p. 240.
- Kubinyi, H. (1994a) Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct. -Act. Relat.*, **13**, 285–294.
- Kubinyi, H. (1994b) Variable selection in QSAR studies. II. A highly efficient combination of

- systematic search and evolution. *Quant. Struct. - Act. Relat.*, **13**, 393–401.
- Kubinyi, H. (1995) From lipophilicity to 3D QSAR – the fascination of computer-aided drug design, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, Barcelona, Spain, pp. 2–16.
- Kubinyi, H. (1996) Evolutionary variable selection in regression and PLS analyses. *J. Chemom.*, **10**, 119–133.
- Kubinyi, H. (1997) A general view on similarity and QSAR studies, in *Computer-Assisted Lead Finding and Optimization* (eds H. van de Waterbeemd, B. Testa and G. Folkers), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 9–28.
- Kubinyi, H. (1998) Similarity and dissimilarity: a medicinal chemist's view, in *3D QSAR in Drug Design*, Vol. 2 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 226–252.
- Kubinyi, H. (2002) From narcosis to hyperspace: the history of QSAR. *Quant. Struct. -Act. Relat.*, **21**, 348–356.
- Kubinyi, H. (2003a) Comparative molecular field analysis (CoMFA), in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1555–1575.
- Kubinyi, H. (2003b) QSAR in drug design, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1532–1554.
- Kubinyi, H., Folkers, G. and Martin, Y.C. (eds) (1998a) *3D QSAR in Drug Design*, Vol. 2, Kluwer/ESCOM, Dordrecht, The Netherlands, p. 416.
- Kubinyi, H., Folkers, G. and Martin, Y.C. (eds) (1998b) *3D QSAR in Drug Design*, Vol. 3, Kluwer/ESCOM, Dordrecht, The Netherlands, p. 352.
- Kubinyi, H., Hamprecht, F.A. and Mietzner, T. (1998) Three-dimensional quantitative similarity–activity relationships (3D-QSiAR) from SEAL similarity matrices. *J. Med. Chem.*, **41**, 2553–2564.
- Kubinyi, H. and Kehrhahn, O.M. (1976) Quantitative structure–activity relationships. 1. The modified Free–Wilson approach. *J. Med. Chem.*, **19**, 578–586.
- Kulkarni, A.S. (1999) Prediction of eye irritation from organic chemicals using membrane-interaction QSAR analysis. *Toxicol. Sci.*, **59**, 335–345.
- Kulkarni, A.S., Han, Y. and Hopfinger, A.J. (2002) Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *J. Chem. Inf. Comput. Sci.*, **42**, 331–342.
- Kulkarni, A.S. and Hopfinger, A.J. (1999) Membrane-interaction QSAR analysis: application to the estimation of eye irritation by organic compounds. *Pharm. Res.*, **16**, 1244–1252.
- Kulkarni, S.K., Newman, A.H. and Houlihan, W.J. (2002) Three-dimensional quantitative structure–activity relationships of mazindol analogues at the dopamine transporter. *J. Med. Chem.*, **45**, 4119–4127.
- Kumar, V. and Madan, A.K. (2004) Topological models for the prediction of cyclin-dependent kinase 2. Inhibitory activity of aminothiazoles. *MATCH Commun. Math. Comput. Chem.*, **51**, 59–78.
- Kumar, V. and Madan, A.K. (2006) Application of graph theory: prediction of cytosolic phospholipase A2 inhibitory activity of propan-2-ones. *J. Math. Chem.*, **39**, 511–521.
- Kumar, V., Sardana, S. and Madan, A.K. (2004) Predicting anti-HIV activity of 2,3-diaryl-1,3-thiazolidin-4-ones: computational approach using reformed eccentric connectivity index. *J. Mol. Model.*, **10**, 399–407.
- Kunz, M. (1986) Entropy and information indices of star forests. *Collect. Czech. Chem. Comm.*, **51**, 1856–1863.
- Kunz, M. (1989) Path and walk matrices of trees. *Collect. Czech. Chem. Comm.*, **54**, 2148–2155.
- Kunz, M. (1990) Molecular connectivity indices revisited. *Collect. Czech. Chem. Comm.*, **55**, 630–633.
- Kunz, M. (1993) On topological and geometrical distance matrices. *J. Math. Chem.*, **13**, 145–151.
- Kunz, M. (1994) Distance matrices yielding angles between arcs of the graphs. *J. Chem. Inf. Comput. Sci.*, **34**, 957–959.
- Kunz, M. and Radl, Z. (1998) Distributions of distances in information strings. *J. Chem. Inf. Comput. Sci.*, **38**, 374–378.
- Kupchik, E.J. (1985) Structure–molar refraction relationships of alkylsilanes using molecular connectivity. *Quant. Struct. -Act. Relat.*, **4**, 123–127.
- Kupchik, E.J. (1986) Structure–molar refraction relationships of alkylsilanes using empirically-modified first order molecular connectivity indices. *Quant. Struct. -Act. Relat.*, **5**, 95–98.
- Kupchik, E.J. (1988) Structure–molar refraction relationships of alkylgermanes using molecular connectivity. *Quant. Struct. -Act. Relat.*, **7**, 57–59.
- Kupchik, E.J. (1989) General treatment of heteroatoms with the Randić molecular connectivity index. *Quant. Struct. -Act. Relat.*, **8**, 98–103.
- Kurnakov, N.S. (1928) *Z. Anorg. Allg. Chem. (German)*, **169**, 113–139.
- Kurunczi, L., Olah, M., Oprea, T.I., Bologa, C. and Simon, Z. (2002) MTD-PLS: a PLS-based variant of

- the MTD method. 2. Mapping ligand–receptor interactions. Enzymatic acetic acid esters hydrolysis. *J. Chem. Inf. Comput. Sci.*, **42**, 841–846.
- Kutter, E. and Hansch, C. (1969) Steric parameters in drug design. Monoamine oxidase inhibitors and antihistamines. *J. Med. Chem.*, **12**, 647–652.
- Kutulya, L.A., Kuz'min, V.E., Stel'makh, I.B., Handrimailova, T.V. and Shtifanyuk, P.P. (1992) Quantitative aspects of chirality. III. Description of the influence of the structure of chiral compounds on their twisting power in the nematic mesophase by means of the dissymmetry function. *J. Phys. Org. Chem.*, **5**, 308–316.
- Kuz'min, V.E., Artemenko, A.G., Polischuk, P.G., Muratov, E.N., Hromov, A.I., Liahovskiy, A.V., Andronati, S.A. and Makan, S.Y. (2005) Hierarchic system of QSAR models (1D–4D) on the base of simplex representation of molecular structure. *J. Mol. Model.*, **11**, 457–467.
- Kuz'min, V.E., Novikova, N.S., Sidelnikova, T.A. and Trigub, L.P. (1994) Topological analysis of the structure–mesomorphous property relationship. *J. Struct. Chem.*, **35**, 471–477.
- Kuz'min, V.E., Stel'makh, I.B., Bekker, M.B. and Pozigun, D.V. (1992a) Quantitative aspects of chirality. I. Method of dissymmetry function. *J. Phys. Org. Chem.*, **5**, 295–298.
- Kuz'min, V.E., Stel'makh, I.B., Yudanov, I.V., Pozigun, D.V. and Bekker, M.B. (1992b) Quantitative aspects of chirality. II. Analysis of dissymmetry function behaviour with different changes in the structure of the model systems. *J. Phys. Org. Chem.*, **5**, 299–307.
- Kuz'min, V.E., Trigub, L.P., Shapiro, Y.E., Mazurov, A. A., Pozigun, V.V., Gorbatyuk, V.Y. and Andronati, S.A. (1995) Shape parameters of peptide molecules as descriptors for solving QSAR problems. *J. Struct. Chem. Eng. Transl.*, **36**, 465–473.
- Kvasnička, V. and Pospichal, J. (1990) Canonical indexing and constructive enumeration of molecular graphs. *J. Chem. Inf. Comput. Sci.*, **30**, 99–105.
- Kvasnička, V. and Pospichal, J. (1995) Simple construction of embedding frequencies of trees and rooted trees. *J. Chem. Inf. Comput. Sci.*, **35**, 121–128.
- Kvasnička, V., Sklenák, Š. and Pospichal, J. (1993a) Application of high order neural networks in chemistry. *Theor. Chim. Acta*, **86**, 257–267.
- Kvasnička, V., Sklenák, Š. and Pospichal, J. (1993b) Neural network classification of inductive and resonance effects of substituents. *J. Am. Chem. Soc.*, **115**, 1495–1500.
- Kyngas, J. and Valjakka, J. (1996) Evolutionary neural networks in quantitative structure–activity relationships of dihydrofolate reductase inhibitors. *Quant. Struct. -Act. Relat.*, **15**, 296–301.
- L'Heureux, P.-J., Carreau, J., Bengio, Y., Delalleau, O. and Yue, S.Y. (2004) Locally linear embedding for dimensionality reduction in QSAR. *J. Comput. Aid. Mol. Des.*, **18**, 475–482.
- L'Huillier, S.A.J. (1861) Mémoire sur la polyèdrométrie. *Annales de Mathématiques*, **3**, 169–189.
- Labanowski, J.K., Motoc, I. and Dammkoehler, R.A. (1991) The physical meaning of topological indices. *Computers Chem.*, **15**, 47–53.
- Labute, P. (1999) Binary QSAR: a new method for the determination of quantitative structure–activity relationships, in *Pacific Symposium on Biocomputing 1999*, Vol. 7 (eds R.B. Altman, A.K. Dunker, L. Hunter, T.E. Klein and K. Lauderdale), World Scientific, NJ, pp. 444–455.
- Labute, P. (2000) A widely applicable set of descriptors. *J. Mol. Graph. Model.*, **18**, 464–477.
- Labute, P. (2001) Probabilistic receptor potential, Internet communication, <http://www.chemcomp.com/journal/cstat.htm>.
- Laffort, P. and Patte, F. (1976) Solubility factors in gas–liquid chromatography: comparison between two approaches and application to some biological studies. *J. Chromat.*, **126**, 625–639.
- Laidboeur, T., Cabrol-Bass, D. and Ivanciuc, O. (1997) Determination of topo-geometrical equivalence classes of atoms. *J. Chem. Inf. Comput. Sci.*, **37**, 87–91.
- Lajiness, M.S. (1990) Molecular similarity-based methods for selecting compounds for screening, in *Computational Chemical Graph Theory* (ed. D.E. Rouvray), Nova Science Publishers, New York, pp. 299–316.
- Lall, R.S. (1981a) Topology and physical properties of acyclic compounds. *Curr. Sci. -India*, **50**, 846–849.
- Lall, R.S. (1981b) Topology and physical properties of *n*-alkanes. *Curr. Sci. -India*, **50**, 668–670.
- Lall, R.S. (1981c) Topology of chemical reactions. I. The fragmentation of hydrocarbons. II. Pericyclic reactions. *MATCH Commun. Math. Comput. Chem.*, **12**, 87–107.
- Lall, R.S. (1982) Steric effect I – Charton's parameter vs topological index. *Curr. Sci. -India*, **51**, 775–777.
- Lall, R.S. (1984) Topology of oxy-organic compounds. *Curr. Sci. -India*, **53**, 642–643.
- Lall, R.S. (1990) Structure–activity relationship of organo-phosphorous insecticides. *Asian J. Chem.*, **2**, 37–42.
- Lamanna, C., Catalano, A., Carocci, A., Di Mola, A., Franchini, C., Tortorella, V., Vanderheyden, P.M. L., Sinicropi, M.S., Watson, K.A. and Sciabola, S.

- (2007) AT₁ receptor ligands: virtual-screening-based design with TOPP descriptors, synthesis, and biological evaluation of pyrrolidine derivatives. *ChemMedChem*, **2**, 1298–1310.
- Lamarche, O. and Platts, J.A. (2003) Atoms in molecules investigation of the pK_{HB} basicity scale: electrostatic and covalent effects in hydrogen bonding. *Chem. Phys. Lett.*, **367**, 123–128.
- Lamarche, O., Platts, J.A. and Hersey, A. (2001) Theoretical prediction of the polarity/polarizability parameter π^H_2 . *Phys. Chem. Chem. Phys.*, **3**, 2747–2753.
- Lambert, F.L. (1966) Polarography of organic halogen compounds. III. Quantitative correlation of the half wave potentials of alkyl bromides with Taft polar and steric constants. *J. Org. Chem.*, **31**, 4184–4188.
- Landon, M.R. and Schaus, S.E. (2006) JEDA: joint entropy diversity analysis. An information-theoretic method for choosing diverse and representative subsets from combinatorial libraries. *Mol. Div.*, **10**, 333–339.
- Lang, P.-Z., Ma, X.-F., Lu, G.-H., Wang, Y. and Bian, Y. (1996) QSAR for the acute toxicity of nitroaromatics to the carp (*Cyprinus carpio*) *Chemosphere*, **32**, 1547–1552.
- Lang, S.A., Kozyukov, A.V., Balakin, K.V., Skorenko, A.V., Ivashchenko, A.A. and Savchuk, N.P. (2002) Classification scheme for the design of serine protease targeted compound libraries. *J. Comput. Aid. Mol. Des.*, **16**, 803–807.
- Langenaeker, W. and Liu, S. (2001) The response of atomic electron densities to point perturbations in the external potential. *J. Mol. Struct. (Theochem)*, **535**, 279–286.
- Langer, T. (1994) Molecular similarity determination of heteroaromatics using CoMFA and multivariate data analysis. *Quant. Struct. -Act. Relat.*, **13**, 402–405.
- Langer, T. and Hoffmann, R.D. (1998a) New principal components derived parameters describing molecular diversity of heteroaromatic residues. *Quant. Struct. -Act. Relat.*, **17**, 211–223.
- Langer, T. and Hoffmann, R.D. (1998b) On the use of chemical function-based alignments as input for 3D-QSAR. *J. Chem. Inf. Comput. Sci.*, **38**, 325–330.
- Langlois, M., Bremont, B., Rousselle, D. and Gaudy, F. (1993) Structural analysis by the comparative molecular field analysis method of the affinity of beta adrenoceptor blocking agents for 5-Ht1A and 5-Ht1B receptors. *Eur. J. Pharmacol.*, **244**, 77–87.
- Langlois, M.H., Audry, E., Croizet, F., Dallet, P., Carpy, A. and Dubost, J.P. (1993) Topological lipophilicity potential: a new tool for a fast evaluation of lipophilicity distribution on a molecular graph, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 354–355.
- Lanteri, S. (1992) Full validation procedures for feature selection in classification and regression problems. *Chemom. Intell. Lab. Syst.*, **15**, 159–169.
- Lapinsh, M., Prusis, P., Mutule, I., Mutulis, F. and Wikberg, J.E.S. (2003) QSAR and proteo-chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes. *J. Med. Chem.*, **46**, 2572–2579.
- Larsson, J., Gottfries, J., Bohlin, L. and Backlund, A. (2005) Expanding the ChemGPS chemical space with natural products. *J. Nat. Prod.*, **68**, 985–991.
- Laskowski, D.A., Goring, C.A.I., McCall, P.J. and Swann, R.L. (1982) Terrestrial environment, in *Environment Risk Analysis for Chemicals* (ed. R.A. Conway), Van Nostrand Reinhold Company, New York, pp. 198–240.
- Lassau, C. and Jungers, J.-C. (1968) N°397 – L'Influence du Solvant sur la Réaction Chimique. La Quaternation des Amines Tertiaires par l'Iodure de Méthyle. *Bull. Soc. Chim. Fran. (French)*, **7**, 2678–2685.
- László, I. (2004) Topological aspects beyond the Hückel theory. *Internet Electron. J. Mol. Des.*, **3**, 182–188.
- Lather, V. and Madan, A.K. (2004) Models for the prediction of adenosine receptors binding activity of 4-amino[1,2,4]triazolo[4,3-a]quinoxalines. *J. Mol. Struct. (Theochem)*, **678**, 1–9.
- Lather, V. and Madan, A.K. (2005a) Application of graph theory: topological models for prediction of CDK-1 inhibitory activity of aloisines. *Croat. Chem. Acta*, **78**, 55–61.
- Lather, V. and Madan, A.K. (2005b) Topological models for the prediction of HIV-protease inhibitory activity of tetrahydropyrimidin-2-ones. *J. Mol. Graph. Model.*, **23**, 339–345.
- Latino, A.R.S. and Aires-de-Sousa, J. (2006) Genome-scale classification of metabolic reactions: a chemoinformatics approach. *Angew. Chem. Int. Ed. Engl.*, **45**, 2066–2069.
- Laurence, C., Berthelot, M., Lucon, M., Helbert, M., Morris, D.G. and Gal, J.-F. (1984) The influence of solvent on the inductive order of substituents from infrared measurements on 4-substituted camphors: a new model of inductive effects. *J. Chem. Soc. Perkin Trans. 2*, 705–710.
- Lavenhar, S.R. and Maczka, C.A. (1985) Structure-activity considerations in risk assessment: a simulation study. *Toxicol. Ind. Health*, **1**, 249–259.
- Lavine, B.K., Davidson, C.E., Breneman, C.M. and Katt, W. (2003) Electronic van der Waals surface

- property descriptors and genetic algorithms for developing structure–activity correlations in olfactory databases. *J. Chem. Inf. Comput. Sci.*, **43**, 1890–1905.
- Lawson, R.G. and Jurs, P.C. (1990) New index for clustering tendency and its application to chemical problems. *J. Chem. Inf. Comput. Sci.*, **30**, 36–41.
- Lazzeretti, P. (2004) Assessment of aromaticity via molecular response properties. *Phys. Chem. Chem. Phys.*, **6**, 217–223.
- Le, S.Y., Nussinov, R. and Maizel, J.V. (1989) Tree graphs of RNA secondary structure and their comparisons. *Comput. Biomed. Res.*, **22**, 461–473.
- Le, T.D. and Weers, J.G. (1995) QSPR and GCA models for predicting the normal boiling points of fluorocarbons. *J. Phys. Chem.*, **99**, 6739–6747.
- Leach, A.R. (1996) *Molecular Modelling. Principles and Applications*, Longman, Singapore, p. 596.
- Leach, A.R., Bradshaw, J., Green, D.V.S., Hann, M.M. and Delany, J.J. III (1999) Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci.*, **39**, 1161–1172.
- Leach, A.R. and Gillet, V.J. (2003) *An Introduction to Chemoinformatics*, Kluwer Academic Publishers, Dordrecht, The Netherlands, p. 259.
- Leach, A.R., Hann, M.M., Burrows, J.N. and Griffen, E.J. (2006) Fragment screening: an introduction. *Mol. BioSyst.*, **2**, 429–446.
- Leahy, D.E. (1986) Intrinsic molecular volume as a measure of the cavity term in linear solvation energy relationships: octanol–water partition coefficients and aqueous solubilities. *J. Pharm. Sci.*, **75**, 629–636.
- Leahy, D.E., Morris, J.J., Taylor, P.J. and Wait, A.R. (1992a) Model solvent systems for QSAR. 3. An LSER analysis of the critical quartet: new light on hydrogen bond strength and directionality. *J. Chem. Soc. Perkin Trans. 2*, 705–722.
- Leahy, D.E., Morris, J.J., Taylor, P.J. and Wait, A.R. (1992b) Model solvent systems for QSAR. Part 2. Fragment values ('f-values') for the 'critical quartet'. *J. Chem. Soc. Perkin Trans. 2*, 723–731.
- Leahy, D.E., Morris, J.J., Taylor, P.J. and Wait, A.R. (1994) Model Solvent systems for QSAR. 4. The hydrogen bond acceptor behavior of heterocycles. *J. Phys. Org. Chem.*, **7**, 743–750.
- Leão, M.B.C., Pavão, A.C., Espinoza, V.A.A., Taft, C.A. and Bulnes, E.P. (2005) A multivariate model of chemical carcinogenesis. *J. Mol. Struct. (Theochem)*, **719**, 129–135.
- Leardi, R. (1994) Application of genetic algorithms to feature selection under full validation conditions and to outlier detection. *J. Chemom.*, **8**, 65–79.
- Leardi, R. (1996) Genetic algorithms in feature selection, in *Genetic Algorithms in Molecular Modeling. Principles of QSAR and Drug Design*, Vol.1 (ed. J. Devillers), Academic Press, London, UK, pp. 67–86.
- Leardi, R. (2001) Genetic algorithms in chemometrics and chemistry: a review. *J. Chemom.*, **15**, 559–569.
- Leardi, R. (ed.) (2003) *Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*, Elsevier, Amsterdam, The Netherlands, p. 384.
- Leardi, R., Boggia, R. and Terrile, M. (1992) Genetic algorithms as a strategy for feature selection. *J. Chemom.*, **6**, 267–281.
- Leardi, R. and Lupiañez Gonzalez, A. (1998) Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.*, **41**, 195–207.
- Lebez, M., Šolmajer, T. and Zupan, J. (2002) Quantitative structure–activity relationship of tricyclic carbapenems: application of artificial intelligence methods for bioactivity prediction. *Croat. Chem. Acta*, **75**, 545–562.
- Ledermann, W. and Vajda, S. (eds) (1980) *Handbook of Applicable Mathematics, Algebra*, Vol. 1, John Wiley & Sons, Ltd, Chichester, UK, p. 524.
- Lee, B. and Richards, F.M. (1971) The interpretation of protein structure: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
- Lee, D.W., Kim, M.K., Kim, I.W., Park, J.H. and No, K. T. (1996) Studies on the chromatographic behaviors of Pd(II)-alpha-isonitroso-beta-diketone imine chelates in reversed-phase liquid chromatography using molecular descriptors. *Bull. Kor. Chem. Soc.*, **17**, 1158–1161.
- Lee, K.W., Kwon, S.Y., Hwang, S., Lee, J.U. and Kim, H.J. (1996) Quantitative structure–activity relationships (QSAR) study on C-7 substituted quinolone. *Bull. Kor. Chem. Soc.*, **17**, 147–152.
- Lee, K.W. and Briggs, J.M. (2001) Comparative molecular field analysis (CoMFA) study of epothilones – tubulin depolymerization inhibitors: pharmacophore development using 3D QSAR methods. *J. Comput. Aid. Mol. Des.*, **15**, 41–55.
- Lee, K.-H. (2004) Current developments in the discovery and design of new drug candidates from plant natural product leads. *J. Nat. Prod.*, **67**, 273–283.
- Lee, S.K., Park, Y.H., Yoon, C.J. and Lee, D.W. (1998) Investigation of relationships between retention behavior and molecular descriptors of quinolones in PRP-1 column. *J. Microcol. Sep.*, **10**, 133–139.

- Lee, S.-L. and Yeh, Y.-N. (1993) On eigenvalues and eigenvectors of graphs. *J. Math. Chem.*, **12**, 121–135.
- Leegwater, D.C. (1989) QSAR-analysis of acute toxicity of industrial pollutants to the guppy using molecular connectivity indices. *Aquat. Toxicol.*, **15**, 157–168.
- Legendre, P. and Legendre, L. (1998) *Numerical Ecology*, Elsevier, Amsterdam, The Netherlands, p. 854.
- Lehtonen, P. (1987) Molecular connectivity indices in the prediction of the retention of oxygen-containing amines in reversed-phase liquid chromatography. *J. Chromat.*, **398**, 143–151.
- Leicester, S.E., Finney, J.L. and Bywater, R.P. (1988) Description of molecular surface shape using Fourier descriptors. *J. Mol. Graph.*, **6**, 104–108.
- Leicester, S.E., Finney, J.L. and Bywater, R.P. (1994a) A quantitative representation of molecular surface shape. I. Theory and development of the method. *J. Math. Chem.*, **16**, 315–341.
- Leicester, S.E., Finney, J.L. and Bywater, R.P. (1994b) A quantitative representation of molecular surface shape. II. Protein classification using Fourier shape descriptors and classical scaling. *J. Math. Chem.*, **16**, 343–365.
- Leip, A. and Lammell, G. (2004) Indicators for persistence and long-range transport potential as derived from multicompartiment chemistry–transport modelling. *Environmental Pollution*, **128**, 205–221.
- Leis, J. and Karelson, M. (2001) A QSPR model for the prediction of the gas-phase free energies of activation of rotation around the N–C(O) bond. *Computers Chem.*, **25**, 171–176.
- Lekishvili, G. (1997) On the characterization of molecular stereostructure. 1. *cis-trans* isomerism. *J. Chem. Inf. Comput. Sci.*, **37**, 924–928.
- Lemmen, C. and Lengauer, T. (2000) Computational methods for the structural alignment of molecules. *J. Comput. Aid. Mol. Des.*, **14**, 215–232.
- Lemmen, C., Lengauer, T. and Klebe, G. (1998) FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.*, **41**, 4502–4520.
- Lendvay, G. (2000) On the correlation of bond order and bond length. *J. Mol. Struct. (Theochem)*, **501–502**, 389–393.
- Lennard-Jones, J.E. (1924) On the determination of molecular fields. 11. The equation of state of a gas. *Proc. Roy. Soc. London A*, **106**, 463–477.
- Lennard-Jones, J.E. (1929) The electronic structure of some diatomic molecules. *Trans. Faraday Soc.*, **25**, 668–686.
- Leo, A. (1987) Some advantages of calculating octanol–water partition coefficients. *J. Pharm. Sci.*, **76**, 166–168.
- Leo, A. (1990) Methods of calculating partition coefficients, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 295–319.
- Leo, A. (1991) Hydrophobic parameter: measurement and calculation. *Methods Enzymol.*, **202**, 544–591.
- Leo, A. (1993) Calculating $\log P_{\text{oct}}$ from structures. *Chem. Rev.*, **93**, 1281–1306.
- Leo, A. and Hansch, C. (1971) Linear free-energy relationships between partitioning solvent systems. *J. Org. Chem.*, **36**, 1539–1544.
- Leo, A., Hansch, C. and Elkins, D. (1971) Partition coefficients and their uses. *Chem. Rev.*, **71**, 525–616.
- Leo, A., Hansch, C. and Jow, P.Y.C. (1976) Dependence of hydrophobicity of apolar molecules on their molecular volume. *J. Med. Chem.*, **19**, 611–615.
- Leo, A., Jow, P.Y.C., Silipo, C. and Hansch, C. (1975) Calculation of hydrophobic constant ($\log P$) from π and f constants. *J. Med. Chem.*, **18**, 865–868.
- Leonard, J.T. and Roy, K. (2004) Classical QSAR modeling of CCR5 receptor binding affinity of substituted pyrazoles. *QSAR Comb. Sci.*, **23**, 387–398.
- Leong, P.M. and Mogenthaler, S. (1995) Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.*, **12**, 503–511.
- Lepoittevin, J.-P. and Roy, K. (2006) On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb. Sci.*, **25**, 235–251.
- Lepovic, M. and Gutman, I. (1998) A collective property of trees and chemical trees. *J. Chem. Inf. Comput. Sci.*, **38**, 823–826.
- Lerche, D., Sørensen, P.B. and Brüggemann, R. (2003) Improved estimation of the ranking probabilities in partial orders using random linear extensions by approximation of the mutual ranking probability. *J. Chem. Inf. Comput. Sci.*, **43**, 1471–1480.
- Lessel, U.F. and Briem, H. (2000) Flexsim-X: a method for the detection of molecules with similar biological activity. *J. Chem. Inf. Comput. Sci.*, **40**, 246–253.
- Lewis, D.F., Ioannides, C. and Parke, D.V. (1995) A quantitative structure–activity relationship study on a series of 10 parasubstituted toluenes binding to cytochrome P450B4 (Cyp2B4), and their hydroxylation rates. *Biochem. Pharmacol.*, **50**, 619–625.
- Lewis, D.F. and Parke, D.V. (1995) The genotoxicity of benzantracenes: a quantitative structure–activity study. *Mut. Res.*, **328**, 207–214.

- Lewis, D.F.V. (1989) The calculation of molar polarizabilities by the CNDO/2 method: correlation with the hydrophobic parameter, log *P*. *J. Comput. Chem.*, **10**, 145–151.
- Lewis, D.F.V. and Dickins, M. (2002) Factors influencing rates and clearance in P450-mediated reactions: QSARs for substrates of the xenobiotic-metabolizing hepatic microsomal P450s. *Toxicology*, **170**, 45–53.
- Lewis, E.S. and Johnson, M.D. (1959) The substituent constants of the diazonium ion group. *J. Am. Chem. Soc.*, **81**, 2070–2072.
- Lewis, G., Mathieu, D. and Phan-Tan-Lu, R. (1999) *Pharmaceutical experimental design*, Marcel Dekker, Inc., New York, p. 498.
- Lewis, G.N. (1916) The atom and the molecule. *J. Am. Chem. Soc.*, **38**, 762–785.
- Lewis, G.N. (1923) *Valence and the Structure of Atoms and Molecules*, Dover, New York.
- Lewis, R.A., Mason, J.S. and McLay, I.M. (1997) Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. *J. Chem. Inf. Comput. Sci.*, **37**, 599–614.
- Leyssens, T., Geerlings, P. and Peeters, D. (2005) The importance of the external potential on group electronegativity. *J. Phys. Chem. A*, **109**, 9882–9889.
- Li, C. and Wang, J. (2005) New invariant of DNA sequences. *J. Chem. Inf. Model.*, **45**, 115–120.
- Li, H. and Lu, M. (2005) The *m*-connectivity index of graphs. *MATCH Commun. Math. Comput. Chem.*, **54**, 417–423.
- Li, H., Ung, C.Y., Yap, C.W., Xue, Y., Li, Z.R., Cao, Z. W. and Chen, Y.Z. (2005) Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem. Res. Toxicol.*, **18**, 1071–1080.
- Li, H., Yap, C.W., Ung, C.Y., Xue, Y., Cao, Z.W. and Chen, Y.Z. (2005) Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model.*, **45**, 1376–1384.
- Li, H., Xu, L., Yang, Y.-Q. and Su, Q. (1996) Quantitative structure–property relationships for colour reagents and their colour reactions with ytterbium using regression analysis and computational neural networks. *Anal. Chim. Acta*, **321**, 97–103.
- Li, J., Liu, H., Yao, X.-J., Liu, M., Hu, Z. and Fan, B.T. (2007) Quantitative structure–activity relationship study of acyl ureas as inhibitors of human liver glycogen phosphorylase using least squares support vector machines. *Chemom. Intell. Lab. Syst.*, **87**, 139–146.
- Li, L.-F. and You, X.-Z. (1993a) A topological index and its application. Part 3. Estimations of the enthalpies of formation of mixed halogen-substituted methanes, silanes and boron mixed halides. *Thermochim. Acta*, **225**, 85–96.
- Li, L.-F. and You, X.-Z. (1993b) Molecular topological index and its application. 1. On the chemical shifts of ^{95}Mo NMR and ^{119}Sn Mössbauer spectroscopy. *Chinese Sci. Bull.*, **38**, 421–425.
- Li, L., Mao, S., Zhao, K. and Tian, A. (2001) Semi-empirical quantum chemical study on structure–activity relationship in monocyclic- β -lactam antibiotics. *J. Mol. Struct. (Theochem)*, **545**, 1–5.
- Li, L.-F., Zhang, Y. and You, X.-Z. (1995) Molecular topological index and its application. 4. Relationships with the diamagnetic susceptibilities of alkyl-IVA group organometallic halides. *J. Chem. Inf. Comput. Sci.*, **35**, 697–700.
- Li, M.-J., Jiang, C., Li, M.-Z. and You, T.-P. (2005) QSAR studies of 20(*S*)-camptothecin analogues as antitumor agents. *J. Mol. Struct. (Theochem)*, **723**, 165–170.
- Li, Q., Chen, X. and Hu, Z. (2004) Quantitative structure–property relationship studies for estimating boiling points of alcohols using calculated molecular descriptors with radial basis function neural networks. *Chemom. Intell. Lab. Syst.*, **72**, 93–100.
- Li, Q., Bender, A., Pei, J. and Lai, L. (2007) A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification. *J. Chem. Inf. Model.*, **47**, 1776–1786.
- Li, S., Fedorowicz, A., Singh, H. and Soderholm, S.C. (2005) Application of the random forest method in studies of local lymph node assay based skin sensitization data. *J. Chem. Inf. Model.*, **45**, 952–964.
- Li, T., Mei, H. and Cong, P. (1999) Combining nonlinear PLS with the numeric genetic algorithm for QSAR. *Chemom. Intell. Lab. Syst.*, **45**, 177–184.
- Li, W.-Y., Guo, Z.-R. and Lien, E.J. (1984) Examination of the interrelationship between aliphatic group dipole moment and polar substituent constants. *J. Pharm. Sci.*, **73**, 553–558.
- Li, X. (2002) The extended Wiener index. *Chem. Phys. Lett.*, **365**, 135–139.
- Li, X. and Gutman, I. (2006) *Mathematical Aspects of Randić-Type Molecular Structure Descriptors*, University of Kragujevac, Kragujevac, Serbia.
- Li, X. and Jalbout, A.F. (2003) Bond order weighted hyper-Wiener index. *J. Mol. Struct. (Theochem)*, **634**, 121–125.
- Li, X., Jalbout, A.F. and Solimannejad, M. (2003) Definition and application of a novel valence

- molecular connectivity index. *J. Mol. Struct. (Theochem)*, **663**, 81–85.
- Li, X., Li, Z. and Hu, M. (2003) A novel set of Wiener indices. *J. Mol. Graph. Model.*, **22**, 161–172.
- Li, X. and Lin, J. (2002) The valence overall Wiener index for unsaturated hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **42**, 1358–1362.
- Li, X., Yu, Q.-S. and Zhu, L. (2000) A novel quantum-topology index. *J. Chem. Inf. Comput. Sci.*, **40**, 399–492.
- Li, X., Zhang, G., Dong, J., Zhou, X., Yan, X. and Luo, M. (2004) Estimation of critical micelle concentration of anionic surfactants with QSPR approach. *J. Mol. Struct. (Theochem)*, **710**, 119–126.
- Li, X., Zhao, H. and Gutman, I. (2005) On the Merrifield–Simmons index of trees. *MATCH Commun. Math. Comput. Chem.*, **54**, 389–402.
- Li, X. and Zhao, H. (2004) Trees with the first three smallest and largest generalized topological indices. *MATCH Commun. Math. Comput. Chem.*, **50**, 57–62.
- Li, Y., Hu, Q. and Zhong, C. (2004) Topological modeling of the Setschenow constant. *Ind. Eng. Chem. Res.*, **43**, 4465–4468.
- Li, Y., Liu, J., Pan, D. and Hopfinger, A.J. (2005) A study of the relationship between cornea permeability and eye irritation using membrane-interaction QSAR analysis. *Toxicol. Sci.*, **88**, 434–446.
- Li, Z., Fu, B., Wang, Y. and Liu, S. (2001) On structural parametrization and molecular modeling of peptide analogues by molecular electronegativity edge vector (VMEE): estimation and prediction for biological activity of dipeptides. *Journal of Chinese Chemical Society*, **48**, 937–944.
- Li, Z., Dai, Y.-M., Wen, S.-N., Nie, C. and Zhou, C. (2005) Relationship between atom valence shell electron quantum topological indices and electronegativity of elements. *Acta Chim. Sin.*, **63**, 1348–1356.
- Liang, C. and Gallagher, D.A. (1998) QSPR prediction of vapor pressure from solely theoretically-derived descriptors. *J. Chem. Inf. Comput. Sci.*, **38**, 321–324.
- Liang, C. and Mislow, K. (1994) Topological chirality of proteins. *J. Am. Chem. Soc.*, **116**, 3588–3592.
- Liang, G.-Z., Zhou, P., Zhou, Y., Zhang, Q.-L. and Li, Z. (2006) New descriptors of amino acids and their applications to peptide quantitative structure–activity relationship. *Acta Chim. Sin.*, **64**, 393–396.
- Liao, B. (2005) A 2D graphical representation of DNA sequence. *Chem. Phys. Lett.*, **401**, 196–199.
- Liao, B. and Ding, K. (2005) Graphical approach to analyzing DNA sequences. *J. Comput. Chem.*, **14**, 1519–1523.
- Liao, B., Ding, K. and Wang, T. (2005) On a six-dimensional representation of RNA secondary structures. *J. Biomol. Struct. Dyn.*, **22**, 455–464.
- Liao, B., Tan, M. and Ding, K. (2005a) A 4D representation of DNA sequences and its applications. *Chem. Phys. Lett.*, **402**, 380–383.
- Liao, B., Tan, M. and Ding, K. (2005b) Application of 2-D graphical representation of DNA sequence. *Chem. Phys. Lett.*, **414**, 296–300.
- Liao, B. and Wang, T. (2004a) 3-D graphical representation of DNA sequences and their numerical characterization. *J. Mol. Struct. (Theochem)*, **681**, 209–212.
- Liao, B. and Wang, T. (2004b) A 3D graphical representation of RNA secondary structure. *J. Biomol. Struct. Dyn.*, **21**, 827–832.
- Liao, B. and Wang, T. (2004c) Analysis of similarity of DNA sequences based on 3D graphical representation. *Chem. Phys. Lett.*, **388**, 195–200.
- Liao, B. and Wang, T. (2004d) New 2D graphical representation of DNA sequences. *J. Comput. Chem.*, **25**, 1364–1368.
- Liao, B., Wang, T. and Ding, K. (2005) On a seven-dimensional representation of RNA secondary structures. *Mol. Simulat.*, **31**, 1063–1071.
- Liao, B., Zhang, Y.S., Ding, K. and Wang, T. (2005) Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation. *J. Mol. Struct. (Theochem)*, **717**, 199–203.
- Liao, B., Zhu, W., Luo, J. and Li, R. (2007) RNA secondary structure mathematical representation without degeneracy. *MATCH Commun. Math. Comput. Chem.*, **57**, 687–695.
- Liao, Q., Yao, J.H. and Yuan, S. (2006) SVM approach for predicting log P. *Mol. Div.*, **10**, 301–309.
- Lias, S.G., Liebman, J.F. and Levin, R.D. (1984) Evaluated gas phase basicities and proton affinities of molecules; heats of formation of protonated molecules. *J. Phys. Chem. Ref. Data*, **13**, 695–808.
- Lide, D.R. (ed.) (1999) *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, FL.
- Lien, E.J. and Guo, Z.-R., Li, R.-L. and Su, C.-T. (1982) Use of dipole moment as a parameter in drug–receptor interaction and quantitative structure–activity relationship studies. *J. Pharm. Sci.*, **71**, 641–655.
- Lien, E.J., Liao, R.C.H. and Shinouda, H.G. (1979) Quantitative structure–activity relationships and dipole moments of anticonvulsants and CNS depressants. *J. Pharm. Sci.*, **68**, 463–465.
- Liljefors, T. (1998) Progress in force-field calculations of molecular interaction fields and intermolecular interactions, in *3D QSAR in Drug Design*, Vol. 2

- (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 3–17.
- Lima, L.M. and Barreiro, E.J. (2005) Bioisosterism: a useful strategy for molecular modification and drug design. *Curr. Med. Chem.*, **12**, 23–49.
- Lin, C.-D and Fan, G.-Q. (1999) Algorithms for the count of linearly independent and minimal conjugated circuits in benzenoid hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **39**, 782–787.
- Lin, S.-T. and Sandler, S.I. (1999) Prediction of octanol–water partition coefficients using a group contribution solvation model. *Ind. Eng. Chem. Res.*, **38**, 4081–4091.
- Lin, S.-K. (1996a) Correlation of entropy with similarity and symmetry. *J. Chem. Inf. Comput. Sci.*, **36**, 367–376.
- Lin, S.-K. (1996) Molecular diversity assessment: logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguishability after rejection of Gibbs paradox of entropy mixing. *Molecules*, **1**, 57–67.
- Lin, T.-C. (2004) Mass-modified Wiener indices and boiling points for lower chloroalkanes. *Acta Chim. Sloven.*, **51**, 611–618.
- Lin, T.-H. and Lin, J.-J. (2001) Three-dimensional quantitative structure–activity relationship for several bioactive peptides searched by a convex hull-comparative molecular field analysis approach. *Computers Chem.*, **25**, 489–498.
- Lin, T.-H. and Tsai, K.-C. (2003) Implementing the Fisher's discriminant ratio in *k*-means clustering algorithm for feature selection and data set trimming. *J. Chem. Inf. Comput. Sci.*, **44**, 76–87.
- Lin, T.-H., Wang, G.-M. and Hsu, Y.-H. (2002) Classification of some active HIV-1 protease inhibitors and their inactive analogues using some uncorrelated three-dimensional molecular descriptors and a fuzzy *c*-means algorithm. *J. Chem. Inf. Comput. Sci.*, **42**, 1490–1504.
- Lin, T.-H., Yu, Y.-S. and Chen, H.-J. (2000) Classification of some active compounds and their inactive analogues using two three-dimensional molecular descriptors derived from computation of three-dimensional convex hulls for structures theoretically generated for them. *J. Chem. Inf. Comput. Sci.*, **40**, 1210–1221.
- Lin, Z., Yin, K., Shi, P., Wang, L.-S. and Yu, H. (2003) Development of QSARs for predicting the joint effects between cyanogenic toxicants and aldehydes. *Chem. Res. Toxicol.*, **16**, 1365–1371.
- Lin, Z., Zhong, P., Yin, K., Wang, L.-S. and Yu, H. (2003) Quantification of joint effect for hydrogen bond and development of QSARs for predicting mixture toxicity. *Chemosphere*, **52**, 1199–1208.
- Lind, P., Lopes, C., Öberg, K. and Eliasson, B. (2004) A QSPR study on optical limiting of organic compounds. *Chem. Phys. Lett.*, **387**, 238–242.
- Lindgren, F. (1994) Third generation PLS. Some elements and applications. PhD Thesis, Umeå University, Umeå, Sweden.
- Lindgren, F., Geladi, P., Berglund, A., Sjöström, M. and Wold, S. (1995) Interactive variable selection (IVS) for PLS. Part II. Chemical applications. *J. Chemom.*, **9**, 331–342.
- Lindgren, F., Geladi, P., Rännar, S. and Wold, S. (1994) Interactive variable selection (IVS) for PLS. Part I. Theory and algorithms. *J. Chemom.*, **8**, 349–363.
- Lindgren, F., Hansen, B., Karcher, W., Sjöström, M. and Eriksson, L. (1996) Model validation by permutation tests: applications to variable selection. *J. Chemom.*, **10**, 521–532.
- Lindgren, F. and Rännar, S. (1998) Alternative partial least-squares (PLS) algorithms, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 105–113.
- Linert, W., Klestorfer, K., Renz, F. and Lukovits, I. (1995) Description of cyclic and branched-acyclic hydrocarbons by variants of the hyper-Wiener index. *J. Mol. Struct. (Theochem)*, **337**, 121–127.
- Linert, W. and Lukovits, I. (1997) Formulas for the hyper-Wiener and hyper-Detour indices of fused bicyclic structures. *MATCH Commun. Math. Comput. Chem.*, **35**, 65–74.
- Linert, W., Renz, F., Klestorfer, K. and Lukovits, I. (1995) An algorithm for the computation of the hyper-Wiener index for the characterization and discrimination of branched acyclic molecules. *Computers Chem.*, **19**, 395–401.
- Linusson, A., Gottfries, J., Lindgren, F. and Wold, S. (2000) Statistical molecular design of building blocks for combinatorial chemistry. *J. Med. Chem.*, **43**, 1320–1328.
- Lipinski, C.A. (2000) Drug-like properties and the cause of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, **44**, 235–249.
- Lipinski, C.A. (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today: Technologies*, **1**, 337–341.
- Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **23**, 3–25.

- Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **46**, 3–36.
- Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (2005) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **23**, 3–25.
- Lipkowitz, K.B., Baker, B. and Larter, R. (1989) Dynamic molecular surface areas. *J. Am. Chem. Soc.*, **111**, 7750–7753.
- Lipkowitz, K.B. and Boyd, D. (eds) (1990) *Reviews in Computational Chemistry*. Vols 1–13, Wiley-VCH Verlag GmbH, Weinheim, Germany.
- Lipkowitz, K.B. and Boyd, D. (eds) (1997) *Reviews in Computational Chemistry*. Vol. 11, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 431.
- Lipkus, A.H. (1997) A ring-imbedding index and its use in substructure searching. *J. Chem. Inf. Comput. Sci.*, **37**, 92–97.
- Lipkus, A.H. (1999) Mining a large database for peptidomimetic ring structures using a topological index. *J. Chem. Inf. Comput. Sci.*, **39**, 582–586.
- Lipkus, A.H. (2001) Exploring chemical rings in a simple topological-descriptor space. *J. Chem. Inf. Comput. Sci.*, **41**, 430–438.
- Lipnick, R.L. (1990) Narcosis: fundamental and baseline toxicity mechanism for nonelectrolyte organic chemicals, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 281–293.
- Lipnick, R.L. (1991) Outliers: their origin and use in the classification of molecular mechanisms of toxicity. *Sci. Total Environ.*, **109/110**, 131–153.
- Lister, D.G., Macdonald, J.N. and Owen, N.L. (1978) *Internal Rotation and Inversion*, Academic Press, London, UK.
- Liu, B. and Gutman, I. (2006) Upper bounds for Zagreb indices of connected graphs. *MATCH Commun. Math. Comput. Chem.*, **55**, 439–446.
- Liu, B. and Gutman, I. (2007) Estimating the Zagreb and the general Randić indices. *MATCH Commun. Math. Comput. Chem.*, **57**, 617–632.
- Liu, D.X., Jiang, H., Chen, K. and Ji, R.Y. (1998) A new approach to design virtual combinatorial library with genetic algorithm based on 3D grid property. *J. Chem. Inf. Comput. Sci.*, **38**, 233–242.
- Liu, F., Liang, Y.-Z. and Cao, C. (2006) QSPR modeling of thermal conductivity detection response factors for diverse organic compound. *Chemom. Intell. Lab. Syst.*, **81**, 120–126.
- Liu, H., Papa, E., Walker, J.D. and Gramatica, P. (2007) *In silico* screening of estrogen-like chemicals based on different nonlinear classification models. *J. Mol. Graph. Model.*, **26**, 135–144.
- Liu, H., Xiang, B. and Qu, L. (2007) The application of rough sets in SAR analysis of N1-site substituted fluoroquinolones. *Chemom. Intell. Lab. Syst.*, **87**, 155–160.
- Liu, H. and Zhong, C. (2005) General correlation for the prediction of theta (lower critical solution temperature) in polymer solutions. *Ind. Eng. Chem. Res.*, **44**, 634–638.
- Liu, H. and Gramatica, P. (2007) QSAR study of selective ligands for the thyroid hormone receptor β . *Bioorg. Med. Chem.*, **15**, 5251–5261.
- Liu, H., Hu, R.J., Zhang, R., Yao, X.-J., Liu, M., Hu, Z. and Fan, B.T. (2005) The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. *J. Comput. Aid. Mol. Des.*, **19**, 33–46.
- Liu, H., Xue, C., Zhang, R., Yao, X.-J., Liu, H., Hu, Z. and Fan, B.T. (2004) Quantitative prediction of log k of peptides in high-performance liquid chromatography based on molecular descriptors by using the heuristic method and support vector machine. *J. Chem. Inf. Comput. Sci.*, **44**, 1979–1986.
- Liu, H., Yao, X.-J., Liu, M., Hu, Z. and Fan, B.T. (2006) Prediction of retention in micellar electrokinetic chromatography based on molecular structural descriptors by using the heuristic method. *Anal. Chim. Acta*, **558**, 86–93.
- Liu, H., Zhang, R., Yao, X.-J., Liu, M., Hu, Z. and Fan, B.T. (2004) QSAR and classification models of a novel series of COX-2 selective inhibitors: 1,5-diarylimidazoles based on support vector machines. *J. Comput. Aid. Mol. Des.*, **18**, 389–399.
- Liu, H., Lu, M. and Tian, F. (2005) On the Randić index. *J. Math. Chem.*, **38**, 345–354.
- Liu, H., Pan, X. and Xu, J.-M. (2006) On the Randić index of unicyclic conjugated molecules. *J. Math. Chem.*, **40**, 135–143.
- Liu, J. and Qian, C. (1995) Hydrophobic coefficients of s-triazine and phenylurea herbicides. *Chemosphere*, **31**, 3951–3959.
- Liu, J., Yang, L., Li, Y., Pan, D. and Hopfinger, A.J. (2006) Constructing plasma protein binding model based on a combination of cluster analysis and 4D-fingerprint molecular similarity analyses. *Bioorg. Med. Chem.*, **14**, 611–621.
- Liu, L., Fu, Y., Liu, R., Li, R.-Q. and Guo, Q.X. (2004) Hammett equation and generalized Pauling's

- electronegativity equation. *J. Chem. Inf. Comput. Sci.*, **44**, 652–657.
- Liu, L. and Guo, Q.X. (1999) Wavelet neural network and its application to the inclusion of β -cyclodextrin with benzene derivatives. *J. Chem. Inf. Comput. Sci.*, **39**, 133–138.
- Liu, Q., Hirono, S. and Moriguchi, I. (1992a) Application of functional link net in QSAR. 1. QSAR for activity data given by continuous variate. *Quant. Struct. -Act. Relat.*, **11**, 135–141.
- Liu, Q., Hirono, S. and Moriguchi, I. (1992b) Application of functional link net in QSAR. 2. QSAR for activity data given by ratings. *Quant. Struct. -Act. Relat.*, **11**, 318–324.
- Liu, R. and So, S.-S. (2001) Development of quantitative structure–property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. *J. Chem. Inf. Comput. Sci.*, **41**, 1633–1639.
- Liu, R., Sun, H. and So, S.-S. (2001) Development of quantitative structure–property relationship models for early ADME evaluation in drug discovery. 2. Blood–brain barrier penetration. *J. Chem. Inf. Comput. Sci.*, **41**, 1623–1632.
- Liu, R. and Zhou, D. (2008) Using molecular fingerprint as descriptors in the QSPR study of lipophilicity. *J. Chem. Inf. Model.*, **48**, 542–549.
- Liu, S., Zhang, R., Liu, M. and Hu, Z. (1997) Neural network-topological indices approach to the prediction of properties of alkene. *J. Chem. Inf. Comput. Sci.*, **37**, 1146–1151.
- Liu, S., Cai, S.-X., Cao, C. and Li, Z. (2000) Molecular electronegative distance vector (MEDV) related to 15 properties of alkanes. *J. Chem. Inf. Comput. Sci.*, **40**, 1337–1348.
- Liu, S., Cao, C. and Li, Z. (1998) Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector λ . *J. Chem. Inf. Comput. Sci.*, **38**, 387–394.
- Liu, S., Liu, H., Xia, Z., Cao, C. and Li, Z. (1999) Molecular distance-edge vector (μ): an extension from alkanes to alcohols. *J. Chem. Inf. Comput. Sci.*, **39**, 951–957.
- Liu, S., Liu, H., Yin, C.-S. and Wang, L.-S. (2003) VSMP: a novel variable selection and modeling method based on the prediction. *J. Chem. Inf. Comput. Sci.*, **43**, 964–969.
- Liu, S., Liu, H., Yu, B., Cao, C. and Li, S.Z. (2001) Investigation on quantitative relationship between chemical shift of carbon-13 nuclear magnetic resonance spectra and molecular topological structure based on a novel atomic distance-edge vector (ADEV). *J. Chemom.*, **15**, 427–438.
- Liu, S., Yin, C.-S., Cai, S.-X. and Li, Z. (2001a) A novel MHDV descriptor for dipeptide QSAR studies. *Journal of Chinese Chemical Society*, **48**, 253–260.
- Liu, S., Yin, C.-S., Cai, S.-X. and Li, Z. (2002a) Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons. *Chemom. Intell. Lab. Syst.*, **61**, 3–15.
- Liu, S., Yin, C.-S., Li, Z. and Cai, S.-X. (2001b) QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comput. Sci.*, **41**, 321–329.
- Liu, S., Yin, C.-S. and Wang, L.-S. (2002b) Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *J. Chem. Inf. Comput. Sci.*, **42**, 749–756.
- Liu, X. and Klein, D.J. (1991) The graph isomorphism problem. *J. Comput. Chem.*, **12**, 1243–1251.
- Liu, X., Wang, B., Huang, Z., Han, S. and Wang, L.-S. (2003) Acute toxicity and quantitative structure–activity relationships of α -branched phenylsulfonyl acetates to *Daphnia magna*. *Chemosphere*, **50**, 403–408.
- Liu, X., Yang, Z. and Wang, L.-S. (2003) Three-dimensional quantitative structure–activity relationship study for phenylsulfonyl carboxylates using CoMFA and CoMSIA. *Chemosphere*, **53**, 945–952.
- Liu, Y. and Brown, S.D. (2004) Wavelet multiscale regression from the perspective of data fusion: new conceptual approaches. *Anal. Bioanal. Chem.*, **380**, 445–452.
- Liu, Y., Guo, X., Xu, J., Pan, L. and Wang, S. (2002) Some notes on 2-D graphical representation of DNA sequence. *J. Chem. Inf. Comput. Sci.*, **42**, 529–533.
- Livingstone, D.J. (1996) *Data Analysis for Chemists: Applications for QSAR and Chemical Product Design*, Oxford University Press, New York, p. 239.
- Livingstone, D.J. (2000) The characterization of chemical structures using molecular properties a survey. *J. Chem. Inf. Comput. Sci.*, **40**, 195–209.
- Livingstone, D.J. (2003) Theoretical property predictions. *Curr. Top. Med. Chem.*, **3**, 1171–1192.
- Livingstone, D.J., Evans, D.A. and Saunders, M.R. (1992) Investigation of a charge transfer substituent constant using computational chemistry and pattern recognition techniques. *J. Chem. Soc. Perkin Trans. 2*, 1545–1550.
- Livingstone, D.J., Ford, M.G., Huuskonen, J.J. and Salt, D.W. (2001) Simultaneous prediction of aqueous solubility and octanol–water partition coefficient based on descriptors derived from molecular structure. *J. Comput. Aid. Mol. Des.*, **15**, 741–752.

- Livingstone, D.J., Hyde, R.M. and Foster, R. (1979) Further study of an organic electron-donor-acceptor related substituent constant. *Eur. J. Med. Chem.*, **14**, 393–397.
- Livingstone, D.J. and Manallack, D.T. (2003) Neural networks in 3D QSAR. *QSAR Comb. Sci.*, **22**, 510–518.
- Livingstone, D.J., Manallack, D.T. and Tetko, I.V. (1997) Data modelling with neural networks: advantages and limitations. *J. Comput. Aid. Mol. Des.*, **11**, 135–142.
- Livingstone, D.J. and Salt, D.W. (1992) Regression analysis for QSAR using neural networks. *Bioorg. Med. Chem. Lett.*, **2**, 213–218.
- Livingstone, D.J. and Salt, D.W. (2005) Judging the significance of multiple linear regression models. *J. Med. Chem.*, **48**, 661–663.
- Liwo, A., Tarnowska, M., Grzonka, Z. and Tempczyk, A. (1992) Modified Free–Wilson method for the analysis of biological activity data. *Computers Chem.*, **16**, 1–9.
- Llacer, M.T., Gálvez, J., García-Domenech, R., Gómez-Lechón, M.J., Más-Arcas, C. and De Julián-Ortiz, V. (2006) Topological virtual screening and pharmacological test of novel cytostatic drugs. *Internet Electron. J. Mol. Des.*, **5**, 306–319.
- Llorente, B., Rivero, N., Carrasco, R. and Martinez, R. S. (1994) A QSAR study of quinolones based on electrotopological state index for atoms. *Quant. Struct. -Act. Relat.*, **13**, 419–425.
- Lloyd, D. (1996) What is aromaticity? *J. Chem. Inf. Comput. Sci.*, **36**, 442–447.
- Lobato, M., Amat, L., Besalú, E. and Carbó-Dorca, R. (1997) Structure–activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indexes. *Quant. Struct. -Act. Relat.*, **16**, 465–472.
- Loew, G.H. and Burt, S.K. (1990) Quantum mechanics and the modeling of drug properties, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 105–123.
- Lohninger, H. (1993) Evaluation of neural networks based on radial basis functions and their application to the prediction of boiling points from structural parameters. *J. Chem. Inf. Comput. Sci.*, **33**, 736–744.
- Lohninger, H. (1994) Estimation of soil partition coefficients of pesticides from their chemical structure. *Chemosphere*, **29**, 1611–1626.
- Lombardo, F., Blake, J.F. and Curatolo, W.J. (1996) Computation of brain-blood partitioning of organic solutes via free energy calculations. *J. Med. Chem.*, **39**, 4750–4755.
- Lombardo, F., Gifford, E. and Shalaeva, M.Y. (2003) *In silico* ADME prediction: data, models facts and myths. *Mini Rev. Med. Chem.*, **3**, 861–875.
- López-Rodríguez, M.L., Murcia, M., Benhamú, B., Viso, A., Campillo, M. and Pardo, L. (2002) Benzimidazole derivatives. 3. 3D-QSAR/CoMFA model and computational simulation for the recognition of 5-HT₄ receptor antagonists. *J. Med. Chem.*, **45**, 4806–4815.
- Lorentz, H.A. (1880a) Über die Beziehung zwischen der Fortpflanzungsgeschwindigkeit des Lichtes der Körperdichte. *Wied. Ann. Phys.*, **9**, 641–665.
- Lorentz, L.V. (1880b) Über die Refraktionsconstante. *Wied. Ann. Phys.*, **11**, 70–103.
- Lounkine, E., Batista, J. and Bajorath, J. (2007) Mapping of activity-specific fragment pathways isolated from random fragment populations reveals the formation of coherent molecular cores. *J. Chem. Inf. Model.*, **47**, 2133–2139.
- Lovasz, L. and Pelikan, J. (1973) On the eigenvalue of trees. *Period Math Hung.*, **3**, 175–182.
- Löw, P. and Saller, H. (1988) PETRA: software package for the calculation of electronic and thermochemical properties of organic molecules, in *Physical Property Prediction in Organic Chemistry* (eds C. Jochum, M.G. Hicks and J. Sunkel), Springer-Verlag, Berlin, Germany, pp. 539–543.
- Löwdin, P.-Q. (1970) On the orthogonality problem. *Adv. Quant. Chem.*, **5**, 185–199.
- Lowe, J.P. (1978) *Quantum Chemistry*, Academic Press, New York, p. 600.
- Lowrey, A.H., Cramer, C.J., Urban, J.J. and Famini, G. R. (1995) Quantum chemical descriptors for linear solvation energy relationships. *Computers Chem.*, **19**, 209–215.
- Lowrey, A.H. and Famini, G.R. (1995) Using theoretical descriptors in quantitative structure–activity relationships HPLC capacity factors for energetic materials. *Struct. Chem.*, **6**, 357–365.
- Lozano, J.J., Pastor, M., Cruciani, G., Gaedt, K., Centeno, N.B., Gago, F. and Sanz, F. (2000) 3D-QSAR methods on the basis of ligand–receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands. *J. Comput. Aid. Mol. Des.*, **14**, 341–353.
- Lu, C., Guo, W., Hu, X., Wang, Y. and Yin, C.-S. (2006a) A Lu index for QSAR/QSPR studies. *Chem. Phys. Lett.*, **417**, 11–15.
- Lu, C., Guo, W., Hu, X., Wang, Y. and Yin, C.-S. (2006b) A novel Lu index to QSPR studies of aldehydes and ketones. *J. Math. Chem.*, **40**, 379–388.
- Lu, C., Guo, W., Wang, Y. and Yin, C.-S. (2006c) Novel distance-based atom-type topological indices DAI

- for QSPR/QSAR studies of alcohols. *J. Mol. Model.*, **12**, 749–756.
- Lu, C., Guo, W., and Yin, C.-S. (2006d) Quantitative structure–retention relationship study of the gas chromatographic retention indices of saturated esters on different stationary phases using novel topological indices. *Anal. Chim. Acta*, **561**, 96–102.
- Lu, C., Wang, Y., Yin, C.-S., Guo, W. and Hu, X. (2006) QSPR study on soil sorption coefficient for persistent organic pollutants. *Chemosphere*, **63**, 1384–1391.
- Lu, G.-H., Yuan, X. and Zhao, Y.-H. (2001) QSAR study on the toxicity of substituted benzenes to the algae (*Scenedesmus obliquus*). *Chemosphere*, **44**, 437–440.
- Lu, M., Liu, H. and Tian, F. (2004) The connectivity index. *MATCH Commun. Math. Comput. Chem.*, **51**, 149–154.
- Lu, W., Dong, N., Naray-Szabo, G. (2006) Predicting anti-HIV-1 activities of HEPT-analog compounds by using support vector classification. *QSAR Comb. Sci.*, **24**, 1021–1025.
- Luan, F., Ma, W., Zhang, X., Zhang, H., Liu, M., Hu, Z. and Fan, B.T. (2006) QSAR study of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls using the heuristic method and support vector machine. *QSAR Comb. Sci.*, **25**, 46–55.
- Luan, F., Zhang, R., Yao, X.-J., Liu, M. Hu, Z. and Fan, B.T. (2005a) Support vector machine-based QSPR for the prediction of van der Waals constants. *QSAR Comb. Sci.*, **24**, 227–239.
- Luan, F., Zhang, R., Zhao, C., Yao, X.-J., Liu, M., Hu, Z. and Fan, B.T. (2005b) Classification of the carcinogenicity of *N*-nitroso compounds based on support vector machines and linear discriminant analysis. *Chem. Res. Toxicol.*, **18**, 198–203.
- Lučić, B., Amić, D. and Trinajstić, N. (2000) Nonlinear multivariate regression outperforms several concisely designed neural networks on three QSPR data sets. *J. Chem. Inf. Comput. Sci.*, **40**, 403–413.
- Lučić, B., Lukovits, I., Nikolić, S. and Trinajstić, N. (2001) Distance-related indexes in the quantitative structure–property relationship modeling. *J. Chem. Inf. Comput. Sci.*, **41**, 527–535.
- Lučić, B., Miličević, A., Nikolić, S. and Trinajstić, N. (2002) Harary index – twelve years later. *Croat. Chem. Acta*, **75**, 847–868.
- Lučić, B., Miličević, A., Nikolić, S. and Trinajstić, N. (2003) On variable Wiener index. *Indian J. Chem.*, **42**, 1279–1282.
- Lučić, B., Nadramija, D., Bašić, I. and Trinajstić, N. (2003) Toward generating simpler QSAR models: nonlinear multivariate regression versus several neural network ensembles and some related methods. *J. Chem. Inf. Comput. Sci.*, **43**, 1094–1102.
- Lučić, B., Nikolić, S., Trinajstić, N. and Juretić, D. (1995a) The structure–property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.*, **35**, 532–538.
- Lučić, B., Nikolić, S., Trinajstić, N., Juretić, D. and Jurić, A. (1995b) A novel QSPR approach to physico-chemical properties of the α -amino acids. *Croat. Chem. Acta*, **68**, 435–450.
- Lučić, B., Nikolić, S., Trinajstić, N., Jurić, A. and Mihalić, Z. (1995c) A structure–property study of the solubility of aliphatic alcohols in water. *Croat. Chem. Acta*, **68**, 417–434.
- Lučić, B. and Trinajstić, N. (1997) New developments in QSPR/QSAR modeling based on topological indices. *SAR & QSAR Environ. Res.*, **7**, 45–62.
- Lučić, B. and Trinajstić, N. (1999) Multivariate regression outperforms several robust architectures of neural networks in QSAR modeling. *J. Chem. Inf. Comput. Sci.*, **39**, 121–132.
- Lučić, B., Trinajstić, N., Sild, S., Karelson, M. and Katritzky, A.R. (1999) A new efficient approach for variable selection based on multiregression: prediction of gas chromatographic retention times and response factors. *J. Chem. Inf. Comput. Sci.*, **39**, 610–621.
- Luco, J.M. (1999) Prediction of the brain–blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *J. Chem. Inf. Comput. Sci.*, **39**, 396–404.
- Luco, J.M. and Ferretti, H.F. (1997) QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.*, **37**, 392–401.
- Luco, J.M., Gálvez, J., García-Domenech, R. and De Julián-Ortiz, V. (2004) Structural invariants for the prediction of relative toxicities of polychloro dibenzo-*p*-dioxins and dibenzofurans. *Mol. Div.*, **8**, 331–342.
- Luco, J.M., Sosa, M.E., Cesco, J.C., Tonn, C.E. and Giordano, O.S. (1994) Molecular connectivity and hydrophobicity in the study of antifeedant activity of clerodane diterpenoids. *Pestic. Sci.*, **41**, 1–6.
- Luco, J.M., Yamin, L.J. and Ferretti, H.F. (1995) Molecular topology and quantum chemical descriptors in the study of reversed-phase liquid chromatography. Hydrogen-bonding behavior of chalcones and flavanones. *J. Pharm. Sci.*, **84**, 903–908.
- Luisi, P. (1977) Molecular conformational rigidity: an approach to quantification. *Naturwissenschaften*, **64**, 569–574.

- Luke, B.T. (1994) Evolutionary programming applied to the development of quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.*, **34**, 1279–1287.
- Luke, B.T. (1999) Comparison of three different QSAR/QSPR generation techniques. *J. Mol. Struct. (Theochem)*, **468**, 13–20.
- Lukovits, I. (1983) Quantitative structure–activity relationships employing independent quantum chemical indices. *J. Med. Chem.*, **26**, 1104–1109.
- Lukovits, I. (1988) Decomposition of the Wiener topological index. Application to drug–receptor interactions. *J. Chem. Soc. Perkin Trans. 2*, 1667–1671.
- Lukovits, I. (1990a) The generalized Wiener index for molecules containing double bonds and the partition coefficients. *Rep. Mol. Theory*, **1**, 127–131.
- Lukovits, I. (1990b) Wiener indices and partition coefficients of unsaturated hydrocarbons. *Quant. Struct. -Act. Relat.*, **9**, 227–231.
- Lukovits, I. (1991) General formulas for the Wiener index. *J. Chem. Inf. Comput. Sci.*, **31**, 503–507.
- Lukovits, I. (1992) Correlation between components of the Wiener index and partition coefficients of hydrocarbons. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **19**, 217–223.
- Lukovits, I. (1994) Formulas for the hyper-Wiener index of trees. *J. Chem. Inf. Comput. Sci.*, **34**, 1079–1081.
- Lukovits, I. (1995a) A formula for the hyper-Wiener index, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, Barcelona, Spain, pp. 53–54.
- Lukovits, I. (1995b) A note on a formula for the hyper-Wiener index of some trees. *Computers Chem.*, **19**, 27–31.
- Lukovits, I. (1995c) An algorithm for computation of bond contributions of the Wiener index. *Croat. Chem. Acta*, **68**, 99–103.
- Lukovits, I. (1996a) Indicators for atoms included in cycles. *J. Chem. Inf. Comput. Sci.*, **36**, 65–68.
- Lukovits, I. (1996b) The Detour index. *Croat. Chem. Acta*, **69**, 873–882.
- Lukovits, I. (1998a) An all-path version of the Wiener index. *J. Chem. Inf. Comput. Sci.*, **38**, 125–129.
- Lukovits, I. (1998b) Wiener index: formulas for non-homeomorphic graphs. *Croat. Chem. Acta*, **71**, 449–458.
- Lukovits, I. (1999) Isomer generation: syntactic rules for detection of isomorphism. *J. Chem. Inf. Comput. Sci.*, **39**, 563–568.
- Lukovits, I. (2000) A compact form of the adjacency matrix. *J. Chem. Inf. Comput. Sci.*, **40**, 1147–1150.
- Lukovits, I. (2001a) A theorem on graph valence shell. *MATCH Commun. Math. Comput. Chem.*, **44**, 279–286.
- Lukovits, I. (2001b) Wiener-type graph invariants, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 31–38.
- Lukovits, I. and Gutman, I. (1994) Edge-decomposition of the Wiener number. *MATCH Commun. Math. Comput. Chem.*, **31**, 133–144.
- Lukovits, I. and Linert, W. (1994) A novel definition of the hyper-Wiener index for cycles. *J. Chem. Inf. Comput. Sci.*, **34**, 899–902.
- Lukovits, I. and Linert, W. (1998) Polarity-numbers of cycle-containing structures. *J. Chem. Inf. Comput. Sci.*, **38**, 715–719.
- Lukovits, I. and Linert, W. (2001) A topological account of chirality. *J. Chem. Inf. Comput. Sci.*, **41**, 1517–1520.
- Lukovits, I. and Lopata, A. (1980) Decomposition of pharmacological activity indices into mutually independent components using principal component analysis. *J. Med. Chem.*, **23**, 449–459.
- Lukovits, I., Miličević, A., Nikolić, S. and Trinajstić, N. (2002) On walk counts and complexity of general graphs. *Internet Electron. J. Mol. Des.*, **1**, 388–400.
- Lukovits, I., Nikolić, S. and Trinajstić, N. (1999) Resistance distance in regular graphs. *Int. J. Quant. Chem.*, **71**, 217–225.
- Lukovits, I., Nikolić, S. and Trinajstić, N. (2000) Note on the resistance distances in the dodecahedron. *Croat. Chem. Acta*, **73**, 957–967.
- Lukovits, I., Nikolić, S. and Trinajstić, N. (2002) On relationships between vertex-degrees, path-numbers and graph valence-shells in trees. *Chem. Phys. Lett.*, **354**, 417–422.
- Lukovits, I., Palfi, K., Bako, I. and Kalman, E. (1997) LKP model of the inhibition mechanism of thiourea compounds. *Corrosion*, **53**, 915–919.
- Lukovits, I. and Razinger, M. (1997) On calculation of the Detour index. *J. Chem. Inf. Comput. Sci.*, **37**, 283–286.
- Lukovits, I. and Trinajstić, N. (2003) Atomic walk counts of negative order. *J. Chem. Inf. Comput. Sci.*, **43**, 1110–1114.
- Luo, J., Liao, B., Li, R. and Zhu, W. (2006) RNA secondary structure 3D graphical representation without degeneracy. *J. Math. Chem.*, **39**, 629–636.
- Luo, X., Stefanski, L.A. and Boos, D.D. (2006) Tuning variable selection procedures by adding noise. *Technometrics*, **48**, 165–175.

- Luo, Y.-R., Benson, S.W. (1990) New electronegativity scale. 11. Comparison with other scales in correlating heats of formation. *J. Phys. Chem.*, **94**, 914–917.
- Luo, Z., Wang, R. and Lai, L. (1996) RASSE: a new method for structure-based drug design. *J. Chem. Inf. Comput. Sci.*, **36**, 1187–1194.
- Luque Ruiz, I., Urbano-Cuadrado, M. and Gómez-Nieto, M.A. (2007) Data fusion of similarity and dissimilarity measurements using Wiener-based indices for the prediction of the NPY Y5 receptor antagonist capacity of benzoxazinones. *J. Chem. Inf. Model.*, **47**, 2235–2241.
- Luzanov, A.V. and Nerukh, D. (2007) Simple one-electron invariants of molecular chirality. *J. Math. Chem.*, **41**, 417–435.
- Lyde, D.R. (ed.) (2007) *Handbook of Chemistry & Physics*, CRC, Boca Raton, FL, p. 2640.
- Lyman, W.J., Reehl, W.F. and Rosenblatt, D.H. (1982) *Handbook of Chemical Property Estimation Methods*, McGraw-Hill, New York.
- Ma, W., Zhang, X., Luan, F., Zhang, H., Zhang, R., Liu, M., Hu, Z. and Fan, B.T. (2005) Support vector machine and the heuristic method to predict the solubility of hydrocarbons in electrolyte. *J. Phys. Chem. A*, **109**, 3485–3492.
- Mabilia, M., Pearlstein, R.A. and Hopfinger, A.J. (1985) Molecular shape analysis and energetics-based intermolecular modelling of benzylpyrimidine dihydrofolate reductase inhibitors. *Eur. J. Med. Chem.*, **20**, 163–174.
- MACCS keys, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA.
- Maciel, G.E. and Natterstad, J.J. (1965) Study of ^{13}C chemical shifts in substituted benzenes. *J. Chim. Phys.*, **42**, 2427–2435.
- Mackay, A.L. (1975) On rearranging the connectivity matrix of a graph – comments. *J. Chim. Phys.*, **62**, 308–309.
- MacPhee, J.A., Panaye, A. and Dubois, J.-E. (1978a) Operational definition of the Taft steric parameter. A homogeneous scale for alkyl groups – experimental extension to highly hindered groups. *Tetrahedron Lett.*, **34**, 3293–3296.
- MacPhee, J.A., Panaye, A. and Dubois, J.-E. (1978b) Steric effects. I. A critical examination of the Taft steric parameter – E_s . Definition of a revised, broader and homogeneous scale. Extension to highly congested alkyl groups. *Tetrahedron*, **34**, 3553–3562.
- MacroModels, Schrödinger, LLC, New York.
- Magee, P.S. (1990) A new approach to active-site binding analysis. Inhibitors of acetylcholinesterase. *Quant. Struct. -Act. Relat.*, **9**, 202–215.
- Magee, P.S. (1991) Positional analysis of binding events, in *QSAR: Rational Approaches to the Design of Bioactive Compounds* (eds C. Silipo and A. Vittoria), Elsevier, Amsterdam, The Netherlands, pp. 549–552.
- Magee, P.S. (1998) Some novel approaches to modeling transdermal penetration and reactivity with epidermal proteins, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 137–168.
- Magee, P.S. (2000a) Exploring the chemistry of quinones by computation. *Quant. Struct. -Act. Relat.*, **19**, 22–28.
- Magee, P.S. (2000b) Exploring the potential for allergic contact dermatitis via computed heats of reaction of haptens with protein end-groups. Heats of reaction of haptens with protein end-groups by computation. *Quant. Struct. -Act. Relat.*, **19**, 356–365.
- Mager, P.P. (1995a) A rigorous QSAR analysis. *J. Chemom.*, **9**, 232–236.
- Mager, P.P. (1995b) Diagnostics statistics in QSAR. *J. Chemom.*, **9**, 211–221.
- Mager, P.P. (1996) A random number experiment to simulate resample model evaluations. *J. Chemom.*, **10**, 221–240.
- Mager, P.P. (2003) Hybrid canonical-correlation neural-network approach applied to nonnucleoside HIV-1 reverse transcriptase inhibitors (HEPT derivatives). *Curr. Med. Chem.*, **10**, 1643–1659.
- Maggiora, G.M. (2006) On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model.*, **46**, 1535.
- Maggiora, G.M., Elrod, D.W. and Trenary, R.G. (1992) Computational neural networks as model free mapping devices. *J. Chem. Inf. Comput. Sci.*, **32**, 732–741.
- Maggiora, G.M. and Johnson, M.A. (1990) Introduction to similarity in chemistry, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G. M. Maggiora), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1–13.
- Magnuson, V.R., Harriss, D.K. and Basak, S.C. (1983) Topological indices based on neighborhood symmetry: chemical and biological applications, in *Studies in Physical and Theoretical Chemistry* (ed. R. B. King), Elsevier, Amsterdam, The Netherlands, pp. 178–191.
- Maier, B.J. (1992) Wiener and Randić topological indices for graphs. *J. Chem. Inf. Comput. Sci.*, **32**, 87–90.
- Maiocchi, A. (2003) The use of molecular descriptors in the design of gadolinium(III) chelates as MRI contrast agents. *Mini Rev. Med. Chem.*, **3**, 845–859.

- Maity, D.K. and Bhattacharyya, S.P. (1996) On the characterization of reaction paths by local descriptors: the behaviour of active BO sum and local information entropy along reaction paths and some consequences. *J. Mol. Struct. (Theochem)*, **367**, 59–66.
- Makara, G.M. (2001) Measuring molecular similarity and diversity: total pharmacophore diversity. *J. Med. Chem.*, **44**, 3563–3571.
- Makhija, M.T. and Kulkarni, A. (2002a) 3D-QSAR and molecular modeling of HIV-1 integrase inhibitors. *J. Comput. Aid. Mol. Des.*, **16**, 181–200.
- Makhija, M.T. and Kulkarni, V.M. (2001a) Eigen value analysis of HIV-1 integrase inhibitors. *J. Chem. Inf. Comput. Sci.*, **41**, 1569–1577.
- Makhija, M.T. and Kulkarni, V.M. (2001b) Molecular electrostatic potentials as input for the alignment of HIV-1 integrase inhibitors in 3D QSAR. *J. Comput. Aid. Mol. Des.*, **15**, 961–978.
- Makhija, M.T. and Kulkarni, V.M. (2002b) QSAR of HIV-1 integrase inhibitors by genetic function approximation method. *Bioorg. Med. Chem.*, **10**, 1483–1497.
- Makovskaya, V., Dean, J.R., Tomlinson, W.R. and Comber, M. (1995) Octanol–water partition coefficients of substituted phenols and their correlation with molecular descriptors. *Anal. Chim. Acta*, **315**, 193–200.
- Maldonado, A.G., Doucet, J.P., Petitjean, M. and Fan, B.T. (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol. Div.*, **10**, 39–79.
- Maldonado, A.G., Doucet, J.P., Petitjean, M. and Fan, B.T. (2007) MolDiA: a novel molecular diversity analysis tool. 1. Principles and architecture. *J. Chem. Inf. Model.*, **47**, 2197–2207.
- Mallion, R.B. (1975) Some graph-theoretical aspects of simple ring current calculations on conjugated systems. *Proc. Roy. Soc. London A*, **341**, 429–449.
- Mallion, R.B., Schwenk, A.J. and Trinajstić, N. (1974) A graphical study of heteroconjugated molecules. *Croat. Chem. Acta*, **46**, 171–182.
- Mallion, R.B. and Trinajstić, N. (2003) Reciprocal spanning-tree density: a new index for characterizing the intricacy of a (poly)cyclic molecular graph. *MATCH Commun. Math. Comput. Chem.*, **48**, 97–116.
- Malloves, C.L. (1973) Some comments on C_p . *Technometrics*, **15**, 661–675.
- Malone, J.G. (1933) The electronic moment as a measure of the ionic nature of covalent bonds. *J. Chim. Phys.*, **1**, 197–199.
- Malta, V.R.S., Pinto, A.V., Molfetta, F.A., Honório, K. M., de Simone, C.A., Pereira, M.A., Santos, R.H.A. and da Silva, A.B.F. (2003) The influence of electronic and steric effects in the structure–activity relationship (SAR) study of quinone compounds with biological activity against *Trypanosoma cruzi*. *J. Mol. Struct. (Theochem)*, **634**, 271–280.
- Manallack, D.T., Ellis, D.D. and Livingstone, D.J. (1994) Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.*, **37**, 3758–3767.
- Manallack, D.T. and Livingstone, D.J. (1993) The use of neural networks for data analysis in QSAR: chance effects, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 128–131.
- Manallack, D.T. and Livingstone, D.J. (1994) Limitations of functional link nets as applied to QSAR data analysis. *Quant. Struct. -Act. Relat.*, **13**, 18–21.
- Manallack, D.T. and Livingstone, D.J. (1995) Relating biological activity to chemical structure using neural networks. *Pestic. Sci.*, **45**, 167–170.
- Manallack, D.T., Tehan, B.G., Gancia, E., Hudson, B. D., Ford, M.G., Livingstone, D.J. and Pitt, W.R. (2003) A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.*, **43**, 674–679.
- Manaut, F., Sanz, F., José, J. and Milesi, M. (1991) Automatic search for maximum similarity between molecular electrostatic potential distributions. *J. Comput. Aid. Mol. Des.*, **5**, 371–380.
- Mandal, A., Johnson, K., Wu, C.F.J. and Bornemeier, D. (2007) Identifying promising compounds in drug discovery: genetic algorithms and some new statistical techniques. *J. Chem. Inf. Model.*, **47**, 981–988.
- Mandelbrot, B.B. (1982) *The Fractal Geometry of Nature*, Freeman, San Francisco, CA.
- Mandloi, M., Sikarwar, A., Sapre, N.S., Karmarkar, S. and Khadikar, P.V. (2000) A comparative QSAR study using Wiener, Szeged, and molecular connectivity indices. *J. Chem. Inf. Comput. Sci.*, **40**, 57–62.
- Mann, G. (1967) Conformation and physical data of alkanes and cycloalkanes. *Tetrahedron*, **23**, 3375–3392.
- Mannhold, R. (2003) Octanol/water partition coefficient, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1300–1313.
- Mannhold, R., Berellini, G., Carosati, E. and Benedetti, P. (2006) Use of MIF-based VolSurf descriptors in physico-chemical and pharmacokinetic studies, in *Molecular Interaction*

- Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 173–196.
- Mannhold, R., Cruciani, G., Dross, K. and Rekker, R. F. (1998) Multivariate analysis of experimental and computational descriptors of molecular lipophilicity. *J. Comput. Aid. Mol. Des.*, **12**, 573–581.
- Mannhold, R. and Dross, K. (1996) Calculation procedures for molecular lipophilicity: a comparative study. *Quant. Struct. -Act. Relat.*, **15**, 403–409.
- Mannhold, R. and Ostermann, C. (2008) Prediction of log *P* with substructure-based methods, in *Molecular Drug Properties*, Vol. 37 (ed. R. Mannhold), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 357–379.
- Mannhold, R., Rekker, R.F., Dross, K., Bijloo, G.J. and de Vries, G. (1998) The lipophilic behaviour of organic compounds. 1. An updating of the hydrophobic fragmental constant approach. *Quant. Struct. -Act. Relat.*, **17**, 517–536.
- Mannhold, R. and van de Waterbeemd, H. (2001) Substructure and whole molecule approaches for calculating log *P*. *J. Comput. Aid. Mol. Des.*, **15**, 337–354.
- Mansfield, M.L. and Covell, D.G. (2002) A new class of molecular shape descriptors. 1. Theory and properties. *J. Chem. Inf. Comput. Sci.*, **42**, 259–273.
- Maran, U., Karelson, M. and Katritzky, A.R. (1999) A comprehensive QSAR treatment of the genotoxicity of heteroaromatics and aromatic amines. *Quant. Struct. -Act. Relat.*, **18**, 3–10.
- Maranas, C.D. (1996) Optimal computer-aided molecular design: a polymer design case study. *Ind. Eng. Chem. Res.*, **35**, 3403–3414.
- Marchand-Geneste, N., Watson, K.A., Alsberg, B.K. and King, R.D. (2002) New approach to pharmacophore mapping and QSAR analysis using inductive logic programming. Application to thermolysin inhibitors and glycogen phosphorylase *b* inhibitors. *J. Med. Chem.*, **45**, 399–409.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1988) *Multivariate Analysis*, Academic Press, London, UK, p. 522.
- Marengo, E., Carpignano, R., Savarino, P. and Viscardi, G. (1992) Comparative study of different structural descriptors and variable selection approaches using partial least squares in quantitative structure–activity relationships. *Chemom. Intell. Lab. Syst.*, **14**, 225–233.
- Marengo, E., Leardi, R., Robotti, E., Righetti, P.G., Antonucci, F. and Cecconi, D. (2003) Application of three-way principal component analysis to the evaluation of two-dimensional maps in proteomics. *J. Proteome Res.*, **2**, 351–360.
- Marengo, E., Robotti, E., Bobba, M., Demartini, M. and Righetti, P.G. (2008) A new method of comparing 2D-PAGE maps based on the computation of Zernike moments and multivariate statistical tools. *Anal. Bioanal. Chem.*, **391**, 1163–1173.
- Marengo, E., Robotti, E., Bobba, M., Liparota, M.C., Rustichelli, C., Zamò, A., Chilosi, M. and Righetti, P.G. (2006) Multivariate statistical tools applied to the characterization of the proteomic profiles of two human lymphoma cell lines by two-dimensional gel electrophoresis. *Electrophoresis*, **27**, 484–494.
- Maria, P.-C., Gal, J.-F., de Franceschi, J. and Fargin, E. (1987) Chemometrics of the solvent basicity: multivariate analysis of the basicity scales relevant to nonprotogenic solvents. *J. Am. Chem. Soc.*, **109**, 483–492.
- Marialke, J., Körner, R., Tietze, S. and Apostolakis, J. (2007) Graph-based molecular alignment (GMA). *J. Chem. Inf. Model.*, **47**, 591–601.
- Marin, R.M., Aguirre, N.F. and Daza, E.E. (2008) Graph theoretical similarity approach to compare molecular electrostatic potentials. *J. Chem. Inf. Model.*, **48**, 109–118.
- Marini, F., Zupan, J. and Magri, A.L. (2005) Class-modeling using Kohonen artificial neural networks. *Anal. Chim. Acta*, **544**, 306–314.
- Mariussen, E., Andersson, P.L., Tysklind, M. and Fonnum, F. (2001) Effect of polychlorinated biphenyls on the uptake of dopamine into rat brain synaptic vesicles: a structure–activity study. *Toxicol. Appl. Pharm.*, **175**, 176–183.
- Marković, S. (1999) Tenth spectral moment for molecular graphs of phenylenes. *J. Chem. Inf. Comput. Sci.*, **39**, 654–658.
- Marković, S. (2003) Approximating the total pi-electron energy of phenylenes in terms of spectral moments. *Indian J. Chem.*, **42**, 1304–1308.
- Marković, S. and Gutman, I. (1991) Dependence of spectral moments of benzenoid hydrocarbons on molecular structure. *J. Mol. Struct. (Theochem)*, **235**, 81–87.
- Marković, S. and Gutman, I. (1999) Spectral moments of the edge adjacency matrix in molecular graphs. Benzenoid hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **39**, 289–293.
- Marković, S., Gutman, I. and Bancevic, Z. (1995) Correlation between Wiener and quasi-Wiener indices in benzenoid hydrocarbons. *J. Serb. Chem. Soc.*, **60**, 633–636.
- Marković, S., Marković, Z., Engelbrecht, J.P. and McCrindle, R.I. (2002) Spectral moments of polycyclic aromatic hydrocarbons. Solution of a

- kinetic problem. *J. Chem. Inf. Comput. Sci.*, **42**, 82–86.
- Marković, S., Marković, Z. and McCrindle, R.I. (2001) Spectral moments of phenylenes. *J. Chem. Inf. Comput. Sci.*, **41**, 112–119.
- Marković, S. and Stajkovic, A. (1997) The evaluation of spectral moments for molecular graphs of phenylenes. *Theor. Chim. Acta*, **96**, 256–260.
- Markowski, W., Dzido, T. and Wawrzynowicz, T. (1978) Correlation between chromatographic parameters and connectivity index in liquid–solid chromatography. *Pol. J. Chem.*, **52**, 2063.
- Marot, C., Chavatte, P. and Lesieur, D. (2000) Comparative molecular field analysis of selective cyclooxygenase-2 (COX-2) inhibitors. *Quant. Struct. -Act. Relat.*, **19**, 127–134.
- Marrero, J. and Gani, R. (2002) Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. *Ind. Eng. Chem. Res.*, **41**, 6623–6633.
- Marrero-Ponce, Y. (2003) Total and local quadratic indices of the molecular pseudograph's atom adjacency matrix: applications to the prediction of physical properties of organic compounds. *Molecules*, **8**, 687–726.
- Marrero-Ponce, Y. (2004a) Linear indices of the “molecular pseudograph's atom adjacency matrix”: definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *J. Chem. Inf. Comput. Sci.*, **44**, 2010–2026.
- Marrero-Ponce, Y. (2004b) Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. *Bioorg. Med. Chem.*, **12**, 6351–6369.
- Marrero-Ponce, Y., Cabrera Pérez, M.A., Zaldivar, V. R., Ofori, E. and Montero, L.A. (2003) Total and local quadratic indices of the “molecular pseudograph's atom adjacency matrix”. Application to prediction of Caco-2 permeability of drugs. *Int. J. Mol. Sci.*, **4**, 512–536.
- Marrero-Ponce, Y. and Castillo-Garit, J.A. (2005) 3D-chiral atom, atom-type, and total non-stochastic and stochastic molecular linear indices and their applications to central chirality codification. *J. Comput. Aid. Mol. Des.*, **19**, 369–383.
- Marrero-Ponce, Y., Castillo-Garit, J.A. and Nodarse, D. (2005a) Linear indices of the macromolecular graphs nucleotides adjacency matrix as a promising approach for bioinformatics studies. Part 1. Prediction of paromomycins affinity constant with HIV-1 W-RNA packaging region. *Bioorg. Med. Chem.*, **13**, 3397–3404.
- Marrero-Ponce, Y., Castillo-Garit, J.A., Olazabal, E., Serrano, H.S., Morales, A., Castañedo, N., Ibarra-Velarde, F., Huesca-Guillén, A., Jorge, E., del Valle, A., Torrens, F. and Catalá, A. (2004a) TOMOCOMD-CARDD, a novel approach for computer-aided ‘rational’ drug design. I. Theoretical and experimental assessment of a promising method for computational screening and *in silico* design of new anthelmintic compounds. *J. Comput. Aid. Mol. Des.*, **18**, 615–634.
- Marrero-Ponce, Y., Castillo-Garit, J.A., Olazabal, E., Serrano, H.S., Morales, A., Castañedo, N., Ibarra-Velarde, F., Huesca-Guillén, A., Sánchez, A.M., Torrens, F. and Castro, E.A. (2005b) Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic. *Bioorg. Med. Chem.*, **13**, 1005–1020.
- Marrero-Ponce, Y., Castillo-Garit, J.A., Torrens, F., Zaldivar, V.R. and Castro, E.A. (2004b) Atom, atom-type, and total linear indices of the “molecular pseudograph's atom adjacency matrix”: application to QSPR/QSAR studies of organic compounds. *Molecules*, **9**, 1100–1123.
- Marrero-Ponce, Y., González Díaz, H., Zaldivar, V.R., Torrens, F. and Castro, E.A. (2004) 3D-chiral quadratic indices of the ‘molecular pseudograph's atom adjacency matrix’ and their application to central chirality codification: classification of ACE inhibitors and prediction of σ -receptor antagonist activities. *Bioorg. Med. Chem.*, **12**, 5331–5342.
- Marrero-Ponce, Y., Huesca-Guillén, A. and Ibarra-Velarde, F. (2005) Quadratic indices of the ‘molecular pseudograph's atom adjacency matrix’ and their stochastic forms: a novel approach for virtual screening and *in silico* discovery of new lead paramphistomicide drugs-like compounds. *J. Mol. Struct. (Theochem)*, **717**, 67–79.
- Marrero-Ponce, Y., Khan, M.T.H., Casañola-Martín, G.M., Ather, A., Sultankhodzhaev, M.N., García-Domenech, R., Torrens, F. and Rotondo, R. (2007a) Bond-based 2D TOMOCOMD-CARDD approach for drug discovery: aiding decision-making in ‘*in silico*’ selection of new lead tyrosinase inhibitors. *J. Comput. Aid. Mol. Des.*, **21**, 167–188.
- Marrero-Ponce, Y., Khan, M.T.H., Casañola-Martín, G.M., Ather, A., Sultankhodzhaev, M.N., Torrens, F. and Rotondo, R. (2007b) Prediction of tyrosinase inhibition activity using atom-based bilinear indices. *ChemMedChem*, **2**, 449–478.
- Marrero-Ponce, Y., Marrero, R.M., Castro, E.A., de Armas, R.R., González Díaz, H., Zaldivar, V.R. and

- Torrens, F. (2004) Protein quadratic indices of the "macromolecular pseudograph's α -carbon atom adjacency matrix". 1. Prediction of arc repressor alanine-mutant's stability. *Molecules*, **9**, 1124–1147.
- Marrero-Ponce, Y., Marrero, R.M., Torrens, F., Martinez, Y., Bernal, M.L., Zaldivar, V.R., Castro, E. A. and Abalo, R.G. (2006) Non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix: a novel approach for computational *in silico* screening and "rational" selection of new lead antibacterial agents. *J. Mol. Model.*, **12**, 255–271.
- Marrero-Ponce, Y., Medina-Marrero, R., Torrens, F., Martinez, Y., Romero-Zaldivar, V. and Castro, E.A. (2005) Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. *Bioorg. Med. Chem.*, **13**, 2881–2899.
- Marrero-Ponce, Y., Montero-Torres, A., Zaldivar, C. R., Veitía, M.I., Mayón Pérez, M. and García Sánchez, R.N. (2005) Non-stochastic and stochastic linear indices of the molecular pseudographs atom adjacency matrix: application to *in silico* studies for the rational discovery of new antimalarial compounds. *Bioorg. Med. Chem.*, **13**, 1293–1304.
- Marrero-Ponce, Y., Torrens, F., Alvarado, Y.J. and Rotondo, R. (2006) Bond-based global and local (bond, group and bond-type) quadratic indices and their applications to computer-aided molecular design. 1. QSPR studies of diverse sets of organic chemicals. *J. Comput. Aid. Mol. Des.*, **20**, 685–701.
- Marriott, S., Reynolds, W.F., Taft, R.W. and Topsom, R.D. (1984) Substituent electronegativity parameters. *J. Org. Chem.*, **49**, 959–965.
- Marriott, S. and Topsom, R.D. (1982) Theoretical studies of the inductive effect. 3. A theoretical scale of field parameters. *Tetrahedron Lett.*, **23**, 1485–1488.
- Marshall, G.R. (1993) Binding-site modeling of unknown receptors, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 80–116.
- Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoehler, R.A. and Dunn, D.A. (1979) The conformational parameter in drug design: the active analog approach, in *Computer-Assisted Drug Design* (eds E.C. Olson and R.E. Christoffersen), American Chemical Society, Washington, DC, pp. 205–226.
- Marshall, G.R. and Cramer, R.D. III (1988) Three-dimensional structure–activity relationships. *Trends Pharmacol. Sci.*, **9**, 285–289.
- Marsili, M. (1988) Computation of volumes and surface areas of organic compounds, in *Physical Property Prediction in Organic Chemistry* (eds C. Jochum, M.G. Hicks and J. Sunkel), Springer-Verlag, Berlin, Germany, pp. 249–254.
- Marsili, M. (2003) Experimental design, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 423–445.
- Marsili, M. and Gasteiger, J. (1980) π charge distribution from molecular topology and π orbital electronegativity. *Croat. Chem. Acta*, **53**, 601–614.
- Marsili, M. and Saller, H. (1993) ANALOGS: a computer program for the design of multivariate sets of analog compounds. *J. Chem. Inf. Comput. Sci.*, **33**, 266–269.
- Martens, H. and Naes, T. (1989) *Multivariate Calibration*, John Wiley & Sons, Ltd, Chichester, UK.
- Martin, T.M. and Young, D.M. (2001) Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (*Pimephales promelas*) using a group contribution method. *Chem. Res. Toxicol.*, **14**, 1378–1385.
- Martin, Y.C. (1978) *Quantitative Drug Design. A Critical Introduction*, Marcel Dekker, New York, p. 425.
- Martin, Y.C. (1979) Advances in the methodology of quantitative drug design, in *Drug Design*, Vol. VIII (ed. E.J. Ariens), Academic Press, New York, pp. 1–72.
- Martin, Y.C. (1998) 3D QSAR: current state, scope, and limitations, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 3–23.
- Martin, Y.C. (2001) Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.*, **3**, 231–250.
- Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A. (1993) A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput. Aid. Mol. Des.*, **7**, 83–102.
- Martin, Y.C., Danaher, E.B., May, C.S. and Weininger, D. (1988) MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical, or geometric properties. *J. Comput. Aid. Mol. Des.*, **2**, 15–29.
- Martin, Y.C., Kofron, J.L. and Traphagen, L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.
- Martin, Y.C., Lin, C.T., Hetti, C. and DeLazzer, J. (1995) PLS analysis of distance matrices to detect

- nonlinear relationships between biological potency and molecular properties. *J. Med. Chem.*, **38**, 3009–3015.
- Martin, Y.C., Lin, C.T. and Wu, J. (1993) Application of CoMFA to D1 dopaminergic agonists: a case study, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 643–659.
- Martin, Y.C. and Lynn, K.R. (1971) Quantitative structure–activity relationships in leucomycin and lincomycin antibiotics. *J. Med. Chem.*, **14**, 1162–1166.
- Martín-Biosca, Y., Molero-Monfort, M., Sagrado, S., Villanueva-Camañas, R.M. and Medina-Hernández, M.J. (2000) Development of predictive retention–activity relationship models of barbiturates by micellar liquid chromatography. *Quant. Struct. -Act. Relat.*, **19**, 247–256.
- Martinek, T.A., Ötvös, F., Dervarics, M., Tóth, G. and Fülöp, F. (2005) Ligand-based prediction of active conformation by 3D-QSAR flexibility descriptors and their application in 3 + 3D-QSAR models. *J. Med. Chem.*, **48**, 3239–3250.
- Mason, H.S. (1943) History of the use of graphic formulas in organic chemistry. *Isis*, **34**, 346–354.
- Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C. and Labaudiniere, R.F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications. Including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.*, **42**, 3251–3264.
- Mason, J.S. and Pickett, S.D. (1997) Partition-based selection. *Persp. Drug Disc. Des.*, **7/8**, 85–114.
- Massart, D.L. and Kaufman, L. (1983) *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, John Wiley & Sons, Inc., New York.
- Massart, D.L., Vandeginste, B.G.M., Buydens, L., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. (1997) *Handbook of Chemometrics and Qualimetrics. Part A*, Elsevier, Amsterdam, The Netherlands, p. 868.
- Massart, D.L., Vandeginste, B.G.M., Buydens, L., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. (1998) *Handbook of Chemometrics and Qualimetrics. Part B*, Elsevier, Amsterdam, The Netherlands, p. 714.
- Mastryukova, T.A. and Kabachnik, M.I. (1971) Correlation constants in the chemistry of organophosphorus compounds. *J. Org. Chem.*, **36**, 1201–1205.
- Masuda, T., Nakamura, K., Jikihara, T., Kasuya, F., Igarashi, K., Fukui, M., Takagi, T. and Fujiwara, H. (1996) 3D quantitative structure–activity relationships for hydrophobic interactions. Comparative molecular field analysis (CoMFA) including molecular lipophilicity potentials as applied to the glycine conjugation of aromatic as well as aliphatic carboxylic acids. *Quant. Struct. -Act. Relat.*, **15**, 194–200.
- Matamala, A.R. and Estrada, E. (2005a) Generalised topological indices: optimisation methodology and physico-chemical interpretation. *Chem. Phys. Lett.*, **410**, 343–347.
- Matamala, A.R. and Estrada, E. (2005b) Simplex optimization of generalized topological index (GTI-simplex): a unified approach to optimize QSPR models. *J. Phys. Chem. A*, **109**, 9890–9895.
- Matito, E., Poater, J., Duran, M. and Solà, M. (2005) An analysis of the changes in aromaticity and planarity along the reaction path of the simplest Diels–Alder reaction. Exploring the validity of different indicators of aromaticity. *J. Mol. Struct. (Theochem)*, **727**, 165–171.
- Matsson, P., Bergström, C.A.S., Nagahara, N., Tavelin, S., Norinder, U. and Artursson, P. (2005) Exploring the role of different drug transport routes in permeability screening. *J. Med. Chem.*, **48**, 604–613.
- Matter, H. (1997) Selecting optimally diverse compounds from structure databases. A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.*, **40**, 1219–1229.
- Matter, H. and Pötter, T. (1999) Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.*, **39**, 1211–1225.
- Matthews, B.W. (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
- Mattioni, B.E. and Jurs, P.C. (2002) Development of quantitative structure–activity relationship and classification models for a set of carbonic anhydrase inhibitors. *J. Chem. Inf. Comput. Sci.*, **42**, 94–102.
- Mattioni, B.E. and Jurs, P.C. (2003) Prediction of dihydrofolate reductase inhibition and selectivity using computational neural networks and linear discriminant analysis. *J. Mol. Graph. Model.*, **21**, 391–419.
- Mattioni, B.E., Kauffman, G.W. and Jurs, P.C. (2003) Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble. *J. Chem. Inf. Comput. Sci.*, **43**, 949–963.
- Mauri, A., Ballabio, D., Consonni, V., Manganaro, A. and Todeschini, R. (2008) Peptides multivariate characterisation using a molecular descriptor

- based approach. *MATCH Commun. Math. Comput. Chem.*, **60**, 671–690.
- Mauri, A., Consonni, V., Pavan, M. and Todeschini, R. (2006) DRAGON software: an easy approach to molecular descriptor calculations. *MATCH Commun. Math. Comput. Chem.*, **56**, 237–248.
- Maw, H.H. and Hall, L.H. (2000) E-state modeling of dopamine transporter binding. Validation of the model for a small data set. *J. Chem. Inf. Comput. Sci.*, **40**, 1270–1275.
- Maw, H.H. and Hall, L.H. (2001) E-state modeling of corticosteroids binding affinity validation of model for small data set. *J. Chem. Inf. Comput. Sci.*, **41**, 1248–1254.
- Maw, H.H. and Hall, L.H. (2002) E-state modeling of HIV-1 protease inhibitor binding independent of 3D information. *J. Chem. Inf. Comput. Sci.*, **42**, 290–298.
- Maxwell, D.M. and Brecht, K.M. (1992) Quantitative structure–activity analysis of acetylcholinesterase inhibition by oxono and thiono analogues of organophosphorus compounds. *Chem. Res. Toxicol.*, **5**, 66–71.
- Mayer, A.Y., Farin, D. and Avair, D. (1986) Cross-sectional areas of alkanolic acids. A comparative study applying fractal theory of adsorption and considerations of molecular shape. *J. Am. Chem. Soc.*, **108**, 7897–7905.
- Mayer, D., Naylor, C.B., Motoc, I. and Marshall, G.R. (1987) A unique geometry of the active site of angiotensin-converting enzyme consistent with structure–activity studies. *J. Comput. Aid. Mol. Des.*, **1**, 3–16.
- Mayer, I. (1986a) Bond orders and valences from *ab initio* wave functions. *Int. J. Quant. Chem.*, **29**, 477–483.
- Mayer, I. (1986b) On bond orders and valences in the *ab initio* quantum chemical theory. *Int. J. Quant. Chem.*, **29**, 73–84.
- Mayer, I. (2007) Bond order and valence indices: a personal account. *J. Comput. Chem.*, **28**, 204–221.
- Mayer, J.M., van de Waterbeemd, H. and Testa, B. (1982) A comparison between the hydrophobic fragmental methods of Rekker and Leo. *Eur. J. Med. Chem.*, **17**, 17–25.
- Mazza, C.B., Sukumar, N., Breneman, C.M. and Cramer, S.M. (2001) Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal. Chem.*, **73**, 5457–5461.
- Mazzatorta, P., Benfenati, E., Neagu, C.-D. and Gini, G. (2003) Tuning neural and fuzzy-neural networks for toxicity modeling. *J. Chem. Inf. Comput. Sci.*, **43**, 513–518.
- Mazzatorta, P., Benfenati, E., Neagu, D. and Gini, G. (2002) The importance of scaling in data mining for toxicity prediction. *J. Chem. Inf. Comput. Sci.*, **42**, 1250–1255.
- Mazzatorta, P., Smiesko, M., Lo Liparo, E. and Benfenati, E. (2005) QSAR model for predicting pesticide aquatic toxicity. *J. Chem. Inf. Model.*, **45**, 1767–1774.
- Mazzatorta, P., Vračko, M. and Benfenati, E. (2003a) ANVAS: artificial neural variables adaptation system for descriptor selection. *J. Comput. Aid. Mol. Des.*, **17**, 335–346.
- Mazzatorta, P., Vračko, M., Jezierska, A. and Benfenati, E. (2003b) Modeling toxicity by using supervised Kohonen neural networks. *J. Chem. Inf. Comput. Sci.*, **43**, 485–492.
- McCabe, G.P. (1975) Computations for variable selection in discriminant analysis. *Technometrics*, **17**, 103–109.
- McCabe, G.P. (1984) Principal variables. *Technometrics*, **26**, 137–144.
- McClellan, A.L. (1963) *Tables of Experimental Dipole Moments*, Freeman, San Francisco, CA.
- McClelland, B.J. (1971) Properties of the latent roots of a matrix: the estimation of π -electron energies. *J. Chim. Phys.*, **54**, 640–643.
- McClelland, B.J. (1974) Graphical method for factorizing secular determinants of Hückel molecular orbital theory. *J. Chem. Soc. Faraday Trans II*, **70**, 1453–1456.
- McClelland, B.J. (1982) Eigenvalues of the topological matrix. Splitting of graphs with symmetrical components and alternant graphs. *J. Chem. Soc. Faraday Trans II*, **78**, 911–916.
- McClure, W.F., Hamid, A., Giesbricht, F.G. and Weeks, W.W. (1984) Fourier analysis enhances NIR diffuse reflectance spectroscopy. *Applied Spectroscopy*, **38**, 322–329.
- McCoy, E.F. and Sykes, M.J. (2003) Quantum-mechanical QSAR/QSPR descriptors from momentum-space wave functions. *J. Chem. Inf. Comput. Sci.*, **43**, 545–553.
- McDaniel, D.H. and Brown, H.C. (1955) A quantitative approach to the *ortho* effects of halogen substituents in aromatic systems. *J. Am. Chem. Soc.*, **77**, 3756–3763.
- McDaniel, D.H. and Brown, H.C. (1958) An extended table of Hammett substituent constants based on the ionization of substituted benzoic acids. *J. Org. Chem.*, **23**, 420–427.
- McDaniel, D.H. and Yingst, A. (1964) The use of basicity and oxidative coupling potential to obtain group electronegativity. *J. Am. Chem. Soc.*, **86**, 1334–1336.

- McElroy, N.R. and Jurs, P.C. (2001) Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **41**, 1237–1247.
- McElroy, N.R. and Jurs, P.C. (2003) QSAR and classification of murine and human soluble epoxide hydrolase inhibition by urea-like compounds. *J. Med. Chem.*, **46**, 1066–1080.
- McFarland, J.W. (1970) On the parabolic relationship between drug potency and hydrophobicity. *J. Med. Chem.*, **13**, 1192–1196.
- McFarland, J.W., Avdeef, A., Berger, C.M. and Raevsky, O.A. (2001) Estimating the water solubilities of crystalline compounds from their chemical structures alone. *J. Chem. Inf. Comput. Sci.*, **41**, 1355–1359.
- McFarland, J.W. and Gans, D.J. (1986) On the significance of clusters in the graphical display of structure–activity data. *J. Med. Chem.*, **29**, 505–514.
- McFarland, J.W. and Gans, D.J. (1990a) Cluster significance analysis: a new QSAR tool for asymmetric data sets. *Drug Inf. J.*, **24**, 705–711.
- McFarland, J.W. and Gans, D.J. (1990b) Linear discriminant analysis and cluster significance analysis, in *Quantitative Drug Design*, Vol. 4 (ed. C. A. Ramsden), Pergamon Press, Oxford, UK, pp. 667–689.
- McFarland, J.W. and Gans, D.J. (1994) On identifying likely determinants of biological activity in high dimensional QSAR problems. *Quant. Struct. -Act. Relat.*, **13**, 11–17.
- McFarland, J.W. and Gans, D.J. (1995) Multivariate data analysis of chemical and biological data cluster significance analysis, in *Chemometrics Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 295–308.
- McGregor, M.J. and Muskal, S.M. (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.*, **39**, 569–574.
- McGregor, M.J. and Muskal, S.M. (2000) Pharmacophore fingerprinting. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci.*, **40**, 117–125.
- McGregor, M.J. and Pallai, P.V. (1997) Clustering of large databases of compounds: using the MDL “keys” as structural descriptors. *J. Chem. Inf. Comput. Sci.*, **37**, 443–448.
- McGregor, T.R. (1979) Connectivity parameters as predictors of retention in gas chromatography. *J. Chromatogr. Sci.*, **17**, 314.
- McHughes, M.C. and Poshusta, R. (1990) Graph-theoretic cluster expansions. Thermochemical properties of alkanes. *J. Math. Chem.*, **4**, 227–249.
- McKay, B.D. (1977) On the spectral characterization of trees. *Ars Comb.*, **3**, 219–232.
- McKinney, J.D., Darden, T., Lyerly, M.A. and Pederson, L.G. (1985) PCB and related compound binding to the Ah receptor(s). Theoretical model based on molecular parameters and molecular mechanics. *Quant. Struct. -Act. Relat.*, **4**, 166–172.
- McKinney, J.D., Richard, A., Waller, C., Newman, M. C. and Gerberick, F. (2000) The practice of structure–activity relationships (SAR) in toxicology. *Toxicol. Sci.*, **56**, 8–17.
- McKone, T.E. (1993) The precision of QSAR methods for estimating intermedia transfer factors in exposure assessments. *SAR & QSAR Environ. Res.*, **1**, 41–51.
- McLay, I., Hann, M., Carosati, E., Cruciani, G. and Baroni, M. (2006) The complexity of molecular interaction: molecular shape fingerprints by the PathFinder approach, in *Molecular Interaction Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany.
- MDC – Molecular Descriptor Correlations, Ver. 1.0, Milano Chemometrics & QSAR Research Group, Univ. Milano-Bicocca, P.za. della Scienza 1, Milano, Italy, http://michem.disat.unimib.it/chm/download/molecular_correlationinfo.htm.
- Medeleanu, M. and Balaban, A.T. (1998) Real-number vertex invariants and Schultz-type indices based on eigenvectors of adjacency and distance matrices. *J. Chem. Inf. Comput. Sci.*, **38**, 1038–1047.
- Medić-Šarić, M., Mornar, A. and Jasprica, I. (2004) Lipophilicity study of salicylamide. *Acta Pharmacol.*, **54**, 91–101.
- Medina-Franco, J.L., Golbraikh, A., Oloff, S., Castillo, R. and Tropsha, A. (2005) Quantitative structure–activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the *k* nearest neighbor method and QSAR-based database mining. *J. Comput. Aid. Mol. Des.*, **19**, 229–242.
- Medina-Franco, J.L., Rodríguez-Morales, S., Juárez-Gordiano, C., Hernández-Campos, A. and Castillo, R. (2004) Docking-based CoMFA and CoMSIA studies of non-nucleoside reverse transcriptase inhibitors of the pyridinone derivative type. *J. Comput. Aid. Mol. Des.*, **18**, 345–360.
- Medven, Z., Güsten, H. and Sabljic, A. (1996) Comparative QSAR study on hydroxyl radical reactivity with unsaturated hydrocarbons: PLS versus MLR. *J. Chemom.*, **10**, 135–147.
- Mei, H., Liao, Z., H., Zhou, Y. and Li, S.Z. (2005) A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers (Peptide Science)* **80**, 775–786.

- Meiler, J., Sanli, E., Junker, J., Meusinger, R., Lindel, T., Will, M., Maier, W. and Köck, M. (2002) Validation of structural proposals by substructure analysis and ^{13}C NMR chemical shift prediction. *J. Chem. Inf. Comput. Sci.*, **42**, 241–248.
- Mekenyan, O., Balaban, A.T. and Bonchev, D. (1985) Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). VII. Condensed benzenoid hydrocarbons and their ^1H NMR chemical shifts. *J. Magn. Reson.*, **63**, 1–13.
- Mekenyan, O. and Basak, S.C. (1994) Topological indices and chemical reactivity, in *Graph Theoretic Approaches to Chemical Reactivity* (eds D. Bonchev and O. Mekenyan), Kluwer Academic, Dordrecht, The Netherlands, pp. 221–239.
- Mekenyan, O. and Bonchev, D. (1986) OASIS method for predicting biological activity of chemical compounds. *Acta Pharm. Jugosl.*, **36**, 225–237.
- Mekenyan, O., Bonchev, D. and Balaban, A.T. (1984a) Hierarchically ordered extended connectivities. Reflection in the ^1H NMR chemical shifts of condensed benzenoid hydrocarbons. *Chem. Phys. Lett.*, **109**, 85–88.
- Mekenyan, O., Bonchev, D. and Balaban, A.T. (1984b) Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). V. New topological indices, ordering of graphs, and recognition of graph similarity. *J. Comput. Chem.*, **5**, 629–639.
- Mekenyan, O., Bonchev, D. and Balaban, A.T. (1985) Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). II. Mathematical proofs for the HOC algorithm. *J. Comput. Chem.*, **6**, 552–561.
- Mekenyan, O., Bonchev, D. and Balaban, A.T. (1988a) Topological indices for molecular fragments and new graph invariants. *J. Math. Chem.*, **2**, 347–375.
- Mekenyan, O., Bonchev, D. and Enchev, V. (1988b) Modeling the interaction of small organic molecules with biomacromolecules (the Oasis approach). V. Toxicity of phenols to algae “*Lemna minor*”. *Quant. Struct. -Act. Relat.*, **7**, 240–244.
- Mekenyan, O., Bonchev, D., Sabljic, A. and Trinajstić, N. (1987) Applications of topological indices to QSAR. The use of the Balaban index and the electropy index for correlations with toxicity of ethers on mice. *Acta Pharm. Jugosl.*, **37**, 75–86.
- Mekenyan, O., Bonchev, D. and Trinajstić, N. (1980) Chemical graph theory: modeling the thermodynamic properties of molecules. *Int. J. Quant. Chem.*, **28**, 369–380.
- Mekenyan, O., Bonchev, D. and Trinajstić, N. (1981) Algebraic characterization of bridged polycyclic compounds. *Int. J. Quant. Chem.*, **19**, 929–955.
- Mekenyan, O., Bonchev, D. and Trinajstić, N. (1983) Structural complexity and molecular properties of cyclic systems with acyclic branches. *Croat. Chem. Acta*, **56**, 237–261.
- Mekenyan, O., Bonchev, D., Trinajstić, N. and Peitchev, D. (1986) Modelling the interaction of small organic molecules with biomacromolecules. II. A generalized concept for biological interactions. *Arzneim. Forsch. (German)*, **36**, 421–424.
- Mekenyan, O., Dimitrov, S. and Bonchev, D. (1963) Graph-theoretical approach to the calculation of physico-chemical properties of polymers. *Eur. Polym. J.*, **19**, 1185–1193.
- Mekenyan, O., Ivanov, J., Veith, G.D. and Bradbury, S. P. (1994) Dynamic QSAR: a new search for active conformations and significant stereoelectronic indices. *Quant. Struct. -Act. Relat.*, **13**, 302–307.
- Mekenyan, O., Karabunarliev, S. and Bonchev, D. (1990a) The Microcomputer OASIS system for predicting the biological activity of chemical compounds. *Computers Chem.*, **14**, 193–200.
- Mekenyan, O., Karabunarliev, S. and Bonchev, D. (1990b) The OASIS concept for predicting biological activity of chemical compounds. *J. Math. Chem.*, **4**, 207–215.
- Mekenyan, O., Mercier, C., Bonchev, D. and Dubois, J.-E. (1993) Comparative study of DARC/PELCO and OASIS methods. 2. Modeling PNMT inhibitory potency of benzylamines and amphetamines. *Eur. J. Med. Chem.*, **28**, 811–819.
- Mekenyan, O., Nikolova, N., Karabunarliev, S., Bradbury, S.P., Ankley, G.T. and Hansen, B. (1999) New development in a hazard identification algorithm for hormone receptor ligands. *Quant. Struct. -Act. Relat.*, **18**, 139–153.
- Mekenyan, O., Nikolova, N. and Schmieder, P. (2003) Dynamic 3D QSAR techniques: applications in toxicology. *J. Mol. Struct. (Theochem)*, **622**, 147–165.
- Mekenyan, O., Peitchev, D., Bonchev, D., Trinajstić, N. and Bangov, I.P. (1986a) Modelling the interaction of small organic molecules with biomacromolecules. I. Interaction of substituted pyridines with anti-3-azopyridine antibody. *Arzneim. Forsch. (German)*, **36**, 176–183.
- Mekenyan, O., Peitchev, D., Bonchev, D., Trinajstić, N. and Dimitrova, J. (1986b) Modelling the interaction of small organic molecules with biomacromolecules. III. Interaction of benzoates with anti-*p*-(*p*′-axophenylazo)-benzoate antibody. *Arzneim. Forsch. (German)*, **36**, 629–635.

- Mekenyan, O., Roberts, D.W. and Karcher, W. (1997) Molecular orbital parameters as predictors of skin sensitization potential of halo- and pseudohalobenzenes acting as S_NAr electrophiles. *Chem. Res. Toxicol.*, **10**, 994–1000.
- Mekenyan, O. and Veith, G.D. (1993) Relationships between descriptors for hydrophobicity and soft electrophilicity in predicting toxicity. *SAR & QSAR Environ. Res.*, **1**, 335–344.
- Mekenyan, O. and Veith, G.D. (1994) The electronic factor in QSAR: MO-parameters, competing interactions, reactivity and toxicity. *SAR & QSAR Environ. Res.*, **2**, 129–143.
- Mekenyan, O. and Veith, G.D. (1997) 3D molecular design: searching for active conformers in QSAR, in *From Chemical Topology to Three-Dimensional Geometry* (ed. A.T. Balaban), Plenum Press, New York, pp. 43–71.
- Mekenyan, O., Veith, G.D., Bradbury, S.P. and Russom, C.L. (1993) Structure–toxicity relationships for α,β -unsaturated alcohols in fish. *Quant. Struct. -Act. Relat.*, **12**, 132–136.
- Mekenyan, O., Veith, G.D., Call, D.J. and Ankley, G.T. (1996) A QSAR evaluation of Ah receptor binding of halogenated aromatic xenobiotics. *Environ. Health Persp.*, **104**, 1302–1310.
- Melkova, Z. (1984) Utilization of the index of molecular connectivity in the study of antitumor activity of a group of benzo[c]fluorene derivatives. *Ceskoslov. Farm.*, **33**, 107–111.
- Melnikov, A.A., Palyulin, V.A. and Zefirov, N.S. (2007) Generation of molecular graphs for QSAR studies: an approach based on supergraphs. *J. Chem. Inf. Model.*, **47**, 2077–2088.
- Meloun, M., Militky, J. and Forina, M. (1994) *Chemometrics for Analytical Chemistry*, Ellis Horwood, Bodmin, UK, p. 400.
- Melssen, W., Üstün, B. and Buydens, L. (2007) SOMPLS: a supervised self-organising map–partial least squares algorithm for multivariate regression problems. *Chemom. Intell. Lab. Syst.*, **86**, 102–120.
- Melville, J.L. and Hirts, J.D. (2007) TMAcc: interpretable correlation descriptors for quantitative structure–activity relationships. *J. Chem. Inf. Model.*, **47**, 626–634.
- Menard, P.R., Lewis, R.A. and Mason, J.S. (1998) Rational screening set design and compound selection: cascaded clustering. *J. Chem. Inf. Comput. Sci.*, **38**, 497–505.
- Menard, P.R., Mason, J.S., Morize, I. and Bauerschmidt, S. (1998) Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Comput. Sci.*, **38**, 1204–1213.
- Mendiratta, S. and Madan, A.K. (1994) Structure–activity study on antiviral 5-vinylpyrimidine nucleoside analogs using Wiener's topological index. *J. Chem. Inf. Comput. Sci.*, **34**, 867–871.
- Meneses, L., Tiznado, W., Contreras, R. and Fuentealba, P. (2004) A proposal for a new local hardness as selectivity index. *Chem. Phys. Lett.*, **383**, 181–187.
- Menezes, F.A.S., Montanari, C.A. and Bruns, R.E. (2000) 3D-WHIM pattern recognition study for bisamidines. A structure–property relationship study. *J. Braz. Chem. Soc.*, **11**, 393–397.
- Menezes, I.R.A., Lopes, J.C.D., Montanari, C.A., Oliva, G., Pavão, F., Vieira, P.C. and Pupo, M.T. (2003) 3D QSAR studies on binding affinities of coumarin natural products for glycosomal GAPDH of *Trypanosoma cruzi*. *J. Comput. Aid. Mol. Des.*, **17**, 277–290.
- Menziani, M.C. and De Benedetti, P.G. (1992) Molecular mechanics and quantum chemical QSAR analysis in carbonic anhydrase heterocyclic sulfonamide interactions. *Struct. Chem.*, **3**, 215–219.
- Mercader, A., Castro, E.A. and Toropov, A.A. (2000) QSPR modeling of the enthalpy of formation from elements by means of correlation weighting of local invariants of atomic orbital molecular graphs. *Chem. Phys. Lett.*, **330**, 612–623.
- Mercader, A., Castro, E.A. and Toropov, A.A. (2001) Maximum topological distances based indices as molecular descriptors for QSPR. 4. Modeling the enthalpy of formation of hydrocarbons from elements. *Int. J. Mol. Sci.*, **2**, 121–132.
- Mercier, C. and Dubois, J.-E. (1979) Comparison of molecular connectivity and DARC/PELCO Methods: performance in antimicrobial, halogenated phenol QSARs. *Eur. J. Med. Chem.*, **14**, 415–423.
- Mercier, C., Mekenyan, O., Dubois, J.-E. and Bonchev, D. (1991) DARC/PELCO and OASIS methods. I. Methodological comparison. Modeling purine pK_a and antitumor activity. *Eur. J. Med. Chem.*, **26**, 575–592.
- Mercier, C., Troullier, G. and Dubois, J.-E. (1990) DARC computer aided design in anticholinergic research. *Quant. Struct. -Act. Relat.*, **9**, 88–93.
- Merkwirth, C., Mauser, H., Schulz-Gasch, T., Roche, O., Stahl, M. and Lengauer, T. (2004) Ensemble methods for classification in chemoinformatics. *J. Chem. Inf. Comput. Sci.*, **44**, 1971–1978.
- Merrifield, R.E. and Simmons, H.E. (1980) The structures of molecular topological spaces. *Theor. Chim. Acta*, **55**, 55–75.
- Merrifield, R.E. and Simmons, H.E. (1998) *Topological Methods in Chemistry*, John Wiley & Sons, Inc., New York, p. 233.

- Merschmundermann, V., Rosenkranz, H.S. and Klopman, G. (1994) The structural basis of the genotoxicity of nitroarenefurans and related compounds. *Mut. Res.*, **304**, 271–284.
- Mestres, J. and Scuseria, G.E. (1995) Genetic algorithms: a robust scheme for geometry optimizations and global minimum structure problems. *J. Comput. Chem.*, **16**, 729–742.
- Meurice, N., Leherter, L. and Vercauteren, D.P. (1998) Comparison of benzodiazepine-like compounds using topological analysis and genetic algorithms. *SAR & QSAR Environ. Res.*, **8**, 195–232.
- Mewes, H.-W. (2003) Sequence and genome bioinformatics, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1812–1844.
- Meyer, A.M. and Richards, W.G. (1991) Similarity of molecular shape. *J. Comput. Aid. Mol. Des.*, **5**, 426–439.
- Meyer, A.Y. (1985a) Molecular mechanics and molecular shape. Part 1. van der Waals descriptors of simple molecules. *J. Chem. Soc. Perkin Trans. 2*, 1161–1169.
- Meyer, A.Y. (1985b) Molecular mechanics and molecular shape. Part II. Beyond the van der Waals descriptors of shape. *J. Mol. Struct. (Theochem)*, **124**, 93–106.
- Meyer, A.Y. (1986a) Molecular mechanics and molecular shape. III. Surface area and cross-sectional areas of organic molecules. *J. Comput. Chem.*, **7**, 144–152.
- Meyer, A.Y. (1986b) Molecular mechanics and molecular shape. Part 4. Size, shape, and steric parameters. *J. Chem. Soc. Perkin Trans. 2*, 1567–1572.
- Meyer, A.Y. (1986c) The size of molecules. *Chem. Soc. Rev.*, **15**, 449–474.
- Meyer, A.Y. (1988a) Molecular mechanics and molecular shape. Part VI. The response of simple molecules to bimolecular association. *J. Mol. Struct. (Theochem)*, **179**, 83–98.
- Meyer, A.Y. (1988b) Molecular mechanics and molecular shape. V. On the computation of the bare surface area of molecules. *J. Comput. Chem.*, **9**, 18–24.
- Meyer, A.Y. (1989) Molecular mechanics and molecular shape. Part VII. Structural factors in the estimation of solvation energies. *J. Mol. Struct. (Theochem)*, **195**, 147–158.
- Meyer, H. (1899) Zur Theorie der Alkoholnarkose. *Arch. Exp. Pathol. Pharmacol.*, **42**, 109–118.
- Meylan, W.M. and Howard, P.H. (1995) Atom/fragment contribution method for estimating octanol–water partition coefficients. *J. Pharm. Sci.*, **84**, 83–92.
- Meylan, W.M. and Howard, P.H. (1996) Improved method for estimating water solubility from octanol/water partition coefficient. *Environ. Toxicol. Chem.*, **15**, 100–106.
- Meylan, W.M. and Howard, P.H. (2000) Estimating log *P* with atom/fragments and water solubility with log *P*. *Persp. Drug Disc. Des.*, **19**, 67–84.
- Meylan, W.M., Howard, P.H. and Boethling, R.S. (1992) Molecular topology/fragment contribution method for predicting soil sorption coefficients. *Environ. Sci. Technol.*, **26**, 1560–1567.
- Mezey, P.G. (1985) Group theory of electrostatic potentials: a tool for quantum chemical drug design. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **12**, 113–122.
- Mezey, P.G. (1987a) Group theory of shapes of asymmetric biomolecules. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **14**, 127–132.
- Mezey, P.G. (1987b) The shape of molecular charge distributions: group theory without symmetry. *J. Comput. Chem.*, **8**, 462–469.
- Mezey, P.G. (1988a) Global and local relative convexity and oriented relative convexity: application to molecular shapes in external fields. *J. Math. Chem.*, **2**, 325.
- Mezey, P.G. (1988b) Graphical shapes: seeing graphs of chemical curves and molecular surfaces. *J. Math. Chem.*, **2**, 377.
- Mezey, P.G. (1988c) Shape group studies of molecular similarity: shape groups and shape graphs of molecular contour surfaces. *J. Math. Chem.*, **2**, 299.
- Mezey, P.G. (1989) The topology of molecular surfaces and shape graphs, in *Computational Chemical Graph Theory and Combinatorics* (ed. D.H. Rouvray), Nova Publications, New York.
- Mezey, P.G. (1990a) A global approach to molecular symmetry: theorems on symmetry relations between ground- and excited-state configurations. *J. Am. Chem. Soc.*, **112**, 3791–3802.
- Mezey, P.G. (1990b) Molecular point symmetry and the phase of the electronic wave function: tools for the prediction of critical points of potential energy surfaces. *Int. J. Quant. Chem.*, **38**, 699–711.
- Mezey, P.G. (1990c) Three-dimensional topological aspects of molecular similarity, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 321–368.
- Mezey, P.G. (ed.) (1991a) *Mathematical Modeling in Chemistry*, Wiley-VCH Verlag GmbH, Weinheim, Germany.
- Mezey, P.G. (1991b) Molecular surfaces, in *Reviews in Computational Chemistry*, Vol. 11 (eds K.B.

- Lipkowitz and D. Boyd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 265–294.
- Mezey, P.G. (1991c) The degree of similarity of three-dimensional bodies: application to molecular shape analysis. *J. Math. Chem.*, **7**, 39–49.
- Mezey, P.G. (1992) Shape-similarity measures for molecular bodies: a three-dimensional topological approach to quantitative shape–activity relations. *J. Chem. Inf. Comput. Sci.*, **32**, 650–656.
- Mezey, P.G. (1993a) Dynamic shape analysis of molecules in restricted domains of a configuration space. *J. Math. Chem.*, **13**, 59–70.
- Mezey, P.G. (1993b) New rules on potential surface topology and critical point search. *J. Math. Chem.*, **14**, 79–90.
- Mezey, P.G. (1993c) *Shape in Chemistry: An Introduction to Molecular Shape and Topology*, VCH Publishers, New York.
- Mezey, P.G. (1993d) Topological shape analysis of chain molecules: an application of the GSTE principle. *J. Math. Chem.*, **12**, 365–374.
- Mezey, P.G. (1994) Iterated similarity sequences and shape ID numbers for molecules. *J. Chem. Inf. Comput. Sci.*, **34**, 244–247.
- Mezey, P.G. (1996) Theorems on molecular shape-similarity descriptors: external T-plasters and interior T-aggregates. *J. Chem. Inf. Comput. Sci.*, **36**, 1076–1081.
- Mezey, P.G. (1997a) Descriptors of molecular shape in 3D, in *From Chemical Topology to Three-Dimensional Geometry* (ed. A.T. Balaban), Plenum Press, New York, pp. 25–42.
- Mezey, P.G. (1997b) Fuzzy measures of molecular shape and size, in *Fuzzy Logic in Chemistry* (ed. D. H. Rouvray), Academic Press, New York, pp. 139–223.
- Mezey, P.G. (1999) Holographic electron density shape theorem and its role in drug design and toxicological risk assessment. *J. Chem. Inf. Comput. Sci.*, **39**, 224–230.
- Mghazli, S., Jaouad, A., Mansour, M., Villemin, D. and Cherqaoui, D. (2001) Neural networks studies: quantitative structure–activity relationships in antifungal 1-[2-(substituted phenyl)allyl] imidazoles and related compounds. *Chemosphere*, **43**, 385–390.
- Michotte, Y. and Massart, D.L. (1977) Molecular connectivity and retention indexes. *J. Pharm. Sci.*, **66**, 1630–1632.
- Miertus, S., Scrocco, E. and Tomasi, J. (1981) Electrostatic interaction of a solute with a continuum. A direct utilization of *ab initio* molecular potentials for the prevision of solvent effects. *Chem. Phys.*, **55**, 117–129.
- Migliavacca, E. (2003) Applied introduction to multivariate methods used in drug discovery. *Mini Rev. Med. Chem.*, **3**, 831–843.
- Migliavacca, E., Ancerewicz, J., Carrupt, P.-A. and Testa, B. (1998) Theoretical parameters to characterize antioxidants. Part 2. The case of melatonin and carvedilol. *Helv. Chim. Acta*, **81**, 1337–1348.
- Migliavacca, E., Carrupt, P.-A. and Testa, B. (1997) Theoretical parameters to characterize antioxidants. Part 1. The case of vitamin E and analogs. *Helv. Chim. Acta*, **80**, 1613–1626.
- Mihalić, Z., Nikolić, S. and Trinajstić, N. (1992) Comparative study of molecular descriptors derived from the distance matrix. *J. Chem. Inf. Comput. Sci.*, **32**, 28–37.
- Mihalić, Z. and Trinajstić, N. (1991) The algebraic modelling of chemical structures: on the development of three-dimensional molecular descriptors. *J. Mol. Struct. (Theochem)*, **232**, 65–78.
- Mihalić, Z. and Trinajstić, N. (1992) A graph-theoretical approach to structure–property relationships. *J. Chem. Educ.*, **69**, 701–712.
- Mihalić, Z., Veljan, D., Amić, D., Nikolić, S., Plavšić, D. and Trinajstić, N. (1992) The distance matrix in chemistry. *J. Math. Chem.*, **11**, 223–258.
- Miletti, F., Storch, L., Sforza, G. and Cruciani, G. (2007) New and original pK_a prediction method using grid molecular interaction fields. *J. Chem. Inf. Model.*, **47**, 2172–2181.
- Miličević, A. and Nikolić, S. (2004) On variable Zagreb indices. *Croat. Chem. Acta*, **77**, 97–101.
- Miličević, A., Nikolić, S., Plavšić, D. and Trinajstić, N. (2003) On the Hosoya Z index of general graphs. *Internet Electron. J. Mol. Des.*, **2**, 160–178.
- Miličević, A., Nikolić, S. and Trinajstić, N. (2004) On reformulated Zagreb indices. *Mol. Div.*, **8**, 393–399.
- Miličević, A. and Raos, N. (2006) Estimation of stability of coordination compounds by using topological indices. *Polyhedron*, **25**, 2800–2808.
- Miller, A.J. (1990a) *Subset Selection in Regression*, Chapman & Hall, London, UK, p. 230.
- Miller, D.W. (2001) Results of a new classification algorithm combining K nearest neighbors and recursive partitioning. *J. Chem. Inf. Comput. Sci.*, **41**, 168–175.
- Miller, D.W. (2003) A chemical class-based approach to predictive model generation. *J. Chem. Inf. Comput. Sci.*, **43**, 568–578.
- Miller, K.J. (1990b) Additivity methods in molecular polarizability. *J. Am. Chem. Soc.*, **112**, 8533–8542.
- Miller, K.J. (1990c) Calculation of the molecular polarizability tensor. *J. Am. Chem. Soc.*, **112**, 8543–8551.

- Miller, K.J. and Savchik, J.A. (1979) A new empirical method to calculate average molecular polarizabilities. *J. Am. Chem. Soc.*, **101**, 7206–7213.
- Millership, J.S. and Woolfson, A.D. (1978) The relation between molecular connectivity and gas chromatographic retention data. *J. Pharm. Pharmacol.*, **30**, 483–485.
- Millership, J.S. and Woolfson, A.D. (1979) A study of the relationship between gas chromatographic retention parameters and molecular connectivity. *J. Pharm. Pharmacol.*, **31**, 44.
- Millership, J.S. and Woolfson, A.D. (1980) Molecular connectivity and gas chromatographic retention parameters. *J. Pharm. Pharmacol.*, **32**, 610–614.
- Mills, E.J. (1884) On melting point and boiling point as related to composition. *Philosophical Magazine*, **17**, 173–187.
- Milne, G.W.A. (1997) Mathematics as a basis for chemistry. *J. Chem. Inf. Comput. Sci.*, **37**, 639–644.
- Minailiuc, O.M. and Diudea, M.V. (2001) TI-MTD model. Applications in molecular design, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science Publishers, Huntington, NY, pp. 363–388.
- Minailiuc, O.M., Katona, G., Diudea, M.V., Strunje, M., Graovac, A. and Gutman, I. (1998) Szeged fragmental indices. *Croat. Chem. Acta*, **71**, 473–488.
- Minkin, V.I., Glukhovtsev, M.N. and Simkin, B.Ya. (1994) *Aromaticity and Antiaromaticity. Electronic and Structural Aspects*, John Wiley & Sons, Inc., New York.
- Minoli, D. (1976) Teorie Combinatorie (Combinatorial graph complexity). *Atti Accad. Naz. Lincei – Rend. (Italian)*, **59**, 651–661.
- Mintz, C. and Acree, W.E., Jr (2007) Comments on ‘an improved characteristic molecular volume parameter for linear solvation energy relationships of acyclic alkanes’. *J. Phys. Org. Chem.*, **20**, 365–367.
- Mishra, R.K. (2001) Getting discriminant functions of antibacterial activity from physico-chemical and topological parameters. *J. Chem. Inf. Comput. Sci.*, **41**, 387–393.
- Mishra, R.K. and Patra, S.M. (1998) Numerical determination of the Kekulé structure count of some symmetrical polycyclic aromatic hydrocarbons and their relationship with π -electronic energy (a computational approach). *J. Chem. Inf. Comput. Sci.*, **38**, 113–124.
- Mislow, K. (1997) Fuzzy restrictions and inherent uncertainties in chirality studies, in *Fuzzy Logic in Chemistry* (ed. D.H. Rouvray), Academic Press, New York, pp. 65–90.
- Mitchell, B.E. and Jurs, P.C. (1997) Prediction of autoignition temperatures of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **37**, 538–547.
- Mitchell, B.E. and Jurs, P.C. (1998a) Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **38**, 489–496.
- Mitchell, B.E. and Jurs, P.C. (1998b) Prediction of infinite dilution activity coefficients of organic compounds in aqueous solution from molecular structure. *J. Chem. Inf. Comput. Sci.*, **38**, 200–209.
- Mitchell, J.B.O., Alex, A. and Snarey, M. (1999) SATIS: atom typing from chemical connectivity. *J. Chem. Inf. Comput. Sci.*, **39**, 751–757.
- Miyashita, Y., Li, Z.L. and Sasaki, S. (1993) Chemical pattern recognition and multivariate analysis for QSAR studies. *TRAC*, **12**, 50–60.
- Miyashita, Y., Ohsako, H., Takayama, C. and Sasaki, S. (1992) Multivariate structure–activity relationships analysis of fungicidal and herbicidal thiocarbamates using partial least squares method. *Quant. Struct. -Act. Relat.*, **11**, 17–22.
- Miyashita, Y., Okuyama, T., Ohsako, H. and Sasaki, S. (1989) Graph theoretical approach to carbon-13 chemical shift sum in alkanes. *J. Am. Chem. Soc.*, **111**, 3469–3470.
- Mlinsek, G., Novič, M., Hodoscek, M. and Solmajer, T. (2001) Prediction of enzyme binding: human thrombin inhibition study by quantum chemical and artificial intelligence methods based on X-ray structures. *J. Chem. Inf. Comput. Sci.*, **41**, 1286–1294.
- MobyDigs (Model by Descriptors in Genetic Selection), Ver. 1.0, Taletè s.r.l., via V. Pisani, 13 – 20124 Milan, Italy.
- MOE – Molecular Operating Environment, Chemical Computing Group, Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada.
- Mohar, B. (1989) Laplacian matrices of graphs, in *MATH/CHEM/COMP 1988* (ed. A. Graovac), Elsevier, Amsterdam, The Netherlands, pp. 1–8.
- Mohar, B. (1989a) Laplacian matrices of graphs. *Stud. Phys. Theor. Chem.*, **63**, 1–8.
- Mohar, B. (1991a) Eigenvalues, diameter, and mean distance in graphs. *Graphs Comb.*, **7**, 53–64.
- Mohar, B. (1991b) The Laplacian spectrum of graphs, in *Graph Theory, Combinatorics, and Applications* (eds Y. Alavi, C. Chartrand and O.R. Ollermann), John Wiley & Sons, Inc., New York, pp. 871–898.
- Mohar, B., Babic, D. and Trinajstić, N. (1993) A novel definition of the Wiener index for trees. *J. Chem. Inf. Comput. Sci.*, **33**, 153–154.
- Mohar, B. and Pisanski, T. (1988) How to compute the Wiener index of a graph. *J. Math. Chem.*, **2**, 267–277.

- Mokrosz, J.L. (1989) Topological indices in correlation analysis. Part 1. Comparison of molecular shape with molecular connectivity for some hydrocarbons. *Quant. Struct. -Act. Relat.*, **8**, 305–309.
- Molchanova, M.S. and Zefirov, N.S. (1998) Irredundant generation of isomeric molecular structures with some known fragments. *J. Chem. Inf. Comput. Sci.*, **38**, 8–22.
- MolConn-Z: A Program for Molecular Topology Analysis, Ver. 3, Hall Associates Consulting, Quincy, MA.
- Moliner, R., García, F., Gálvez, J., García-Domenech, R. and Serrano, C. (1991) Nuevos Índices Topológicos en Conectividad Molecular. Su Aplicación a Algunas Propiedades Fisicoquímicas de un Grupo de Hidrocarburos Alifáticos. *An. R. Acad. Farm.*, **57**, 287–298.
- Molnar, S.P. and King, J.W. (1995) Structure– pK_a correlation via the integrated molecular transform. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **22**, 201–206.
- Molnar, S.P. and King, J.W. (1998) Parametric transform and moment indices in the molecular dynamics of *n*-alkanes. *Int. J. Quant. Chem.*, **70**, 1185–1194.
- Molpro Quantum Chemistry Package, Werner, H.-J. and Knowles, P.J., University of Sussex, Brighton, UK.
- Monev, V. (2004) Introduction to similarity searching in chemistry. *MATCH Commun. Math. Comput. Chem.*, **51**, 7–38.
- Monge, A., Arrault, A., Marot, C. and Morin-Allory, L. (2006) Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Div.*, **10**, 389–403.
- Montanari, C.A., Cass, Q.B., Tiritan, M.E. and Souza, A.L.S.d. (2000) A QSERR study on enantioselective separation of enantiomeric sulfoxides. *Anal. Chim. Acta*, **419**, 93–100.
- Montanari, C.A., Tute, M.S., Beezer, A.E. and Mitchell, J.C. (1996) Determination of receptor-bound drug conformations by QSAR using flexible fitting to derive a molecular similarity index. *J. Comput. Aid. Mol. Des.*, **10**, 67–73.
- Montero-Torres, A., García Sánchez, R.N., Marrero-Ponce, Y., Machado-Tugores, Y., Nogal-Ruiz, J.J., Martínez-Fernández, A.R., Aran, V.J., Ochoa, C., Meneses-Marcel, A. and Torrens, F. (2006) Non-stochastic quadratic fingerprints and LDA-based QSAR models in hit and lead generation through virtual screening: theoretical and experimental assessment of a promising method for the discovery of new antimalarial compounds. *Eur. J. Med. Chem.*, **41**, 483–493.
- Montero-Torres, A., Vega, M.C., Marrero-Ponce, Y., Rolon, M., Gomez-Barrio, A., Escario, J.A., Aran, V. J., Martínez-Fernández, A.R. and Meneses-Marcel, A. (2005) A novel non-stochastic quadratic fingerprints-based approach for the ‘*in silico*’ discovery of new antitrypanosomal compounds. *Bioorg. Med. Chem.*, **13**, 6264–6275.
- Monti, E., Gariboldi, M., Maiocchi, A., Marengo, E., Cassino, C., Gabano, E. and Osella, D. (2005) Cytotoxicity of *cis*-platinum(II) conjugate models. The effect of chelating arms and leaving groups on cytotoxicity: a quantitative structure–activity relationship approach. *J. Med. Chem.*, **48**, 857–866.
- Moody, M.L., Willauer, H.D., Griffin, S.T., Huddleston, J.G. and Rogers, R.D. (2005) Solvent property characterization of poly(ethylene glycol)/dextran aqueous biphasic systems using the free energy of transfer of a methylene group and a linear solvation energy relationship. *Ind. Eng. Chem. Res.*, **44**, 3749–3760.
- Moon, T., Chi, M.H., Kim, D.-H., Yoon, C.N. and Choi, Y.-S. (2000) Quantitative structure–activity relationships (QSAR) study of flavonoid derivatives for inhibition of cytochrome P450 1A2. *Quant. Struct. -Act. Relat.*, **19**, 257–263.
- Moon, T., Song, J.-S., Lee, J.K. and Yoon, C.N. (2003) QSAR analysis of SH2-binding phosphopeptides: using interaction energies and cross-correlation coefficients. *J. Chem. Inf. Comput. Sci.*, **43**, 1570–1575.
- Moorthy, N.S.H.N., Karthikeyan, C. and Trivedi, P. (2007) QSAR studies of cytotoxic acridine 5,7-diones: a comparative study using P-VSA descriptors and topological descriptors. *Indian J. Chem.*, **46**, 177–184.
- MOPAC 6, Air Force Academy, Colorado Spring, CO.
- Morais, H., Ramos, C., Forgács, E., Jakab, A., Cserhádi, T., Oliveira, J., Illés, T. and Illés, Z. (2001) Three-dimensional principal component analysis used for the study of enzyme kinetics. An empirical approximation for the determination of the dimensions of component matrices. *Quant. Struct. -Act. Relat.*, **20**, 241–247.
- Morales, D.A. and Araujo, O. (1993) On the search for the best correlation between graph theoretical invariants and physico-chemical properties. *J. Math. Chem.*, **13**, 95–106.
- Moran, P.A.P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.
- Moreau, G. (1997) Atomic chirality, a quantitative measure of the chirality of the environment of an atom. *J. Chem. Inf. Comput. Sci.*, **37**, 929–938.

- Moreau, G. and Broto, P. (1980a) Autocorrelation of molecular structures. Application to SAR studies. *Nouv. J. Chim.*, **4**, 757–764.
- Moreau, G. and Broto, P. (1980b) The autocorrelation of a topological structure: a new molecular descriptor. *Nouv. J. Chim.*, **4**, 359–360.
- Moreau, G. and Turpin, C. (1996) Use of similarity analysis to reduce large molecular libraries to smaller sets of representative molecules. *Analisis*, **24**, M17–M21.
- Morell, C., Grand, A. and Toro-Labbé, A. (2005) New dual descriptor for chemical reactivity. *J. Phys. Chem. A*, **109**, 205–212.
- Morgan, H.L. (1965) The generation of a unique machine description for chemical structures – a technique developed at Chemical Abstracts Service. *J. Chem. Doc.*, **5**, 107–113.
- Moriguchi, I. (1975) Quantitative structure–activity studies. I. Parameters relating to hydrophobicity. *Chem. Pharm. Bull.*, **23**, 247–257.
- Moriguchi, I., Hirono, S., Liu, Q. and Nakagome, I. (1992a) Fuzzy adaptive least squares and its application to structure–activity studies. *Quant. Struct. -Act. Relat.*, **11**, 325–331.
- Moriguchi, I., Hirono, S., Liu, Q., Nakagome, I. and Matsushita, Y. (1992b) Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.*, **40**, 127–130.
- Moriguchi, I., Hirono, S., Nakagome, I. and Hirano, H. (1994) Comparison of reliability of log *P* values for drugs calculated by several methods. *Chem. Pharm. Bull.*, **42**, 976–978.
- Moriguchi, I. and Kanada, Y. (1977) Quantitative structure–activity studies. Part III. Use of van der Waals volume in structure–activity studies. *Chem. Pharm. Bull.*, **25**, 926–935.
- Moriguchi, I., Kanada, Y. and Komatsu, K. (1976) van der Waals volume and the related parameters for hydrophobicity in structure–activity studies. *Chem. Pharm. Bull.*, **24**, 1799–1806.
- Morikawa, T. and Balaban, A.T. (1992) Topological formulas and upper/lower bounds in chemical polygonal graphs, particularly in benzenoid polyhexes. *MATCH Commun. Math. Comput. Chem.*, **28**, 235–247.
- Morón, J.A., Campillo, M., Perez, V., Unzeta, M. and Pardo, L. (2000) Molecular determinants of MAO selectivity in a series of indolylmethylamine derivatives: biological activities, 3D-QSAR/CoMFA analysis, and computational simulation of ligand recognition. *J. Med. Chem.*, **43**, 1684–1691.
- Morovitz, H. (1955) Some order–disorder considerations in living systems. *Bull. Math. Biophys.*, **17**, 81–86.
- Mosier, P.D. and Jurs, P.C. (2002) QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. *J. Chem. Inf. Comput. Sci.*, **42**, 1460–1470.
- Mosier, P.D., Jurs, P.C., Custer, L.L., Durham, S.K. and Pearl, G.M. (2003) Predicting the genotoxicity of thiophene derivatives from molecular structure. *Chem. Res. Toxicol.*, **16**, 721–732.
- Mössner, S.G., Lopez de Alda, M.J., Sander, L.C., Lee, M.L. and Wise, S.A. (1999) Gas chromatographic retention behavior of polycyclic aromatic sulfur heterocyclic compounds (dibenzothiophene, naphtho[*b*]thiophenes, benzo[*b*]naphthothiophenes and alkyl-substituted derivatives) on stationary phases of different selectivity. *J. Chromat.*, **841**, 207–228.
- Motoc, I. (1983a) Molecular shape descriptors, in *Steric Effects in Drug Design, Topics in Current Chemistry*, Vol. 114 (eds M. Charton and I. Motoc), Springer-Verlag, Berlin, Germany, pp. 93–105.
- Motoc, I. (1983b) Quantitative comparison of the shape of bioorganic molecules. *Z. Naturforsch.*, **38a**, 1342–1345.
- Motoc, I. (1984a) Biological receptor maps. 2. Steric maps of benzoate antibody and carbonic anhydrase. *Quant. Struct. -Act. Relat.*, **3**, 47–51.
- Motoc, I. (1984b) Biological receptor maps. I. Steric maps. The SIBIS method. *Quant. Struct. -Act. Relat.*, **3**, 43–47.
- Motoc, I. and Balaban, A.T. (1981) Topological indices: intercorrelations, physical meaning, correlational ability. *Rev. Roum. Chim.*, **26**, 593–600.
- Motoc, I. and Balaban, A.T. (1982) Testing the geometrical meaning of Taft-type steric constants. *Rev. Roum. Chim.*, **27**, 735–739.
- Motoc, I., Balaban, A.T., Mekenyan, O. and Bonchev, D. (1982) Topological indices: inter-relations and composition. *MATCH Commun. Math. Comput. Chem.*, **13**, 369–404.
- Motoc, I. and Dragomir, O. (1981) Molecular interactions in biological systems. Steric interactions. The SIBIS algorithm. *Math. Chem.*, **12**, 117–126.
- Motoc, I., Holban, S., Vancea, R. and Simon, Z. (1977) Minimal steric difference calculated as nonoverlapping volumes. Correlations with enzymatic hydrolyses of ribonucleosides. *Studia Biophys.*, **66**, 75–78.
- Motoc, I. and Marshall, G.R. (1985) van der Waals volume fragmental constants. *Chem. Phys. Lett.*, **116**, 415–419.
- Mount, J., Ruppert, J., Welch, W. and Jain, A.N. (1999) *IcePick*: a flexible surface-based system for molecular diversity. *J. Med. Chem.*, **42**, 60–66.

- Mouvier, G. and Dubois, J.-E. (1968) N° 224 – Réactivité des Composés Éthyléniques: Réaction de Bromation. XVIII. Applications des Relations Linéaires D'énergie Libre au Cas des Alcènes. *Bull. Soc. Chim. Fran. (French)*, **4**, 1441–1445.
- Mowshowitz, A. (1968a) Entropy and the complexity of graphs. I. An index of the relative complexity of a graph. *Bull. Math. Biophys.*, **30**, 175–204.
- Mowshowitz, A. (1968b) Entropy and the complexity of graphs. II. The information content of digraphs and infinite graphs. *Bull. Math. Biophys.*, **30**, 225–240.
- Mowshowitz, A. (1968c) Entropy and the complexity of graphs. III. Graphs with prescribed information content. *Bull. Math. Biophys.*, **30**, 387–414.
- Mowshowitz, A. (1968d) Entropy and the complexity of graphs. IV. Entropy measures and graphical structure. *Bull. Math. Biophys.*, **30**, 533–546.
- Mpamhanga, C.P., Chen, B., McLay, I.M., Ormsby, D. L. and Lindvall, M.K. (2005) Retrospective docking study of PDE4B ligands and an analysis of the behavior of selected scoring functions. *J. Chem. Inf. Model.*, **45**, 1061–1074.
- Mpamhanga, C.P., Chen, B., McLay, I.M. and Willett, P. (2006) Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.*, **46**, 686–698.
- Mracec, M., Juchel, L. and Mracec, M. (2006) QSAR analysis of a series of imidazole derivatives acting on the H₃ receptor. *Rev. Roum. Chim.*, **51**, 287–292.
- Mracec, M., Muresan, S., Simon, Z. and Naray-Szabo, G. (1997) QSARs with orthogonal descriptors on psychotomimetic phenylalkylamines. *Quant. Struct.-Act. Relat.*, **16**, 459–464.
- Mracec, M., Mracec, M., Kurunczi, L., Nusser, T., Simon, Z., Náray-Szabó, G. (1996) QSAR study with steric (MTD), electronic and hydrophobicity parameters on psychotomimetic phenylalkylamines. *J. Mol. Struct. (Theochem)*, **367**, 139–149.
- Mu, L. and Feng, C. (2004) Novel connectivity index of edge valence and its applications. *Journal of Chemical Industrial Engineering (China)* **55**, 531–540.
- Mu, L. and Feng, C. (2007) Quantitative structure–property relations (QSPRs) for predicting standard absolute entropy, S^0_{298} , of inorganic compounds. *MATCH Commun. Math. Comput. Chem.*, **57**, 111–134.
- Mu, L., Feng, C. and Xu, L. (2006) Study on QSPR of alcohols with a novel edge connectivity index ^mF . *MATCH Commun. Math. Comput. Chem.*, **56**, 217–230.
- Muegge, I. (2002) Pharmacophore features of potential drugs. *Chem. Eur. J.*, **8**, 1977–1981.
- Muegge, I. (2003) Selection criteria for drug-like compounds. *Med. Res. Rev.*, **23**, 302–321.
- Muegge, I., Heald, S.L. and Brittelli, D. (2001) Simple selection criteria for drug-like chemical matter. *J. Med. Chem.*, **44**, 1841–1846.
- Muijselaar, P.G.H., Claessens, H.A. and Cramers, C. A. (1994) Application of the retention index concept in micellar electrokinetic capillary chromatography. *Anal. Chem.*, **66**, 635–644.
- Mukherjee, S., Mukherjee, A. and Saha, A. (2005) QSAR modeling on binding affinity of diverse estrogenic flavonoids: electronic, topological and spatial functions in quantitative approximation. *J. Mol. Struct. (Theochem)*, **715**, 85–90.
- Mullay, J. (1984) Atomic and group electronegativities. *J. Am. Chem. Soc.*, **106**, 5842–5847.
- Mullay, J. (1985) Calculation of group electronegativity. *J. Am. Chem. Soc.*, **107**, 7271–7275.
- Müller, K. (1997a) On the paradigm shift from rational to random design. *J. Mol. Struct. (Theochem)*, **398–399**, 467–471.
- Müller, M. (1997b) Quantum chemical modelling of soil sorption coefficients: multiple linear regression models. *Chemosphere*, **35**, 365–377.
- Müller, M. and Klein, W. (1991) Estimating atmospheric degradation processes by SARs. *Sci. Total Environ.*, **109/110**, 261–273.
- Müller, M. and Kördel, W. (1996) Comparison of screening methods for the estimation of adsorption coefficients on soil. *Chemosphere*, **32**, 2493–2504.
- Müller, W.R., Szymanski, K., von Knop, J., Mihalić, Z. and Trinajstić, N. (1993) The walk ID number revisited. *J. Chem. Inf. Comput. Sci.*, **33**, 231–233.
- Müller, W.R., Szymanski, K., von Knop, J., Mihalić, Z. and Trinajstić, N. (1995) Note on isocodal graphs. *J. Chem. Inf. Comput. Sci.*, **35**, 871–873.
- Müller, W.R., Szymanski, K., von Knop, J., Nikolić, S. and Trinajstić, N. (1990a) On the enumeration and generation of polyhex hydrocarbons. *J. Comput. Chem.*, **11**, 223–235.
- Müller, W.R., Szymanski, K., von Knop, J. and Trinajstić, N. (1987) An algorithm for construction of the molecular distance matrix. *J. Comput. Chem.*, **8**, 170–173.
- Müller, W.R., Szymanski, K., von Knop, J. and Trinajstić, N. (1990b) Molecular topological index. *J. Chem. Inf. Comput. Sci.*, **30**, 160–163.
- Mulliken, R.S. (1928a) The assignment of quantum numbers for electrons in molecules. I. *Phys. Rev.*, **32**, 186–222.
- Mulliken, R.S. (1928b) The assignment of quantum numbers for electrons in molecules. II. Correlation

- of molecular and atomic electron states. *Phys. Rev.*, **32**, 761–772.
- Mulliken, R.S. (1934) A new electroaffinity scale, together with data on valence states and an ionization potential and electron affinities. *J. Chim. Phys.*, **2**, 782–793.
- Mulliken, R.S. (1935a) Electronic structures of molecules. XI. Electroaffinity, molecular orbitals and dipole moments. *J. Chim. Phys.*, **3**, 573–585.
- Mulliken, R.S. (1935b) Electronic structures of molecules. X. Aldehydes, ketones and related molecules. *J. Chim. Phys.*, **3**, 564–573.
- Mulliken, R.S. (1935c) Electronic structures of molecules. XII. Electroaffinity and molecular orbitals, polyatomic applications. *J. Chim. Phys.*, **3**, 586–591.
- Mulliken, R.S. (1955a) Electronic population analysis on LCAO-MO molecular wave functions. I. *J. Chim. Phys.*, **23**, 1833–1840.
- Mulliken, R.S. (1955b) Electronic population analysis on LCAO-MO molecular wave functions. II. Overlap populations, bond orders, and covalent bond energies. *J. Chim. Phys.*, **23**, 1841–1846.
- Mulliken, R.S. (1955c) Electronic population analysis on LCAO-MO molecular wave functions. III. Effects of hybridization on overlap and gross AO populations. *J. Chim. Phys.*, **23**, 2338–2342.
- Mulliken, R.S. (1955d) Electronic population analysis on LCAO-MO molecular wave functions. IV. Bonding and antibonding in LCAO and valence-bond theories. *J. Chim. Phys.*, **23**, 2343–2346.
- Munk Jørgensen, A.M. and Pedersen, J.T. (2001) Structural diversity of small molecule libraries. *J. Chem. Inf. Comput. Sci.*, **41**, 338–345.
- Muñoz-Muriedas, J., Perspicace, S., Bech, N., Guccione, S., Orozco, M. and Luque, F.J. (2005) Hydrophobic molecular similarity from MST fractional contributions to the octanol/water partition coefficient. *J. Comput. Aid. Mol. Des.*, **19**, 401–419.
- Murcia-Soler, M., Perez-Gimenez, F., García-March, F.J., Salabert-Salvador, M.T., Díaz-Villanueva, W. and Castro-Bleda, M.J. (2003) Drugs and nondrugs: an effective discrimination with topological methods and artificial neural networks. *J. Chem. Inf. Comput. Sci.*, **43**, 1688–1702.
- Murcia-Soler, M., Pérez-Giménez, F., Nalda-Molina, R., Salabert-Salvador, M.T., García-March, F.J., Cercos-del-Pozo, R.A. and Garrigues, T.M. (2001) QSAR analysis of hypoglycemic agents using the topological indices. *J. Chem. Inf. Comput. Sci.*, **41**, 1345–1354.
- Muresan, S., Bologa, C., Mracec, M., Chiriac, A., Jastorff, B., Simon, Z. and Naray-Szabo, G. (1995) Comparative QSAR study with electronic and steric parameters for cAMP derivatives with large substituents in position 2, position 6 and position 8. *J. Mol. Struct. (Theochem)*, **342**, 161–171.
- Murray, C.W., Auton, T.R. and Eldridge, M.D. (1998) Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand–receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput. Aid. Mol. Des.*, **12**, 503–519.
- Murray, J.S., Abu-Awwad, F. and Politzer, P. (1999) Prediction of aqueous solvation free energies from properties of solute molecular surface electrostatic potentials. *J. Phys. Chem. A*, **103**, 1853–1856.
- Murray, J.S., Brinck, T., Lane, P., Paulsen, K. and Politzer, P. (1994) Statistically-based interaction indices derived from molecular surface electrostatic potentials: a general interaction properties function (GIPF). *J. Mol. Struct. (Theochem)*, **307**, 55–64.
- Murray, J.S., Brinck, T. and Politzer, P. (1993) Partition coefficients of nitroaromatics expressed in terms of their molecular surface areas and electrostatic potentials. *J. Phys. Chem.*, **97**, 13807–13809.
- Murray, J.S., Brinck, T. and Politzer, P. (1996) Relationships of molecular surface electrostatic potentials to some macroscopic properties. *Chem. Phys.*, **204**, 289–299.
- Murray, J.S., Gagarin, S.G. and Politzer, P. (1995) Representation of C_{60} solubilities in terms of computed molecular surface electrostatic potentials and areas. *J. Phys. Chem.*, **99**, 12081–12083.
- Murray, J.S., Lane, P., Brinck, T., Paulsen, K., Grice, M.E. and Politzer, P. (1993a) Relationships of critical constants and boiling points to computed molecular surface properties. *J. Phys. Chem.*, **97**, 9369–9373.
- Murray, J.S., Lane, P., Brinck, T. and Politzer, P. (1993b) Relationships between computed molecular properties and solute–solvent interactions in supercritical solutions. *J. Phys. Chem.*, **97**, 5144–5148.
- Murray, J.S., Lane, P. and Politzer, P. (1998) Effects of strongly electron-attracting components on molecular surface electrostatic potentials: application to predicting impact sensitivities of energetic molecules. *Mol. Phys.*, **93**, 187–194.
- Murray, J.S. and Politzer, P. (1991) Correlations between the solvent hydrogen-bond-donating parameter α and the calculated molecular surface electrostatic potential. *J. Org. Chem.*, **56**, 6715–6717.
- Murray, J.S. and Politzer, P. (1998) Statistical analysis of the molecular surface electrostatic potential: an

- approach to describing noncovalent interactions in condensed phases. *J. Mol. Struct. (Theochem)*, **425**, 107–114.
- Murray, J.S., Politzer, P. and Famini, G.R. (1998) Theoretical alternatives to linear solvation energy relationships. *J. Mol. Struct. (Theochem)*, **454**, 299–306.
- Murray, J.S., Ranganathan, S. and Politzer, P. (1991) Correlations between the solvent hydrogen bond acceptor parameter β and the calculated molecular electrostatic potential. *J. Org. Chem.*, **56**, 3734–3737.
- Murray, M. (1989) Inhibition of hepatic drug metabolism by phenothiazine tranquilizers: quantitative structure–activity relationships and selective inhibition of cytochrome P-450 isoform-specific activities. *Chem. Res. Toxicol.*, **2**, 240–246.
- Murray, W.J. (1977) Molecular connectivity and steric parameters. *J. Pharm. Sci.*, **66**, 1352–1354.
- Murray, W.J., Hall, L.H. and Kier, L.B. (1975) Molecular connectivity. III. Relationship to partition coefficients. *J. Pharm. Sci.*, **64**, 1978–1981.
- Murray, W.J., Kier, L.B. and Hall, L.H. (1976) Molecular connectivity. 6. Examination of the parabolic relationship between molecular connectivity and biological activity. *J. Med. Chem.*, **19**, 573–578.
- Murrell, J.N. and Harget, A.J. (1972) *Semi-Empirical Self-Consistent-Field Molecular Orbital Theory of Molecules*, Wiley-Interscience, London, UK.
- Murugan, R., Grendze, M.P., Toomey, J.E., Jr, Katritzky, A.R., Karelson, M., Lobanov, V.S. and Rachwal, P. (1994) Predicting physical properties from molecular structure. *Chemtech*, **24**, 17–23.
- Musumarra, G., Condorelli, D.F., Costa, A.S. and Fichera, M. (2001) A multivariate insight into the *in vitro* antitumour screen database of the National Cancer Institute: classification of compounds, similarities among cell lines and the influence of molecular targets. *J. Comput. Aid. Mol. Des.*, **15**, 219–234.
- Mutelet, F., Ekulu, G. and Rogalski, M. (2002) Characterization of crude oils by inverse gas chromatography. *J. Chromat.*, **969**, 207–213.
- Mutelet, F., Ekulu, G., Solimando, R. and Rogalski, M. (2004) Solubility parameters of crude oils and asphaltenes. *Energy & Fuels*, **18**, 667–673.
- Mwense, M., Wang, X.Z., Buontempo, F.V., Horan, N., Young, A. and Osborn, D. (2004) Prediction of noninteractive mixture toxicity of organic compounds based on a fuzzy set method. *J. Chem. Inf. Comput. Sci.*, **44**, 1763–1773.
- Myers, R.H. (1986) *Classical and Modern Regression with Applications*, Duxbury Press, Boston, MA.
- Myrdal, P., Ward, G.H., Simamora, P. and Yalkowsky, S.H. (1993) AQUAFAC: aqueous functional group activity coefficients. *SAR & QSAR Environ. Res.*, **1**, 53–61.
- Myrdal, P.B. and Yalkowsky, S.H. (1997) Estimating pure component vapor pressures of complex organic molecules. *Ind. Eng. Chem. Res.*, **36**, 2494–2499.
- Myshkin, E. and Wang, B. (2003) Chemometrical classification of ephrin ligands and Eph kinases using GRID/CPCA approach. *J. Chem. Inf. Comput. Sci.*, **43**, 1004–1010.
- Nagle, J.K. (1990) Atomic polarizability and electronegativity. *J. Am. Chem. Soc.*, **112**, 4741–4747.
- Nagy, P.J., Tokarski, J. and Hopfinger, A.J. (1994) Molecular shape and QSAR analyses of a family of substituted dichlorodiphenyl aromatase inhibitors. *J. Chem. Inf. Comput. Sci.*, **34**, 1190–1197.
- Nair, A.C., Jayatilke, P., Wang, X., Miertus, S. and Welsh, W.J. (2002) Computational studies on tetrahydropyrimidine-2-one HIV-1 protease inhibitors: improving three-dimensional quantitative structure–activity relationship comparative molecular field analysis models by inclusion of calculated inhibitor- and receptor-based properties. *J. Med. Chem.*, **45**, 973–983.
- Nakayama, A., Hagiwara, K., Hashimoto, S. and Shimoda, S. (1993) QSAR of fungicidal delta(3)-1,2,4-thiadiazolines reactivity activity correlation of SH inhibitors. *Quant. Struct. -Act. Relat.*, **12**, 251–255.
- Nakayama, S., Shigezumi, S. and Yoshida, M. (1988) Method for clustering proteins by use of all possible pairs of amino acids as structural descriptors. *J. Chem. Inf. Comput. Sci.*, **28**, 72–78.
- Nanbo, A. and Nanbo, T. (2002) Mechanistic study on N-demethylation catalyzed with P450 by quantitative structure–activity relationship using electronic properties of 4-substituted N,N-dimethylaniline. *Quant. Struct. -Act. Relat.*, **21**, 613–616.
- Nandy, A. (1994) A new graphical representation and analysis of DNA sequence structure. I. Methodology and application to globin genes. *Curr. Sci. -India*, **66**, 309–314.
- Nandy, A. (1996a) Graphical analysis of DNA sequence structure. III. Indications of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci. -India*, **70**, 661–668.
- Nandy, A. (1996b) Two-dimensional graphical representation of DNA sequences and intron–exon discrimination in intron-rich sequences. *Comput. Appl. Biosci.*, **12**, 55–62.

- Nandy, A. and Basak, S.C. (2000) Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J. Chem. Inf. Comput. Sci.*, **40**, 915–919.
- Nandy, A. and Basak, S.C. (2005) Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J. Chem. Inf. Comput. Sci.*, **40**, 915–919.
- Nandy, A., Basak, S.C. and Gute, B.D. (2007) Graphical representation and numerical characterization of H5N1 avian flu neuraminidase gene sequence. *J. Chem. Inf. Model.*, **47**, 945–951.
- Nandy, A., Harle, M. and Basak, S.C. (2006) Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC*, (**ix**), 211–238.
- Nandy, A. and Nandy, P. (1995) Graphical analysis of DNA sequence structure. II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. *Curr. Sci. -India*, **68**, 75–85.
- Nandy, A. and Nandy, P. (2003) On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models. *Chem. Phys. Lett.*, **368**, 102–107.
- Nandy, A., Nandy, P. and Basak, S.C. (2002) Quantitative descriptor for SNP related gene sequences. *Internet Electron. J. Mol. Des.*, **1**, 367–373.
- Naray-Szabo, G. and Balogh, T. (1993) Viewpoint 7 – the average molecular electrostatic field as a QSAR descriptor. 4. Hydrophobicity scales for amino acid residues alpha. *J. Mol. Struct. (Theochem)*, **284**, 243–248.
- Narumi, H. (1987) New topological indices for finite and infinite systems. *MATCH Commun. Math. Comput. Chem.*, **22**, 195–207.
- Narumi, H. and Hosoya, H. (1980) Topological index and thermodynamics properties. II. Analysis of topological factors on the absolute entropy of acyclic saturated hydrocarbons. *Bull. Chem. Soc. Jap.*, **53**, 1228–1237.
- Narumi, H. and Hosoya, H. (1985) Topological index and thermodynamics properties. III. Classification of various topological aspects of properties of acyclic saturated hydrocarbons. *Bull. Chem. Soc. Jap.*, **58**, 1778–1786.
- Narumi, H. and Katayama, M. (1984) Simple topological index – a newly devised index characterizing the topological nature of structural isomers of saturated hydrocarbons. *Memories of Faculty of Engineering of Hokkaido, University*, **16**, 209–214.
- Natarajan, R., Basak, S.C. and Neumann, T.S. (2007) Novel approach for the numerical characterization of molecular chirality. *J. Chem. Inf. Model.*, **47**, 771–775.
- Natarajan, R., Kamalakanan, P. and Nirdosh, I. (2003) Applications of topological indices to structure–activity relationship modelling and selection of mineral collectors. *Indian J. Chem.*, **42**, 1330–1346.
- Natarajan, R., Nirdosh, I., Basak, S.C. and Mills, D.R. (2002) QSAR modeling of flotation collectors using principal components extracted from topological indices. *J. Chem. Inf. Comput. Sci.*, **42**, 1425–1430.
- Navajas, C., Poso, A., Tuppurainen, K. and Gynther, J. (1996) Comparative molecular field analysis (CoMFA) of MX compounds using different semiempirical methods. LUMO field and its correlation with mutagenic activity. *Quant. Struct. - Act. Relat.*, **15**, 189–193.
- Needham, D.E., Wei, I.C. and Seybold, P.G. (1988) Molecular modeling of the physical properties of the alkanes. *J. Am. Chem. Soc.*, **110**, 4186–4194.
- Nefati, H., Cense, J.-M. and Legendre, J.J. (1996) Prediction of the impact sensitivity by neural networks. *J. Chem. Inf. Comput. Sci.*, **36**, 804–810.
- Nefati, H., Diawara, B. and Legendre, J.J. (1993) Predicting the impact sensitivity of explosive molecules using neuromimetic networks. *SAR & QSAR Environ. Res.*, **1**, 131–136.
- Nelson, S.D. and Seybold, P.G. (2001) Molecular structure–property relationships for alkenes. *J. Mol. Graph. Model.*, **20**, 36–53.
- Nelson, T.M. and Jurs, P.C. (1994) Prediction of aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.*, **34**, 601–609.
- Nemba, R.M. and Balaban, A.T. (1998) Algorithm for the direct enumeration of chiral and achiral skeletons of a homosubstituted derivative of a monocyclic cycloalkane with a large and factorizable ring size n . *J. Chem. Inf. Comput. Sci.*, **38**, 1145–1150.
- Nendza, M. and Müller, M. (2000) Discriminating toxicant classes by mode of action. 2. Physico-chemical descriptors. *Quant. Struct. -Act. Relat.*, **19**, 581–598.
- Netzeva, T.I., Aptula, A.O., Benfenati, E., Cronin, M.T.D., Gini, G., Lessigiarska, I., Maran, U., Vračko, M. and Schüürmann, G. (2005) Description of the electronic structure of organic chemicals using semiempirical and *ab initio* methods for development of toxicological QSARs. *J. Chem. Inf. Model.*, **45**, 106–114.
- Netzeva, T.I., Aptula, A.O., Chaudary, S.H., Duffy, J. C., Schultz, T.W., Schüürmann, G. and Cronin, M. T.D. (2003) Structure–activity relationships for the toxicity of substituted poly-hydroxylated benzenes

- to *Tetrahymena pyriformis*: influence of free radical formation. *QSAR Comb. Sci.*, **22**, 575–582.
- Netzeva, T.I., Dearden, J.C., Edwards, R., Worgan, A. D.P. and Cronin, M.T.D. (2004) QSAR analysis on the toxicity of aromatic compounds to *Chlorella vulgaris* in a novel short-term assay. *J. Chem. Inf. Comput. Sci.*, **44**, 258–265.
- Netzeva, T.I., Pavan, M. and Worth, A.P. (2008) Review of (quantitative) structure–activity relationships for acute aquatic toxicity. *QSAR Comb. Sci.*, **27**, 77–90.
- Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliaskova, N., Patlewicz, G., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., van de Sandt, J.J.M., Tong, W.D., Veith, G.D. and Yang, C. (2005) Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. *ATLA*, **33**, 155–173.
- Neudert, R. and Davies, A.N. (2003) Spectroscopic databases, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 700–721.
- Nevalainen, T. and Kolehmainen, E. (1994) New QSAR models for polyhalogenated aromatics. *Environ. Toxicol. Chem.*, **13**, 1699–1706.
- Newman, M.E.J. (2004) Analysis of weighted networks. *Phys. Rev. E*, **70**, 056131.
- Newman, M.E.J. (2005) A measure of betweenness centrality based on random walks. *Social Networks*, **27**, 39–54.
- Newman, M.S. (1950) Some observations concerning steric factors. *J. Am. Chem. Soc.*, **72**, 4783–4786.
- Nguyen-Cong, V. and Rode, B.M. (1996) Quantum pharmacological analysis of structure–activity relationships for mefloquine antimalarial drugs using optimal transformations. *J. Chem. Inf. Comput. Sci.*, **36**, 114–117.
- Nguyen-Cong, V., Vandang, G. and Rode, B.M. (1996) Using multivariate adaptive regression splines to QSAR studies of dihydroartemisinin derivatives. *Eur. J. Med. Chem.*, **31**, 797–803.
- Nicholls, A., MacCuish, N.E. and MacCuish, J.D. (2004) Variable selection and model validation of 2D and 3D molecular descriptors. *J. Comput. Aid. Mol. Des.*, **18**, 451–474.
- Nicklaus, M.C. (2003) Pharmacophore and drug discovery, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1687–1711.
- Nicklaus, M.C., Milne, G.W.A. and Burke, T.R. (1992) QSAR of conformationally flexible molecules: comparative molecular field analysis of protein-tyrosine kinase inhibitors. *J. Comput. Aid. Mol. Des.*, **6**, 487–504.
- Nicolotti, O., Gillet, V.J., Fleming, P.J. and Green, D.V.S. (2002) Optimization in quantitative structure–activity relationships: deriving accurate and interpretable QSARs. *J. Med. Chem.*, **45**, 5069–5080.
- Nicolotti, O., Pellegrini-Calace, M., Carrieri, A., Altomare, C., Centeno, N.B., Sanz, F. and Carotti, A. (2001) Neuronal nicotinic receptor agonists: a multi-approach development of the pharmacophore. *J. Comput. Aid. Mol. Des.*, **15**, 859–872.
- Niculescu, S.P. (2003) Artificial neural networks and genetic algorithms in QSAR. *J. Mol. Struct. (Theochem)*, **622**, 71–83.
- Nidiry, E.S.J. (2003) Quantitative structure–fungitoxicity relationships of some monohydric alcohols. *J. Agr. Food Chem.*, **51**, 5337–5343.
- Nie, C., Dai, Y.-M., Wen, S.-N., Li, Z., Zhou, C. and Peng, G.-W. (2005) Topological homologous regularity for additive property of alkanes. *Acta Chim. Sin.*, **63**, 1449–1455.
- Niemi, G.J., Basak, S.C., Veith, G.D. and Grunwald, G.D. (1992) Prediction of octanol/water partition coefficient (K_{ow}) with algorithmically derived variables. *Environ. Toxicol. Chem.*, **11**, 893–900.
- Nikolić, S., Kovacevic, G., Miličević, A. and Trinajstić, N. (2003) The Zagreb indices 30 years after. *Croat. Chem. Acta*, **76**, 113–124.
- Nikolić, S., Medicsaric, M. and Matijevicsosa, J. (1993) A QSAR study of 3-(phthalimidoalkyl)-pyrazolin-5-ones. *Croat. Chem. Acta*, **66**, 151–160.
- Nikolić, S., Miličević, A. and Trinajstić, N. (2005) Graphical matrices in chemistry. *Croat. Chem. Acta*, **78**, 241–250.
- Nikolić, S., Miličević, A., Trinajstić, N. and Jurić, A. (2004) On use of the variable Zagreb $^mM_2$ index in QSPR: boiling points of benzenoid hydrocarbons. *Molecules*, **9**, 1208–1221.
- Nikolić, S., Plavšić, D. and Trinajstić, N. (1992) On the Z-counting polynomial for edge-weighted graphs. *J. Math. Chem.*, **9**, 381–387.
- Nikolić, S., Plavšić, D. and Trinajstić, N. (2001) On the Balaban-like topological indices. *MATCH Commun. Math. Comput. Chem.*, **44**, 361–386.
- Nikolić, S. and Raos, N. (2001) Estimation of stability constants of mixed amino acid complexes with copper(II) from topological indices. *Croat. Chem. Acta*, **74**, 621–631.
- Nikolić, S., Tolić, I.M., Trinajstić, N. and Baučić, I. (2000) On the Zagreb indices as complexity indices. *Croat. Chem. Acta*, **73**, 909–921.
- Nikolić, S. and Trinajstić, N. (1997) On the concept of a chemical model. *Croat. Chem. Acta*, **70**, 777–786.

- Nikolić, S. and Trinajstić, N. (1998) Modeling the aqueous solubility of aliphatic alcohols. *SAR & QSAR Environ. Res.*, **9**, 117–126.
- Nikolić, S. and Trinajstić, N. (2000) Complexity of molecules. *J. Chem. Inf. Comput. Sci.*, **40**, 920–926.
- Nikolić, S., Trinajstić, N., Amić, D., Bešlo, D. and Basak, S.C. (2001a) Modeling the solubility of aliphatic alcohols in water. Graph connectivity indices versus line graph connectivity indices, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 63–81.
- Nikolić, S., Trinajstić, N. and Baučić, I. (1998) Comparison between the vertex- and edge-connectivity indices for benzenoid hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **38**, 42–46.
- Nikolić, S., Trinajstić, N. and Ivaniš, S. (1999a) The connectivity indices of regular graphs. *Croat. Chem. Acta*, **72**, 875–883.
- Nikolić, S., Trinajstić, N., Jurić, A. and Mihalić, Z. (1996a) The Detour matrix and the Detour index of weighted graphs. *Croat. Chem. Acta*, **69**, 1577–1591.
- Nikolić, S., Trinajstić, N., Jurić, A., Mihalić, Z. and Krilov, G. (1996b) Complexity of some interesting (chemical) graphs. *Croat. Chem. Acta*, **69**, 883–897.
- Nikolić, S., Trinajstić, N. and Mihalić, Z. (1993) Molecular topological index: an extension to heterosystems. *J. Math. Chem.*, **12**, 251–264.
- Nikolić, S., Trinajstić, N. and Mihalić, Z. (1995) The Wiener index: development and applications. *Croat. Chem. Acta*, **68**, 105–129.
- Nikolić, S., Trinajstić, N. and Mihalić, Z. (1999b) The Detour matrix and the Detour index, in *Topological Indices and Related Descriptors in QSAR and QSPR* (eds J. Devillers and A.T. Balaban), Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp. 279–306.
- Nikolić, S., Trinajstić, N., Mihalić, Z. and Carter, S. (1991) On the geometric-distance matrix and the corresponding structural invariants of molecular systems. *Chem. Phys. Lett.*, **179**, 21–28.
- Nikolić, S., Trinajstić, N. and Randić, M. (2001b) Wiener index revisited. *Chem. Phys. Lett.*, **333**, 319–321.
- Nikolić, S., Trinajstić, N., Tolić, I.M., Rücker, G. and Rücker, C. (2003) On molecular complexity indices, in *Complexity in Chemistry: Introduction and Fundamentals*, Vol. 7 (eds D. Bonchev and D.E. Rouvray), Taylor & Francis, London, UK, pp. 29–89.
- Nikolova, N. and Jaworska, J.S. (2003) Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.*, **22**, 1006–1026.
- Nikolova-Jeliazkova, N. and Jaworska, J.S. (2005) An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN. *ATLA*, **33**, 461–470.
- Nikolovska-Coleska, Z., Suturkova, L., Dorevski, K., Krbavčić, A. and Solmajer, T. (1998) Quantitative structure–activity relationship of flavonoid inhibitors of p56lck protein tyrosine kinase: a classical/quantum chemical approach. *Quant. Struct. -Act. Relat.*, **17**, 7–13.
- Nilakantan, R., Bauman, N., Dixon, J.S. and Venkataraghavan, R. (1987) Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.*, **27**, 82–85.
- Nilakantan, R., Bauman, N. and Venkataraghavan, R. (1993) New method for rapid characterization of molecular shapes: applications in drug design. *J. Chem. Inf. Comput. Sci.*, **33**, 79–85.
- Nilsson, J., Homan, E.J., Smilde, A.K., Grol, C.J. and Wikström, H. (1998) A multiway 3D QSAR analysis of a series of (S)-N-[(1-ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides. *J. Comput. Aid. Mol. Des.*, **12**, 81–93.
- Niño, M.V., Daza, E.E.C. and Tello, M. (2001) A criteria to classify biological activity of benzimidazoles from a model of structural similarity. *J. Chem. Inf. Comput. Sci.*, **41**, 495–504.
- Nirmalakhandan, N.N. and Speece, R.E. (1988a) Prediction of aqueous solubility of organic chemicals based on molecular structure. *Environ. Sci. Technol.*, **22**, 328–338.
- Nirmalakhandan, N.N. and Speece, R.E. (1988b) Structure–activity relationships. Quantitative techniques for predicting the behavior of chemicals in the ecosystem. *Environ. Sci. Technol.*, **22**, 606–615.
- Nirmalakhandan, N.N. and Speece, R.E. (1989a) Prediction of aqueous solubility of organic chemicals based on molecular structure. 2. Application to PNAAs, PCBs, PCDDs, etc. *Environ. Sci. Technol.*, **23**, 708–713.
- Nirmalakhandan, N.N. and Speece, R.E. (1989b) QSAR model for predicting Henry's constant. *Environ. Sci. Technol.*, **22**, 1349–1357.
- Nirmalakhandan, N.N. and Speece, R.E. (1993) Prediction of activated carbon adsorption capacities for organic vapors using quantitative structure–activity relationship methods. *Environ. Sci. Technol.*, **27**, 1512–1516.
- Nissink, J.W.M., Verdonk, M.L. and Klebe, G. (2000) Simple knowledge-based descriptors to predict protein–ligand interactions. Methodology and validation. *J. Comput. Aid. Mol. Des.*, **14**, 787–803.
- Nissink, J.W.M., Verdonk, M.L. and Klebe, G. (2001) Knowledge-based descriptors to predict protein–

- ligand interactions. Derivation, use, and application to knowledge-based molecular alignment, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona (Spain), pp. 115–124.
- Niwa, T. (2003) Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.*, **43**, 113–119.
- No, K.T., Cho, K.H., Jhon, M.S. and Scheraga, H.A. (1993) An empirical method to calculate average molecular polarizabilities from the dependence of effective atomic polarizabilities on net atomic charge. *J. Am. Chem. Soc.*, **115**, 2005–2014.
- No, K.T., Grant, J.A., Jhon, M.S. and Scheraga, H.A. (1990a) Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 2. Application to ionic and aromatic molecules as models for polypeptides. *J. Phys. Chem.*, **94**, 4740–4746.
- No, K.T., Grant, J.A. and Scheraga, H.A. (1990b) Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 1. Application to neutral molecules as models for polypeptides. *J. Phys. Chem.*, **94**, 4732–4739.
- Noeske, T., Sasse, B.C., Stark, H., Parsons, C.G., Weil, T. and Schneider, G. (2006) Predicting compound selectivity by self-organizing maps: cross-activities of metabotropic glutamate receptor antagonists. *ChemMedChem*, **1**, 1066–1068.
- Nohair, M. and Zakarya, D. (2003) Prediction of solubility of aliphatic alcohols using the restricted components of autocorrelation method (RCAM). *J. Mol. Model.*, **9**, 365–371.
- Nohair, M., Zakarya, D. and Berrada, A. (2002) Autocorrelation method adapted to generate new atomic environments: application for the prediction of ¹³C chemical shifts of alkanes. *J. Chem. Inf. Comput. Sci.*, **42**, 586–591.
- Nord, L.I., Fransson, D. and Jacobsson, S.P. (1998) Prediction of liquid chromatographic retention times of steroids by three-dimensional structure descriptors and partial least squares modeling. *Chemom. Intell. Lab. Syst.*, **44**, 257–269.
- Nordqvist, A., Nilsson, J., Lindmark, T., Eriksson, A., Garberg, P. and Kihlén, M. (2004) A general model for prediction of Caco-2 cell permeability. *QSAR Comb. Sci.*, **23**, 303–310.
- Norel, R., Fisher, D., Wolfson, H.J. and Nussinov, R. (1994) Molecular surface recognition by a computer vision-based technique. *Protein Engineering*, **7**, 39–46.
- Norinder, U. (1991) Theoretical amino acid descriptors. Application to bradykinin potentiating peptides. *Peptides*, **12**, 1223–1227.
- Norinder, U. (1992) Experimental design based quantitative structure toxicity relationship of some local anesthetics using the PLS method. *J. Appl. Toxicol.*, **12**, 143–147.
- Norinder, U. (1993) Multivariate Free–Wilson analysis of some N-alkylmorphinan-6-one opioids using PLS. *Quant. Struct. -Act. Relat.*, **12**, 119–123.
- Norinder, U. (1994) Theoretical descriptors of nucleic acid bases. Application to DNA promotor sequences. *Quant. Struct. -Act. Relat.*, **13**, 295–301.
- Norinder, U. (1996a) 3D-QSAR investigation of the Tripos benchmark steroids and some protein-tyrosine kinase inhibitors of styrene type using the TDQ approach. *J. Chemom.*, **10**, 533–545.
- Norinder, U. (1996b) Single and domain mode variable selection in 3D QSAR applications. *J. Chemom.*, **10**, 95–105.
- Norinder, U. (1998) Recent progress in CoMFA methodology and related techniques, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 25–39.
- Norinder, U., Florvall, L. and Ross, S.B. (1994) A PLS quantitative structure–activity relationship study of some monoamine oxidase inhibitors of the phenyl alkylamine type. *Eur. J. Med. Chem.*, **29**, 191–195.
- Norinder, U. and Haeberlein, M. (2002) Computational approaches to the prediction of the blood–brain distribution. *Adv. Drug Deliv. Rev.*, **54**, 291–313.
- Norinder, U. and Haeberlein, M. (2003) Calculated molecular properties and multivariate statistical analysis in absorption prediction, in *Drug Bioavailability*, Vol. 18, Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 358–397.
- Norinder, U. and Hogberg, T. (1992) PLS-based quantitative structure–activity relationship for substituted benzamides of clebopride type. Application of experimental design in drug design. *Acta Chem. Scand.*, **46**, 363–366.
- Norinder, U., Österberg, T. and Artursson, P. (1997) Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. *Pharm. Res.*, **14**, 1786–1791.
- Norinder, U., Österberg, T. and Artursson, P. (1999) Theoretical calculation and prediction of intestinal absorption of drugs in humans using MolSurf parametrization and PLS statistics. *Eur. J. Pharm. Sci.*, **8**, 49–56.
- Norinder, U., Sjöberg, P. and Österberg, T. (1998) Theoretical calculation and prediction of brain–

- blood partitioning of organic solutes using MolSurf parametrization and PLS statistics. *J. Pharm. Sci.*, **87**, 952–959.
- Norrington, F.E., Hyde, R.M., Williams, S.G. and Wootton, R. (1975) Physico-chemical activity relations in practice. 1. A rational and self-consistent data bank. *J. Med. Chem.*, **18**, 604–607.
- Nouwen, J., Lindgren, F., Hansen, B. and Karcher, W. (1996) Fast screening of large databases using clustering and PCA based on structure fragments. *J. Chemom.*, **10**, 385–398.
- Nouwen, J., Lindgren, F., Hansen, B. and Karcher, W. (1997) Classification of environmentally occurring chemicals using structural fragments and PLS discriminant analysis. *Environ. Sci. Technol.*, **31**, 2313–2318.
- Novič, M., Nikolovska-Coleska, Z. and Solmajer, T. (1997) Quantitative structure–activity relationship of flavonoid p56^{lck} protein tyrosine kinase inhibitors. A neural network approach. *J. Chem. Inf. Comput. Sci.*, **37**, 990–998.
- Novič, M. and Vračko, M. (2001) Comparison of spectrum-like representation of 3D chemical structure with other representations when used for modelling biological activity. *Chemom. Intell. Lab. Syst.*, **59**, 33–44.
- Novič, M. and Zupan, J. (1996) A new general approach and uniform structure representation, in *Software Development in Chemistry*, Vol. 10 (ed. J. Gasteiger), Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main, Germany, pp. 47–58.
- Novikov, V.P. and Raevsky, O.A. (1982) Representation of molecular structure as a spectrum of interatomic distances for the study of structure–biological activity relations. *Khimico-Farmaceuticheskii Zhurnal*, **16**, 574–581.
- Noy, M. (2003) Graphs determined by polynomial invariants. *Theor. Comp. Sci.*, **307**, 365–384.
- Nusser, T., Balogh, T. and Naray-Szabo, G. (1993) The average molecular electrostatic field as a QSAR descriptor. 5. Hydrophobicity indexes for small molecules. *J. Mol. Struct.*, **297**, 127–132.
- NWChem, EMSL, Pacific Northwest National Laboratory, Richland, WA.
- Nys, G.G. and Rekker, R.F. (1973) Statistical analysis of a series of partition coefficients with special reference to the predictability of folding of drug molecules. The introduction of hydrophobic fragmental constants (*f*-values). *Eur. J. Med. Chem.*, **8**, 521–535.
- Nys, G.G. and Rekker, R.F. (1974) The concept of hydrophobic fragmental constants (*f*-values). II. Extension of its applicability to the calculation of lipophilicities of aromatic and heteroaromatic structures. *Eur. J. Med. Chem.*, **9**, 361–375.
- Nyström, Å., Andersson, P.M. and Lundstedt, T. (2000) Multivariate data analysis of topographically modified α -melanotropin analogues using auto and cross auto covariances (ACC). *Quant. Struct. - Act. Relat.*, **19**, 264–269.
- O'Brien, S.E. and Popelier, P.L.A. (2001) Quantum molecular similarity. 3. QTMS descriptors. *J. Chem. Inf. Comput. Sci.*, **41**, 764–775.
- Öberg, T. (2004a) A QSAR for baseline toxicity: validation, domain of application, and prediction. *Chem. Res. Toxicol.*, **17**, 1630–1637.
- Öberg, T. (2004b) Boiling points of halogenated aliphatic compounds: a quantitative structure–property relationship for prediction and validation. *J. Chem. Inf. Comput. Sci.*, **44**, 187–192.
- Oberrauch, E. and Mazzanti, V. (1990) Partial-least-squares models for the octane number of alkanes based on subgraph descriptors. *Anal. Chim. Acta*, **235**, 177–188.
- Ohlenbusch, G. and Frimmel, F.H. (2001) Investigations on the sorption of phenols to dissolved organic matter by a QSAR study. *Chemosphere*, **45**, 323–327.
- Okamoto, A.K., Gaudio, A.C., Marques, A., dos, S. and Takahata, Y. (2005) QSAR study of inhibition by coumarins of IQ induced mutation in *S. typhimurium* TA98. *J. Mol. Struct. (Theochem)*, **725**, 231–238.
- Okamoto, Y. and Brown, H.C. (1958) Rates of solvolysis of phenyldimethylcarbinyl chlorides containing substituents ($-NMe_3^+$, $-CO_2^-$) bearing a charge. *J. Am. Chem. Soc.*, **80**, 4976–4979.
- Okamoto, Y., Inukai, T. and Brown, H.C. (1958a) Rates of solvolysis of phenyldimethylcarbinyl chlorides containing *meta* directing substituents. *J. Am. Chem. Soc.*, **80**, 4969–4972.
- Okamoto, Y., Inukai, T. and Brown, H.C. (1958b) Rates of solvolysis of phenyldimethylcarbinyl chlorides in methyl, ethyl and isopropyl alcohols. Influence of the solvent on the value of the electrophilic substituent constant. *J. Am. Chem. Soc.*, **80**, 4972–4976.
- Okey, R.W. and Martis, M.C. (1999) Molecular level studies on the origin of toxicity: identification of key variables and selection of descriptors. *Chemosphere*, **38**, 1419–1427.
- Okey, R.W. and Stensel, H.D. (1996) A QSAR-based biodegradability model. A QSBR. *Water Res.*, **30**, 2206–2214.
- Okey, R.W., Stensel, H.D. and Martis, M.C. (1996) Modeling nitrification inhibition. *Water Sci. Technol.*, **33**, 101–107.

- Okouchi, S. and Saegusa, H. (1989) Prediction of soil sorption coefficients of hydrophobic organic pollutants by adsorbability index. *Bull. Chem. Soc. Jap.*, **62**, 922–924.
- Okouchi, S., Saegusa, H. and Nojima, O. (1992) Prediction of environmental parameters by adsorbability index: water solubilities of hydrophobic organic pollutants. *Environment International*, **18**, 249–261.
- Oláh, J., Blockhuys, F., Veszprémi, T. and Van Alsenoy, C. (2006) On the usefulness of bond orders and overlap populations to chalcogen–nitrogen systems. *Eur. J. Inorg. Chem.*, 69–77.
- Olah, M., Bologa, C. and Oprea, T.I. (2004a) An automated PLS search for biologically relevant QSAR descriptors. *J. Comput. Aid. Mol. Des.*, **18**, 437–449.
- Olah, M., Bologa, C. and Oprea, T.I. (2004b) Strategies for compound selection. *Curr. Drug Discov. Technol.*, **1**, 211–220.
- Olivero, J. and Kannan, K. (1999) Quantitative structure–retention relationships of polychlorinated naphthalenes in gas chromatography. *J. Chromat.*, **849**, 621–627.
- Olivero-Verbel, J. and Pacheco-Londoño, L. (2002) Structure–activity relationships for the anti-HIV activity of flavonoids. *J. Chem. Inf. Comput. Sci.*, **42**, 1241–1246.
- Oloff, S., Zhang, S., Sukumar, N., Breneman, C.M. and Tropsha, A. (2006) Chemometric analysis of ligand receptor complementarity: identifying complementary ligands based on receptor information (CoLiBRI). *J. Chem. Inf. Model.*, **46**, 844–851.
- Olsen, E. and Nielsen, F. (2001) Predicting vapour pressures of organic compounds from their chemical structure for classification according to the VOC-directive and risk assessment in general. *Molecules*, **6**, 370–389.
- Olsson, T. and Oprea, T.I. (2001) Chemoinformatics: a tool for decision-makers in drug discovery. *Current Opinion in Drug Discovery & Development*, **4**, 308–313.
- Onicescu, O. (1966) Energie informationelle. *Comp. Rend. Acad. Sci. (Paris, French)*, **263**, 841–842.
- Oprea, T.I. (2000) Property distribution of drug-related chemical databases. *J. Comput. Aid. Mol. Des.*, **14**, 251–264.
- Oprea, T.I. (2001) Rapid estimation of hydrophobicity for virtual combinatorial library analysis. *SAR & QSAR Environ. Res.*, **12**, 129–141.
- Oprea, T.I. (2002a) Current trends in lead discovery: are we looking for the appropriate properties? *Mol. Div.*, **5**, 199–208.
- Oprea, T.I. (2002b) On the information content of 2D and 3D descriptors for QSAR. *J. Braz. Chem. Soc.*, **13**, 811–815.
- Oprea, T.I. (2002c) Virtual screening in lead discovery: a viewpoint. *Molecules*, **7**, 51–62.
- Oprea, T.I. (2003) Chemoinformatics and the quest for leads in drug discovery, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1509–1531.
- Oprea, T.I. (2004) 3D QSAR modeling in drug design, in *Computational Medicinal Chemistry for Drug Discovery* (eds P. Bultinck H. De Winter, W. Langenaeker and J.P. Tollenaere), Marcel Dekker, New York, pp. 571–616.
- Oprea, T.I., Ciubotariu, D., Sulea, T. and Simon, Z. (1993) Comparison of the minimal steric difference (MTD) and comparative molecular field analysis (CoMFA) methods for analysis of binding of steroids to carrier proteins. *Quant. Struct. -Act. Relat.*, **12**, 21–26.
- Oprea, T.I., Davis, A.M., Teague, S.J. and Leeson, P.D. (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.*, **41**, 1308–1315.
- Oprea, T.I. and Garcia, A.E. (1996) Three-dimensional quantitative structure–activity relationships of steroid aromatase inhibitors. *J. Comput. Aid. Mol. Des.*, **10**, 186–200.
- Oprea, T.I. and Gottfries, J. (2001a) ChemGPS: a chemical space navigation tool, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 437–446.
- Oprea, T.I. and Gottfries, J. (2001b) Chemography: the art of navigating in chemical space. *J. Comb. Chem.*, **3**, 157–166.
- Oprea, T.I., Gottfries, J., Sherbukhin, V., Svensson, P. and Kühler, T.C. (2000) Chemical information management in drug discovery: optimizing the computational and combinatorial chemistry interfaces. *J. Mol. Graph. Model.*, **18**, 512–524.
- Oprea, T.I., Kurunczi, L., Olah, M. and Simon, Z. (2001) MTD-PLS: a PLS-based variant of the MTD method. A 3D-QSAR analysis of receptor affinities for a series of halogenated dibenzoxin and biphenyl derivatives. *SAR & QSAR Environ. Res.*, **12**, 75–92.
- Oprea, T.I., Kurunczi, L. and Timofei, S. (1997) QSAR studies of disperse azo dyes. Towards the negation of the pharmacophore theory of dye–fiber interaction? *Dyes & Pigments*, **33**, 41–64.
- Oprea, T.I. and Matter, H. (2004) Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.*, **8**, 349–358.

- Oprea, T.I. and Waller, C.L. (1997) Theoretical and practical aspects of three-dimensional quantitative structure–activity relationships, in *Reviews in Computational Chemistry*, Vol. 11 (eds K.B. Lipkowitz and D. Boyd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 127–182.
- Oprea, T.I., Zamora, I. and Ungell, A.-L. (2002) Pharmacokinetically based mapping device for chemical space navigation. *J. Comb. Chem.*, **4**, 258–266.
- Ordorica, M.A., Velazquez, M.L., Ordorica, J.G., Escobar, J.L. and Lehmann, P.A. (1993) A principal component and cluster significance analysis of the antiparasitic potency of praziquantel and some analogs. *Quant. Struct. -Act. Relat.*, **12**, 246–250.
- Ormerod, A., Willett, P. and Bawden, D. (1989) Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct. -Act. Relat.*, **8**, 115–129.
- Ormerod, A., Willett, P. and Bawden, D. (1990) Further comparative studies of fragment weighting schemes for substructural analysis. *Quant. Struct. -Act. Relat.*, **9**, 302–312.
- Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C. (1995) Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.*, **38**, 2681–2691.
- Ósk Jónsdóttir, S., Jørgensen, F.S. and Brunak, S. (2005) Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics*, **21**, 2145–2160.
- Osmialowski, K., Halkiewicz, J. and Kaliszan, R. (1986) Quantum chemical parameters in correlation analysis of gas–liquid chromatographic retention indices of amines. II. Topological electronic index. *J. Chromat.*, **361**, 63–69.
- Osmialowski, K., Halkiewicz, J., Radecki, A. and Kaliszan, R. (1985) Quantum chemical parameters in correlation analysis of gas–liquid chromatographic retention indices of amines. *J. Chromat.*, **346**, 53–60.
- Osmialowski, K. and Kaliszan, R. (1991) Studies of performance of graph theoretical indices in QSAR analysis. *Quant. Struct. -Act. Relat.*, **10**, 125–134.
- Osten, D.W. (1988) Selection of optimal regression models via cross-validation. *J. Chemom.*, **2**, 39–48.
- Österberg, T. and Norinder, U. (2000) Prediction of polar surface area and drug transport processes using simple parameters and PLS statistics. *J. Chem. Inf. Comput. Sci.*, **40**, 1408–1411.
- Oth, J.F.M. and Gilles, J.-M. (1968) Mobilité Conformationnelle et Isomérisation de Valence Rapide Réversible dans le [16] Annulène. *Tetrahedron Lett.*, **1968**, 6259–6264.
- Otsuji, Y., Kubo, M. and Imoto, E. (1960) Reactivities of heterocyclic compounds. IX. A method to systematize the reactivities of the substituents in aromatic and heteroaromatic compounds. *Engineering and Natural Sciences - Osaka University*, **7**, 61–70.
- Ouyang, Z., Yuan, S., Brandt, J. and Zheng, C. (1999) An effective topological symmetry perception and unique numbering algorithm. *J. Chem. Inf. Comput. Sci.*, **39**, 299–303.
- Overton, E. (2001) *Studien über die Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie*, Verlag Gustav Fischer, Jena, Germany, p. 141.
- Overton, E. (1991) *Studies on Narcosis*, Chapman & Hall, London, UK (English translation).
- Pacios, L.F. (2001) Distinct molecular surfaces and hydrophobicity of amino acid residues in proteins. *J. Chem. Inf. Comput. Sci.*, **41**, 1427–1435.
- Padrón, J.A., Carrasco, R. and Pellón, R.F. (2002) Molecular descriptor based on a molar refractivity partition using Randić-type graph-theoretical invariant. *J. Pharm. Pharmaceut. Sci.*, **5**, 258–265.
- Pagliara, A., Caron, G., Lisa, G., Fan, W., Gaillard, P., Carrupt, P.-A., Testa, B. and Abraham, M.H. (1997) Solvatochromic analysis of di-n-butyl ether/water partition coefficients as compared to other solvent systems. *J. Chem. Soc. Perkin Trans. 2*, 2639–2643.
- Pagliara, A., Carrupt, P.-A., Caron, G., Gaillard, P. and Testa, B. (1997) Lipophilicity profiles of ampholytes. *Chem. Rev.*, **97**, 3385–3400.
- Pagliara, A., Khamis, E., Trinh, A., Carrupt, P.-A. and Tsai, R.-S., Testa, B. (1995) Structural properties governing retention mechanisms on RP-HPLC stationary phases used for lipophilicity measurements. *J. Liquid Chromat.*, **18**, 1721–1745.
- Pal, D.K., Purkayastha, S.K., Sengupta, C. and De, A. U. (1992) Quantitative structure–property relationships with TAU indices. Part I. Research octane numbers of alkane fuel molecules. *Indian J. Chem.*, **31**, 109–114.
- Pal, D.K., Sengupta, C. and De, A.U. (1988) A new topochemical descriptor (TAU) in molecular connectivity concept. Part I. Aliphatic compounds. *Indian J. Chem.*, **27**, 734–739.
- Pal, D.K., Sengupta, C. and De, A.U. (1989) Introduction of a novel topochemical index and exploitation of group connectivity concept to achieve predictability in QSAR and RDD. *Indian J. Chem.*, **28**, 261–267.
- Pal, D.K., Sengupta, C. and De, A.U. (1990) QSAR with TAU (·) indices. Part I. Polymethylene primary diamines as amebicidal agents. *Indian J. Chem.*, **29**, 451–454.
- Palacios, J.L. (2001) Resistance distance in graphs and random walks. *Int. J. Quant. Chem.*, **81**, 29–33.

- Palm, K., Luthman, K., Ungell, A.-L., Strandlund, G. and Artursson, P. (1996) Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.*, **85**, 32–39.
- Palm, K., Luthman, K., Ungell, A.-L., Strandlund, G., Beigi, F., Lundahl, P. and Artursson, P. (1998) Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *J. Med. Chem.*, **41**, 5382–5392.
- Palm, V.A. (1972) *Fundamentals of the Quantitative Theory of Organic Reactions*, Khimiya, Leningrad, Russia.
- Palyulin, V.A., Baskin, I.I., Petelin, D.E. and Zefirov, N.S. (1995) Novel descriptors of molecular structure in QSAR and QSPR studies, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, Barcelona, Spain, pp. 51–52.
- Palyulin, V.A., Radchenko, E.V. and Zefirov, N.S. (2000) Molecular field topology analysis method in QSAR studies of organic compounds. *J. Chem. Inf. Comput. Sci.*, **40**, 659–667.
- Pan, Y., Huang, N., Cho, S. and MacKerell, A.D., Jr (2003) Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.*, **43**, 267–272.
- Panaye, A., Doucet, J.P. and Fan, B.T. (1993) Topological approach of ^{13}C NMR spectral simulation: application to fuzzy substructures. *J. Chem. Inf. Comput. Sci.*, **33**, 258–265.
- Panaye, A., MacPhee, J.A. and Dubois, J.-E. (1980) Steric effects. II. Relationship between topology and steric parameter E_s – topology as a tool for the correlation and prediction of steric effects. *Tetrahedron*, **36**, 759–768.
- Panek, J.J., Jezierska, A. and Vračko, M. (2005) Kohonen network study of aromatic compounds based on electronic and nonelectronic structure descriptors. *J. Chem. Inf. Model.*, **45**, 264–272.
- Paolini, J.P. (1990) The bond order–bond length relationship. *J. Comput. Chem.*, **11**, 1160–1163.
- Papa, E., Battaini, F. and Gramatica, P. (2005) Ranking of aquatic toxicity of esters modelled by QSAR. *Chemosphere*, **58**, 559–570.
- Papa, E., Castiglioni, S., Gramatica, P., Nikolayenko, V., Kayumov, O. and Calamari, D. (2004) Screening the leaching tendency of pesticides applied in the Amu Darya Basin (Uzbekistan). *Water Res.*, **38**, 3485–3494.
- Papa, E., Dearden, J.C. and Gramatica, P. (2007) Linear QSAR regression models for the prediction of bioconcentration factors by physico-chemical properties and structural theoretical molecular descriptors. *Chemosphere*, **67**, 351–358.
- Papa, E., Villa, F. and Gramatica, P. (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (fathead minnow). *J. Chem. Inf. Model.*, **45**, 1256–1266.
- Papadopoulos, M.C. and Dean, P.M. (1991) Molecular structure matching by simulated annealing. IV. Classification of atom correspondences in sets of dissimilar molecules. *J. Comput. Aid. Mol. Des.*, **5**, 119–133.
- Papp, Á., Gulyás-Forró, A., Gulyás, Z., Dormán, G., Ürge, L. and Darvas, F. (2006) Explicit diversity index (EDI): a novel measure for assessing the diversity of compound databases. *J. Chem. Inf. Model.*, **46**, 1898–1904.
- Paris, C.G. (2003) Databases of chemical structures, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 523–555.
- Pariser, R. and Parr, R.G. (1953a) A semi-empirical theory of the electronic spectra and electronic structure of complex unsaturated molecules. I. *J. Chim. Phys.*, **21**, 466–471.
- Pariser, R. and Parr, R.G. (1953b) A semi-empirical theory of the electronic spectra and electronic structure of complex unsaturated molecules. II. *J. Chim. Phys.*, **21**, 767–776.
- Park, D.-S., Grodnitzky, J.A. and Coats, J.R. (2002) QSAR evaluation of cyanohydrins' fumigation toxicity to house fly (*Musca domestica*) and lesser grain borer (*Rhyzopertha dominica*). *J. Agr. Food Chem.*, **50**, 5617–5620.
- Parr, R.G. and Pearson, R.G. (1983) Absolute hardness: companion parameter to absolute electronegativity. *J. Am. Chem. Soc.*, **105**, 7512–7516.
- Parr, R.G., Szentpály, L.V. and Liu, S. (1999) Electrophilicity index. *J. Am. Chem. Soc.*, **121**, 1922–1924.
- Parr, R.G. and Yang, W. (1984) Density functional approach to the frontier-electron theory of chemical reactivity. *J. Am. Chem. Soc.*, **106**, 4049–4050.
- Parr, R.G. and Yang, W. (1989) *Density-Functional Theory of Atoms and Molecules*, Oxford Science Publications, New York, p. 334.
- Parthasarathi, R., Subramanian, V. and Chattaraj, P.K. (2003) Effect of electric field on the global and local reactivity indices. *Chem. Phys. Lett.*, **382**, 48–56.
- Pascual, R., Borrell, J.I. and Teixidó, J. (2003) Analysis of selection methodologies for combinatorial library design. *Mol. Div.*, **6**, 121–133.

- Pascual, R., Mateu, M., Gasteiger, J., Borrell, J.I. and Teixidó, J. (2003) Design and analysis of a combinatorial library of HEPT analogues: comparison of selection methodologies and inspection of the actually covered chemical space. *J. Chem. Inf. Comput. Sci.*, **43**, 199–207.
- Pascual-Ahuir, J.L. and Silla, E. (1990) GEOL: an improved description of molecular surfaces. I. Building the spherical surface set. *J. Comput. Chem.*, **11**, 1047–1060.
- Pasha, F.A., Srivastava, H.K. and Singh, P.P. (2005) Semiempirical QSAR study and ligand receptor interaction of estrogens. *Mol. Div.*, **9**, 215–220.
- Pasti, L., Jouan-Rimbaud, D., Massart, D.L. and de Noord, O.E. (1998) Application of Fourier transform to multivariate calibration of near-infrared data. *Anal. Chim. Acta*, **364**, 253–263.
- Pastor, M. (2006) Alignment-independent descriptors from molecular interaction fields, in *Molecular Interaction Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 117–143.
- Pastor, M. and Alvarez-Builla, J. (1991) The EDISFAR programs. Rational drug series design. *Quant. Struct. -Act. Relat.*, **10**, 350–358.
- Pastor, M. and Alvarez-Builla, J. (1994) New developments of EDISFAR programs. Experimental design in QSAR practice. *J. Chem. Inf. Comput. Sci.*, **34**, 570–575.
- Pastor, M., Cruciani, G. and Clementi, S. (1997) Smart region definition: a new way to improve the predictive ability and interpretability of three-dimensional quantitative structure–activity relationships. *J. Med. Chem.*, **40**, 1455–1464.
- Pastor, M., Cruciani, G., McLay, I.M., Pickett, S.D. and Clementi, S. (2000) Grid-independent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.*, **43**, 3233–3243.
- Patani, G.A. and LaVoie, E.J. (1996) Bioisosterism: a rational approach in drug design. *Chem. Rev.*, **96**, 3147–3176.
- Patankar, S.J. and Jurs, P.C. (2000) Prediction of IC₅₀ values for ACAT inhibitors from molecular structure. *J. Chem. Inf. Comput. Sci.*, **40**, 706–723.
- Patankar, S.J. and Jurs, P.C. (2002) Prediction of glycine/NMDA receptor antagonist inhibition from molecular structure. *J. Chem. Inf. Comput. Sci.*, **42**, 1053–1068.
- Patankar, S.J. and Jurs, P.C. (2003a) Classification of HIV protease inhibitors on the basis of their antiviral potency using radial basis function neural networks. *J. Comput. Aid. Mol. Des.*, **17**, 155–171.
- Patankar, S.J. and Jurs, P.C. (2003b) Classification of inhibitors of protein tyrosine phosphatase 1B using molecular structure based descriptors. *J. Chem. Inf. Comput. Sci.*, **43**, 885–899.
- Patel, H. and Cronin, M.T.D. (2001) A novel index for the description of molecular linearity. *J. Chem. Inf. Comput. Sci.*, **41**, 1228–1236.
- Patel, H., Schultz, T.W. and Cronin, M.T.D. (2002) Physico-chemical interpretation and prediction of the dimyristoyl phosphatidyl choline–water partition coefficient. *J. Mol. Struct. (Theochem)*, **593**, 9–18.
- Patel, H., ten Berge, W. and Cronin, M.T.D. (2002) Quantitative structure–activity relationships (QSARs) for the prediction of skin permeation of exogenous chemicals. *Chemosphere*, **48**, 603–613.
- Patel, H.C., Duca, J.S., Hopfinger, A.J., Glendening, C.D. and Thompson, E.D. (1999) Quantitative component analysis of mixtures for risk assessment: application to eye irritation. *Chem. Res. Toxicol.*, **12**, 1050–1056.
- Patil, G.S., Bora, M. and Dutta, N.N. (1995) Empirical correlations for prediction of permeability of gases liquids through polymers. *J. Memb. Sci.*, **101**, 145–152.
- Patlewicz, G., Aptula, A.O., Roberts, D.W. and Uriarte, E. (2008) A minireview of available skin sensitization (Q)SARs/expert systems. *QSAR Comb. Sci.*, **27**, 60–76.
- Patrinos, A.N. and Hakimi, S.L. (1973) The distance matrix of a graph and its tree realization. *Quarterly Applied Mathematics*, **30**, 255–269.
- Pattarino, F., Marengo, E., Trotta, M. and Gasco, M.R. (2000) Combined use of lecithin and decyl polyglucoside in microemulsions: domain of existence and cosurfactant effect. *J. Disp. Sci. Technol.*, **21**, 345–363.
- Patte, F., Etcheto, M. and Laffort, P. (1982) Solubility factors for 240 solutes and 207 stationary phases in gas–liquid chromatography. *Anal. Chem.*, **54**, 2239–2247.
- Patterson, D.E., Cramer, R.D., III, Ferguson, A.M., Clark, R.D. and Weinberger, L.E. (1996) Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.*, **39**, 3049–3059.
- Pauling, L. (1932) The additivity of the energies of normal covalent bonds. *Proc. Nat. Acad. Sci. USA*, **14**, 414–416.
- Pauling, L. (1936) The diamagnetic anisotropy of aromatic molecules. *J. Chim. Phys.*, **4**, 673–677.
- Pauling, L. (1939) *The Nature of the Chemical Bond*, Cornell University Press, Ithaca, NY.

- Pauling, L. (1947) Atomic radii and interatomic distances in metals. *J. Am. Chem. Soc.*, **69**, 542–553.
- Pauling, L., Brockway, L.O. and Beach, J.Y. (1935) The dependence of interatomic distance on single bond–double bond resonance. *J. Am. Chem. Soc.*, **57**, 2705–2709.
- Pauling, L. and Kamb, B. (1986) A revised set of values of single-bond radii derived from the observed interatomic distances in metals by correction for bond number and resonance energy. *Proc. Nat. Acad. Sci. USA*, **83**, 3569–3571.
- Pauling, L. and Pressman, D. (1945) The serological properties of simple substances. IX. Hapten inhibition of precipitation of antisera homologous to the *o*-, *m*-, and *p*-azophenylarsonic acid groups. *J. Am. Chem. Soc.*, **67**, 1003–1012.
- Pauling, L. and Sherman, J. (1933) The nature of the chemical bond. VI. The calculation from thermochemical data of the energy of resonance of molecules among several electronic structures. *J. Chim. Phys.*, **1**, 606–617.
- Pauling, L. and Wheland, G.W. (1933) The nature of the chemical bond. V. The quantum-mechanical calculation of the resonance energy of benzene and naphthalene and hydrocarbon free radicals. *J. Chim. Phys.*, **1**, 362–374.
- Pauling, L. and Wilson, E.B. (1935) *Introduction to Quantum Mechanics*, McGraw-Hill, New York.
- Pavan, M., Consonni, V. and Todeschini, R. (2005) Partial ranking models by genetic algorithms variable subset selection (GA-VSS) approach for environmental priority settings. *MATCH Commun. Math. Comput. Chem.*, **54**, 583–609.
- Pavan, M., Mauri, A. and Todeschini, R. (2004) Total ranking models by the genetic algorithms variable subset selection (GA-VSS) approach for environmental priority settings. *Anal. Bioanal. Chem.*, **380**, 430–444.
- Pavan, M., Netzeva, T.I. and Worth, A.P. (2008) Review of literature-based quantitative structure–activity relationship models for bioconcentration. *QSAR Comb. Sci.*, **27**, 21–31.
- Pavan, M. and Todeschini, R. (2004) New indices for analyzing partial ranking diagrams. *Anal. Chim. Acta*, **515**, 167–181.
- Pavan, M. and Todeschini, R. (eds) (2008) *Scientific Data Ranking Methods: Theory and Applications*. Elsevier, Amsterdam, The Netherlands, 214.
- Pavan, M. and Worth, A.P. (2008) Review of estimation models for biodegradation. *QSAR Comb. Sci.*, **27**, 32–40.
- Pavani, R. and Ranghino, G. (1982) A method to compute the volume of a molecule. *Computers Chem.*, **6**, 133–135.
- Pavlikova, M., Lacko, I., Devinsky, F. and Mlynarcik, D. (1995) Quantitative relationships between structure, aggregation properties and antimicrobial activity of quaternary ammonium bolaamphiphiles. *Collect. Czech. Chem. Comm.*, **60**, 1213–1228.
- Pavlović, L. and Gutman, I. (1997) Wiener numbers of phenylenes: an exact result. *J. Chem. Inf. Comput. Sci.*, **37**, 355–358.
- Payares, P., Díaz, D., Olivero, J., Vivas, R. and Gómez, I. (1997) Prediction of the gas chromatographic relative retention times of flavonoids from molecular structure. *J. Chromat.*, **771**, 213–219.
- Pearlman, R.S. (1980) Molecular surface areas and volumes and their use in structure/activity relationships, in *Physical Chemical Properties of Drugs* (eds S.H. Yalkowsky, A.A. Sinkula and S.C. Valvani), Marcel Dekker, New York, pp. 321–347.
- Pearlman, R.S. (1993) 3D molecular structures: generation and use in 3D searching, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 41–79.
- Pearlman, R.S. (1999) Novel software tools for addressing chemical diversity. Internet communication <http://www.netsci.org/Science/Combichem/feature08.html>.
- Pearlman, R.S. and Smith, K.M. (1998) Novel software tools for chemical diversity, in *3D QSAR in Drug Design*, Vol. 2 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 339–353.
- Pearlman, R.S. and Smith, K.M. (1999) Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.*, **39**, 28–35.
- Pearson, K. (1920) Notes on the history of correlation. *Biometrika*, **13**, 25–45.
- Pearson, R.G. (1997) *Chemical Hardness*, Wiley-VCH Verlag GmbH, Weinheim, Germany, 198.
- Pednekar, D.V., Kelkar, M.A., Pimple, S.R. and Akamanchi, K.G. (2004) 3D QSAR studies of inhibitors of epidermal growth factor receptor (EGFR) using CoMFA and GFA methodologies. *Med. Chem. Res.*, **13**, 605–618.
- Peijnenburg, W.J.G.M., Debeer, K.G., Dehaan, M.W., Denhollander, H.A., Stegeman, M.H. and Verboom, H.H. (1992) Development of a structure–reactivity relationship for the photohydrolysis of substituted aromatic halides. *Environ. Sci. Technol.*, **26**, 2116–2121.
- Peijnenburg, W.J.G.M., Debeer, K.G., Denhollander, H.A., Stegeman, M.H. and Verboom, H.H. (1993) Kinetics, products, mechanisms and QSARs for the hydrolytic transformation of aromatic nitriles

- in anaerobic sediment slurries. *Environ. Toxicol. Chem.*, **12**, 1149–1161.
- Peijnenburg, W.J.G.M., Thart, M.J., Denhollander, H. A., Vandemeent, D., Verboom, H.H. and Wolfe, N. L. (1992) QSARs for predicting reductive transformation rate constants of halogenated aromatic hydrocarbons in anoxic sediment systems. *Environ. Toxicol. Chem.*, **11**, 301–314.
- Pellegrin, V. (1983) Molecular formulas of organic compounds. The nitrogen rule and degree of unsaturation. *J. Chem. Educ.*, **60**, 626–633.
- Pelletier, D.J., Gehlhaar, D., Tilloy-Ellul, A., Johnson, T.O. and Greene, N. (2007) Evaluation of a published *in silico* model and construction of a novel Bayesian model for predicting phospholipidosis inducing potential. *J. Chem. Inf. Model.*, **47**, 1196–1205.
- Peltason, L. and Bajorath, J. (2007) SAR index: quantifying the nature of structure–activity relationships. *J. Med. Chem.*, **50**, 5571–5578.
- Peng, X.-L., Fang, K.-T., Hu, Q.-N. and Liang, Y.-Z. (2004) Impersonality of the connectivity index and recomposition of topological indices according to different properties. *Molecules*, **9**, 1089–1099.
- Pepperrell, C.A. (1994) *Three-Dimensional Chemical Similarity Searching*, Research Studies Press–Wiley, Taunton, UK, p. 304.
- Pepperrell, C.A. and Willett, P. (1991) Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. *J. Comput. Aid. Mol. Des.*, **5**, 455–474.
- Perdih, A. (2000a) On topological indices indicating branching. Part 1. The principal component analysis of alkane properties and indices. *Acta Chim. Sloven.*, **47**, 231–259.
- Perdih, A. (2000b) On topological indices indicating branching. Part 2. The suitability of some physico-chemical properties of alkanes as references to assess branching. *Acta Chim. Sloven.*, **47**, 293–316.
- Perdih, A. (2000c) On topological indices indicating branching. Part 3. Assessment of some indices for their suitability to represent branching. *Acta Chim. Sloven.*, **47**, 435–452.
- Perdih, A. (2003) Towards branching indices. *Indian J. Chem.*, **42A**, 1246–1257.
- Perdih, A. and Perdih, B. (2002a) Some topological indices derived from the $v^m d^n$ matrix. Part 1. Wiener like indices of BI_M type. *Acta Chim. Sloven.*, **49**, 67–110.
- Perdih, A. and Perdih, B. (2002b) Some topological indices derived from the $v^m d^n$ matrix. Part 2. The “mean degree of vertices” summation-derived indices of BI_M type. *Acta Chim. Sloven.*, **49**, 291–308.
- Perdih, A. and Perdih, B. (2002c) Some topological indices derived from the $v^m d^n$ matrix. Part 3. The largest eigenvalues of the $v^m d^n$ matrix as topological indices of the BI_M -type. *Acta Chim. Sloven.*, **49**, 309–330.
- Perdih, A. and Perdih, B. (2002d) Some topological indices derived from the $v^m d^n$ matrix. Part 4. The largest eigenvalues of the “mean degree of vertices” matrices as topological indices of the BI_M -type. *Acta Chim. Sloven.*, **49**, 467–482.
- Perdih, A. and Perdih, B. (2002e) Some topological indices derived from the $v^m d^n$ matrix. Part 5. Summation-derived susceptibilities for branching as BI_A type indices. *Acta Chim. Sloven.*, **49**, 497–514.
- Perdih, A. and Perdih, B. (2003a) On the structural interpretation of topological indices. *Indian J. Chem.*, **42**, 1219–1226.
- Perdih, A. and Perdih, B. (2003b) Some topological indices derived from the $v^m d^n$ matrix. Part 6. Summation-derived difference type indices of BI_A class. *Acta Chim. Sloven.*, **50**, 83–94.
- Perdih, A. and Perdih, B. (2003c) Some topological indices derived from the $v^m d^n$ matrix. Part 7. The $V_{ij}(m,n)$ indices. *Acta Chim. Sloven.*, **50**, 95–114.
- Perdih, A. and Perdih, B. (2003d) Some topological indices derived from the $v^m d^n$ matrix. Part 8. the $I_{ij}(m,n)$ indices. *Acta Chim. Sloven.*, **50**, 161–184.
- Perdih, A. and Perdih, B. (2003e) Some topological indices derived from the $v^m d^n$ matrix. Part 9. The $M_j(m,n)$ and $M_{ij}(m,n)$ indices. *Acta Chim. Sloven.*, **50**, 513–538.
- Perdih, A. and Perdih, B. (2004) Topological indices derived from the $G(a,b,c)$ matrix, useful as physico-chemical property indices. *Acta Chim. Sloven.*, **51**, 589–609.
- Pérez González, M., Gonzalez, H., Molina Ruiz, R., Cabrera, M.A. and Ramos de Armas, R. (2003) TOPS-MODE based QSARs derived from heterogeneous series of compounds. Applications to the design of new herbicides. *J. Chem. Inf. Comput. Sci.*, **43**, 1192–1199.
- Pérez González, M., Helguera Morales, A. and Collado, I.G. (2006) A topological substructural molecular design to predict soil sorption coefficients for pesticides. *Mol. Div.*, **10**, 109–118.
- Pérez González, M., Helguera Morales, A. and Diaz, H.G. (2004) A TOPS-MODE approach to predict permeability coefficients. *Polymer*, **45**, 2073–2079.
- Perdih, A. and Perdih, B. (2003) TOPS-MODE versus DRAGON descriptors to predict permeability coefficients through low-density polyethylene. *J. Comput. Aid. Mol. Des.*, **17**, 665–672.

- Pérez González, M., Helguera, A.M. and Rodriguez, Y.M. (2004) TOPS-MODE and DRAGON descriptors in QSAR. 1. Skin permeation. *Internet Electron. J. Mol. Des.*, **3**, 750–758.
- Pérez González, M. and Moldes Teran, M.d.C. (2004) A TOPS-MODE approach to predict adenosine kinase inhibition. *Bioorg. Med. Chem. Lett.*, **14**, 3077–3079.
- Pérez González, M., Suarez, P.L., Fall, Y. and Gomez, G. (2005) Quantitative structure–activity relationship studies of vitamin D receptor affinity for analogues of 1 α ,25-dihydroxyvitamin D₃. 1. WHIM descriptors. *Bioorg. Med. Chem. Lett.*, **15**, 5165–5169.
- Pérez González, M., Terán, C., Teijeira, M. and Besada, P. (2005a) Geometry, topology, and atom-weights assembly descriptors to predicting A₁ adenosine receptors agonists. *Bioorg. Med. Chem. Lett.*, **15**, 2641–2645.
- Pérez González, M., Terán, C., Teijeira, M. and Besada, P. (2006) Geometry, topology, and atom-weights assembly descriptors to predicting A₁ adenosine receptors agonists. *Bioorg. Med. Chem. Lett.*, **15**, 2641–2645.
- Pérez González, M., Terán, C., Teijeira, M. and Gonzalez-Moa, M.J. (2005b) GETAWAY descriptors to predicting A_{2A} adenosine receptors agonists. *Eur. J. Med. Chem.*, **40**, 1080–1086.
- Pérez González, M., Toropov, A.A., Duchowicz, P.R. and Castro, E.A. (2004) QSPR calculation of normal boiling points of organic molecules based on the use of correlation weighting of atomic orbitals with extended connectivity of zero- and first-order graphs of atomic orbitals. *Molecules*, **9**, 1019–1033.
- Pérez, C., Pastor, M., Ortiz, A.R. and Gago, F. (1998) Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Med. Chem.*, **41**, 836–852.
- Perez, J.J. (2005) Managing molecular diversity. *Chem. Soc. Rev.*, **34**, 143–152.
- Pérez, P. and Contreras, R. (1998) A theoretical analysis of the gas-phase protonation of hydroxylamine, methyl-derivatives and aliphatic amino acids. *Chem. Phys. Lett.*, **293**, 239–244.
- Perez-Gimenez, F., Antón-Fos, G.M., García-March, F.J., Salabert-Salvador, M.T., Cercos-del-Pozo, R.A. and Jaenoltra, J. (1995) Prediction of chromatographic parameters for some anilines by molecular connectivity. *Chromatographia*, **41**, 167–174.
- Perrin, D.D., Dempsey, B. and Serjeant, E.P. (1981) *pKa Prediction for Organic Acids and Bases*, Chapman & Hall, London, UK.
- Perruccio, F., Mason, J.S., Sciabola, S. and Baroni, M. (2006) FLAP: 4-point pharmacophore fingerprints from GRID, in *Molecular Interaction Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany.
- Peruzzo, P.J., Marino, D.J.G., Castro, E.A. and Toropov, A.A. (2001) Calculation of pK values of flavylum salts from the optimization of correlation weights of local graph invariants. *J. Mol. Struct. (Theochem)*, **572**, 53–60.
- Peterson, D.L. and Yalkowsky, S.H. (2001) Comparison of two methods for predicting aqueous solubility. *J. Chem. Inf. Comput. Sci.*, **41**, 1531–1534.
- Petitjean, M. (1992) Applications of the radius–diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **32**, 331–337.
- Petitjean, M. (1996) Three-dimensional pattern recognition from molecular distance minimization. *J. Chem. Inf. Comput. Sci.*, **36**, 1038–1049.
- Petitjean, M. (2004) From shape similarity to shape complementarity: toward a docking theory. *J. Math. Chem.*, **35**, 147–158.
- Petitjean, M. and Dubois, J.-E. (1990) Topological statistics on a large structural file. *J. Chem. Inf. Comput. Sci.*, **30**, 332–343.
- Petrauskas, A. and Kolovanov, E.A. (2000) ACD/log P method description. *Persp. Drug Disc. Des.*, **19**, 99–116.
- Piazza, R., Pino, A., Marchini, S., Passerini, L., Chiorboli, C. and Tosato, M.L. (1995) Modelling physico-chemical properties of halogenated benzenes: QSAR optimization through variables selection. *SAR & QSAR Environ. Res.*, **4**, 59–71.
- Pickett, S.D., Luttmann, C., Guerin, V., Laoui, A. and James, E. (1998) DIVSEL and COMPLIB – strategies for the design and comparison of combinatorial libraries using pharmacophoric descriptors. *J. Chem. Inf. Comput. Sci.*, **38**, 144–150.
- Pickett, S.D., Mason, J.S. and McLay, I.M. (1996) Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.*, **36**, 1214–1223.
- Pickett, S.D., McLay, I.M. and Clark, D.E. (2000) Enhancing the hit-to-lead properties of lead optimization libraries. *J. Chem. Inf. Comput. Sci.*, **40**, 263–272.
- Piclin, N., Pintore, M., Wechman, C. and Chrétien, J. R. (2004) Classification of a large anticancer data set by adaptive fuzzy partition. *J. Comput. Aid. Mol. Des.*, **18**, 577–586.
- Pidgeon, C., Ong, S., Liu, H., Pidgeon, M., Dantzig, A., Munroe, J., Hornback, W., Kasher, J.S., Glunz,

- L. and Szerba, T. (1995) IAM chromatography: an *in vitro* screen for predicting drug membrane permeability. *J. Med. Chem.*, **38**, 590–594.
- Piggott, J.R. and Withers, S.J. (1993) Modern statistics and quantitative structure–activity relationships in flavor. *ACS Symp. Ser.*, **528**, 100–108.
- Pimentel, G.C. and McClellan, A.L. (1960) *The Hydrogen Bond*, Freeman, San Francisco, CA, p. 575.
- Pimple, S.R., Kelkar, M.A., Pednekar, D.V. and Akamanchi, K.G. (2004) Studies on novel non-imidazole human H_4 receptor antagonists using GFA and Free–Wilson analysis. *Med. Chem. Res.*, **13**, 619–630.
- Pinheiro, A.A.C., Borges, R.S., Santos, L.S. and Alves, C.N. (2004) A QSAR study of 8.O.4'-neolignans with antifungal activity. *J. Mol. Struct. (Theochem)*, **672**, 215–219.
- Pinheiro, J.C., Ferreira, M.M.C. and Romero, O.A.S. (2001) Antimalarial activity of dihydroartemisinin derivatives against *P. falciparum* resistant to mefloquine: a quantum chemical and multivariate study. *J. Mol. Struct. (Theochem)*, **572**, 35–44.
- Pinheiro, J.C., Kiralj, R. and Ferreira, M.M.C. (2003) Artemisinin derivatives with antimalarial activity against *Plasmodium falciparum* designed with the aid of quantum chemical and partial least squares methods. *QSAR Comb. Sci.*, **22**, 830–842.
- Pino, A., Giuliani, A. and Benigni, R. (2003) Toxicity mode-of-action: discrimination via infrared spectra and eigenvalues of the modified adjacency matrix. *Quant. Struct. -Act. Relat.*, **22**, 1–5.
- Pinto, M.F.S., Romero, O.A.S. and Pinheiro, J.C. (2001) Pattern recognition study of structure–activity relationship of halophenols and halonitrophenols against fungus *T. mentagrophytes*. *J. Mol. Struct. (Theochem)*, **539**, 303–310.
- Pintore, M., Piclin, N., Benfenati, E., Gini, G. and Chrétien, J.R. (2003) Database mining with adaptive fuzzy partition: application to the prediction of pesticide toxicity on rats. *Environ. Toxicol. Chem.*, **22**, 983–991.
- PipeLine Pilot, Ver. 4.5.2, Scitegic, Inc., 9665 Chesapeake Dr., Suite 401, San Diego, CA.
- Pirard, B. and Pickett, S.D. (2000) Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Comput. Sci.*, **40**, 1431–1440.
- Pires, J.M., Floriano, W.B. and Gaudio, A.C. (1997) Extension of the frontier reactivity indices to groups of atoms and application to quantitative structure–activity relationship studies. *J. Mol. Struct. (Theochem)*, **389**, 159–167.
- Piruzyan, L.A., Kabankin, A.S., Gabrielyan, L.I., Ostapchuk, N.V., Pyn'ko, N.E. and Radkevich, L.A. (2004) Hepatoprotector effect and relationship between structure and detoxicant activity of adamantane derivatives. *Pharm. Chem. J.*, **38**, 136–142.
- Pis Diez, R., Duchowicz, P.R., Castañeta, H., Castro, E.A., Fernández, F.M. and Albesa, A.G. (2006) A theoretical study of a family of new quinoxaline derivatives. *J. Mol. Graph. Model.*, **25**, 487–494.
- Pisanski, T., Plavšić, D. and Randić, M. (2000) On numerical characterization of cyclicity. *J. Chem. Inf. Comput. Sci.*, **40**, 520–523.
- Pisanski, T. and Žerovnik, J. (1994) Weights on edges of chemical graphs determined by paths. *J. Chem. Inf. Comput. Sci.*, **34**, 395–397.
- Pissurlenkar, R.R.S., Malde, A.K., Khedkar, S.A. and Coutinho, E.C. (2007) Encoding type and position in peptide QSAR: application to peptides binding to class I MHC molecule HLA-A0201. *QSAR Comb. Sci.*, **26**, 189–203.
- Pitman, M.C., Huber, W.K., Horn, H., Krämer, A., Rice, J.E. and Swope, W.C. (2001) FLASHFLOOD: a 3D field-based similarity search and alignment method for flexible molecules. *J. Comput. Aid. Mol. Des.*, **15**, 587–612.
- Pitzer, K.S. (1940) The vibration frequencies and thermodynamic functions of long chain hydrocarbons. *J. Chim. Phys.*, **8**, 711–720.
- Pitzer, K.S. (1955) The volumetric and thermodynamic properties of fluids. I. Theoretical basis and virial coefficients. *J. Am. Chem. Soc.*, **77**, 3427–3433.
- Pitzer, K.S., Lippmann, D.Z., Curl, R.F., Huggins, C. M. and Peterson, D.E. (1955) The volumetric and thermodynamic properties of fluids. II. compressibility factor, vapor pressure and entropy of vaporization. *J. Am. Chem. Soc.*, **77**, 3433–3440.
- Pitzer, K.S. and Scott, D.W. (1941) The thermodynamics of branched-chain paraffins. The heat capacity, heat of fusion and vaporization, and entropy of 2,3,4-trimethylpentane. *J. Am. Chem. Soc.*, **63**, 2419–2422.
- Pixner, P., Heiden, W., Merx, H., Moeckel, G., Möller, A. and Brickmann, J. (1994) Empirical method for the quantification and localization of molecular hydrophobicity. *J. Chem. Inf. Comput. Sci.*, **34**, 1309.
- Pizarro Millán, C., Forina, M., Casolino, C. and Leardi, R. (1998) Extraction of representative subsets by potential functions method and genetic algorithms. *Chemom. Intell. Lab. Syst.*, **40**, 33–52.
- Plass, M., Valkó, K. and Abraham, M.H. (1998) Determination of solute descriptors of tripeptide derivatives based on high-throughput gradient high-performance liquid chromatography retention data. *J. Chromat.*, **803**, 51–60.

- Platt, D.E. and Silverman, B.D. (1996) Registration, orientation and similarity of molecular electrostatic potentials through multipole matching. *J. Comput. Chem.*, **17**, 358–366.
- Platt, J.R. (1947) Influence of neighbor bonds on additive bond properties in paraffins. *J. Chim. Phys.*, **15**, 419–420.
- Platt, J.R. (1952) Prediction of isomeric differences in paraffin properties. *J. Phys. Chem.*, **56**, 328–336.
- Platt, J.R. (1954) The box model and electron densities in conjugated systems. *J. Chim. Phys.*, **22**, 1448–1455.
- Platts, J.A. (2000a) Theoretical prediction of hydrogen bond basicity. *Phys. Chem. Chem. Phys.*, **2**, 3115–3120.
- Platts, J.A. (2000b) Theoretical prediction of hydrogen bond donor capacity. *Phys. Chem. Chem. Phys.*, **2**, 973–980.
- Platts, J.A., Abraham, M.H., Butina, D. and Hersey, A. (2000) Estimation of molecular linear free energy relationship descriptors by a group contribution approach. 2. Prediction of partition coefficients. *J. Chem. Inf. Comput. Sci.*, **40**, 71–80.
- Platts, J.A., Butina, D., Abraham, M.H. and Hersey, A. (1999) Estimation of molecular linear free energy relation descriptors using a group contribution approach. *J. Chem. Inf. Comput. Sci.*, **39**, 835–845.
- Plavšić, D. (1999) On the definition and calculation of the molecular descriptor R'/R . *Chem. Phys. Lett.*, **304**, 111–116.
- Plavšić, D. and Graovac, A. (2001) On calculation of molecular descriptors based on various graphical bond orders, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 39–61.
- Plavšić, D., Lerš, N. and Sertic-Bionda, K. (2000) On the relation between W'/W index, hyper-Wiener index, and Wiener number. *J. Chem. Inf. Comput. Sci.*, **40**, 516–519.
- Plavšić, D., Nikolić, S., Trinajstić, N. and Klein, D.J. (1993a) Relation between the Wiener index and the Schultz index for several classes of chemical graphs. *Croat. Chem. Acta*, **66**, 345–353.
- Plavšić, D., Nikolić, S., Trinajstić, N. and Mihalić, Z. (1993b) On the Harary index for the characterization of chemical graphs. *J. Math. Chem.*, **12**, 235–250.
- Plavšić, D., Šoškić, M., Daković, Z., Gutman, I. and Graovac, A. (1997) Extension of the Z matrix to cycle-containing and edge-weighted molecular graphs. *J. Chem. Inf. Comput. Sci.*, **37**, 529–534.
- Plavšić, D., Šoškić, M., Landeka, I., Gutman, I. and Graovac, A. (1996a) On the relation between the path numbers ${}^1Z$, ${}^2Z$ and the Hosoya Z index. *J. Chem. Inf. Comput. Sci.*, **36**, 1118–1122.
- Plavšić, D., Šoškić, M., Landeka, I. and Trinajstić, N. (1996b) On the relation between the P'/P index and the Wiener number. *J. Chem. Inf. Comput. Sci.*, **36**, 1123–1126.
- Plavšić, D., Šoškić, M. and Lerš, N. (1998) On the calculation of the molecular descriptor χ'/χ . *J. Chem. Inf. Comput. Sci.*, **38**, 889–892.
- Plavšić, D., Trinajstić, N., Amić, D. and Šoškić, M. (1998) Comparison between the structure–boiling point relationships with different descriptors for condensed benzenoids. *New J. Chem.*, **22**, 1075–1078.
- Pleiss, M.A. and Grunewald, G.L. (1983) An extension of the f-fragment method for the calculation of hydrophobic constants ($\log P$) of conformationally defined systems. *J. Med. Chem.*, **26**, 1760–1764.
- Plesnik, J. (1984) On the sum of all distances in a graph or digraph. *J. Graph Theory*, **8**, 1–21.
- Pliška, V., Testa, B. and van de Waterbeemd, H. (eds) (2008) *Lipophilicity in Drug Action and Toxicology*, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 438.
- Plummer, E.L. (1995) Successful application of the QSAR paradigm in discovery programs. *ACS Symp. Ser.*, **606**, 240–253.
- Poater, J., Fradera, X., Duran, M. and Solà, M. (2003a) An insight into the local aromaticities of polycyclic aromatic hydrocarbons and fullerenes. *Chem. Eur. J.*, **9**, 1113–1122.
- Poater, J., Fradera, X., Duran, M. and Solà, M. (2003b) The delocalization index as an electronic aromaticity criterion: application to a series of planar polycyclic aromatic hydrocarbons. *Chem. Eur. J.*, **9**, 400–406.
- Poater, J., Garcia-Cruz, I., Illas, F. and Solà, M. (2004) Discrepancy between common local aromaticity measures in a series of carbazole derivatives. *Phys. Chem. Chem. Phys.*, **6**, 314–318.
- Podlipnik, C. and Koller, J. (2001) Fast evaluation of molecular 3D shape similarity. *Acta Chim. Sloven.*, **48**, 325–331.
- Podlipnik, C., Solmajer, T. and Koller, J. (2003) Lipophilic connectivity indices. *MATCH Commun. Math. Comput. Chem.*, **49**, 7–14.
- Podlipnik, C., Solmajer, T. and Koller, J. (2006) Similarity of radial distribution function's intervals. *MATCH Commun. Math. Comput. Chem.*, **56**, 261–270.
- Pogliani, L. (1992a) Molecular connectivity model for determination of isoelectric point of amino acids. *J. Pharm. Sci.*, **81**, 334–336.
- Pogliani, L. (1992b) Molecular connectivity: treatment of electronic structure of amino acids. *J. Pharm. Sci.*, **81**, 967–969.

- Pogliani, L. (1993a) Molecular connectivity model for determination of physico-chemical properties of α -amino acids. *J. Phys. Chem.*, **97**, 6731–6736.
- Pogliani, L. (1993b) Molecular connectivity model for determination of T_1 relaxation times of α -carbons of amino acids and cyclic dipeptides. *Computers Chem.*, **17**, 283–286.
- Pogliani, L. (1994a) Molecular connectivity descriptors of the physico-chemical properties of the α -amino acids. *J. Phys. Chem.*, **98**, 1494–1499.
- Pogliani, L. (1994b) On a graph theoretical characterization of *cis/trans* isomers. *J. Chem. Inf. Comput. Sci.*, **34**, 801–804.
- Pogliani, L. (1994c) Structure–property relationships of amino acids and some dipeptides. *Amino Acids*, **6**, 141–153.
- Pogliani, L. (1995a) Modeling the solubility and activity of amino acids with the LCCI method. *Amino Acids*, **9**, 217–228.
- Pogliani, L. (1995b) Molecular modeling by linear combinations of connectivity indexes. *J. Phys. Chem.*, **99**, 925–937.
- Pogliani, L. (1996a) A strategy for molecular modeling of a physico-chemical property using a linear combination of connectivity indexes. *Croat. Chem. Acta*, **69**, 95–109.
- Pogliani, L. (1996b) Modeling purines and pyrimidines with the linear combination of connectivity indices–molecular connectivity “LCCI-MC” method. *J. Chem. Inf. Comput. Sci.*, **36**, 1082–1091.
- Pogliani, L. (1996c) Modeling with special descriptors derived from a medium-sized set of connectivity indices. *J. Phys. Chem.*, **100**, 18065–18077.
- Pogliani, L. (1997a) Modeling biochemicals with leading molecular connectivity terms. *Med. Chem. Res.*, **7**, 380–393.
- Pogliani, L. (1997b) Modeling enthalpy and hydration properties of inorganic compounds. *Croat. Chem. Acta*, **70**, 803–817.
- Pogliani, L. (1997c) Modeling properties of biochemical compounds with connectivity terms. *Amino Acids*, **13**, 237–255.
- Pogliani, L. (1999a) Modeling properties with higher-level molecular connectivity descriptors. *J. Chem. Inf. Comput. Sci.*, **39**, 104–111.
- Pogliani, L. (1999b) Modeling with semiempirical molecular connectivity terms. *J. Phys. Chem. A*, **103**, 1598–1610.
- Pogliani, L. (1999c) Properties of molecular connectivity terms and physico-chemical properties. *J. Mol. Struct. (Theochem)*, **466**, 1–19.
- Pogliani, L. (2000a) From molecular connectivity indices to semiempirical connectivity terms: recent trends in graph theoretical descriptors. *Chem. Rev.*, **100**, 3827–3858.
- Pogliani, L. (2000b) Modeling with molecular pseudoconnectivity descriptors. A useful extension of the intrinsic *I*-state concept. *J. Phys. Chem. A*, **104**, 9029–9045.
- Pogliani, L. (2001a) How far are molecular connectivity descriptors from I_s molecular pseudoconnectivity descriptors? *J. Chem. Inf. Comput. Sci.*, **41**, 836–847.
- Pogliani, L. (2001b) The concept of graph mass in molecular graph theory. A case in data reduction analysis, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science, Huntington, NY, pp. 109–146.
- Pogliani, L. (2002a) Algorithmically compressed data and the topological conjecture for the inner-core electrons. *J. Chem. Inf. Comput. Sci.*, **42**, 1028–1042.
- Pogliani, L. (2002b) Higher-level descriptors in molecular connectivity. *Croat. Chem. Acta*, **75**, 409–432.
- Pogliani, L. (2002c) Topics in molecular modeling: dual indices, quality of modeling and missing information, truncation. *J. Mol. Struct. (Theochem)*, **581**, 87–109.
- Pogliani, L. (2003a) Introducing the complete graphs for the inner-core electrons. *Indian J. Chem.*, **42**, 1347–1353.
- Pogliani, L. (2003b) Model with dual indices and complete graphs. The heterogeneous description of the dipole moments and polarizabilities. *New J. Chem.*, **27**, 919–927.
- Pogliani, L. (2004) Modeling with indices obtained from complete graphs. *Croat. Chem. Acta*, **77**, 193–201.
- Pogliani, L. (2005a) A chemical graph model study of the partition coefficient of halogenated carbocompounds. *Croat. Chem. Acta*, **78**, 189–194.
- Pogliani, L. (2005b) Model of the physical properties of halides with complete graph-based indices. *Int. J. Quant. Chem.*, **102**, 38–52.
- Pogliani, L. (2006a) The evolution of the valence delta in molecular connectivity theory. *Internet Electron. J. Mol. Des.*, **5**, 364–375.
- Pogliani, L. (2006b) The hydrogen perturbation in molecular connectivity computations. *J. Comput. Chem.*, **27**, 868–882.
- Pogrebnyak, A.V., Oganessian, E.T. and Glushko, A.A. (2002) MATRIX, a new algorithm for predicting biological activity of organic molecules based on multidimensional analysis of physico-chemical descriptors of modern pharmaceuticals. I. General principles. *Russ. J. Org. Chem.*, **38**, 1564–1575.

- Poincaré, H. (1900) Second complément à l'Analysis Situs. *Proc. London Math. Soc.*, **32**, 277–308.
- Poirrette, A.R., Artymiuk, P.J., Rice, D.W. and Willett, P. (1997) Comparison of protein surfaces using a genetic algorithm. *J. Comput. Aid. Mol. Des.*, **11**, 557–569.
- Polanski, J. (1997) The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.*, **37**, 553–561.
- Polanski, J. (2003) Molecular shape analysis, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 302–319.
- Polanski, J., Bak, A., Gieleciak, R. and Magdziarz, T. (2004) Self-organizing neural networks for modeling robust 3D and 4D QSAR: application to dihydrofolate reductase inhibitors. *Molecules*, **9**, 1148–1159.
- Polanski, J. and Bonchev, D. (1986a) The minimum distance number of trees. *MATCH Commun. Math. Comput. Chem.*, **21**, 341–344.
- Polanski, J. and Bonchev, D. (1986b) The Wiener number of graphs. I. General theory and changes due to graph operations. *MATCH Commun. Math. Comput. Chem.*, **21**, 133–186.
- Polanski, J. and Bonchev, D. (1990) Theory of the Wiener number of graphs. II. Transfer graphs and some of their metric properties. *MATCH Commun. Math. Comput. Chem.*, **25**, 3–39.
- Polanski, J., Gasteiger, J., Wagener, M. and Sadowski, J. (1998) The comparison of molecular surfaces by neural networks and its applications to quantitative structure–activity studies. *Quant. Struct. -Act. Relat.*, **17**, 27–36.
- Polanski, J. and Gieleciak, R. (2003) The comparative molecular surface analysis (CoMSA) with modified uniform variable elimination-PLS (UVE-PLS) method: application to the steroids binding the aromatase enzyme. *J. Chem. Inf. Comput. Sci.*, **43**, 656–666.
- Polanski, J., Gieleciak, R. and Bak, A. (2002) The comparative molecular surface analysis (COMSA) – a nongrid 3D QSAR method by a coupled neural network and PLS system: predicting pK_a values of benzoic and alkanolic acids. *J. Chem. Inf. Comput. Sci.*, **42**, 184–191.
- Polanski, J. and Gutman, I. (1979) *MATCH Commun. Math. Comput. Chem.*, **5**, 149.
- Polanski, J. and Rouvray, D.E. (1976a) Graph-theoretical treatment of aromatic hydrocarbons. I. The formal graph-theoretical description. *MATCH Commun. Math. Comput. Chem.*, **2**, 63–90.
- Polanski, J. and Rouvray, D.E. (1976b) Graph-theoretical treatment of aromatic hydrocarbons. II. The analysis of all-benzenoid systems. *MATCH Commun. Math. Comput. Chem.*, **2**, 91–109.
- Polanski, J. and Walczak, B. (2000) The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Computers Chem.*, **24**, 615–625.
- Polansky, O.E. (1991) Elements of graph theory for chemists, in *Chemical Graph Theory. Introduction and Fundamentals* (eds D. Bonchev and D.H. Rouvray), Abacus Press/Gordon and Breach Science Publishers, New York, pp. 42–96.
- Polansky, O.E. and Derflinger, G. (1963) Über den Zusammenhang von Bindungslängen und Elektronegativitäten. *Theor. Chim. Acta*, **1**, 308–315.
- Polansky, O.E., Randić, M. and Hosoya, H. (1989) Transfer matrix approach to the Wiener number of catacondensed benzenoids. *MATCH Commun. Math. Comput. Chem.*, **24**, 3–28.
- Politzer, P. (1987) A relationship between the charge capacity and the hardness of neutral atoms and groups. *J. Chim. Phys.*, **86**, 1072–1073.
- Politzer, P., Lane, P., Murray, J.S. and Brinck, T. (1992) Investigation of relationships between solute molecule surface electrostatic potentials and solubilities in supercritical fluids. *J. Phys. Chem.*, **96**, 7938–7943.
- Politzer, P. and Murray, J.S. (1994) *Quantitative Treatments of Solute/Solvent Interactions*, Elsevier, Amsterdam, The Netherlands.
- Politzer, P. and Murray, J.S. (1998) Relationships between lattice energies and surface electrostatic potentials and areas of anions. *J. Phys. Chem. A*, **102**, 1018–1020.
- Politzer, P. and Murray, J.S. (2002) The fundamental nature and role of the electrostatic potential in atoms and molecules. *Theor. Chem. Acc.*, **108**, 134–142.
- Politzer, P., Murray, J.S. and Abu-Awwad, F. (2000) Prediction of solvation free energies from computed properties of solute molecular surfaces. *Int. J. Quant. Chem.*, **76**, 643–647.
- Politzer, P., Murray, J.S. and Flodmark, P. (1996) Relationship between measured diffusion coefficients and calculated molecular surface properties. *J. Phys. Chem.*, **100**, 5538–5540.
- Politzer, P., Murray, J.S., Grice, M.E., DeSalvo, M. and Miller, E. (1997) Calculation of heats of sublimation and solid phase heats of formation. *Mol. Phys.*, **91**, 923–928.
- Politzer, P., Murray, J.S., Lane, P. and Brinck, T. (1993) Relationships between solute molecular properties and solubility in supercritical CO_2 . *J. Phys. Chem.*, **97**, 729–732.

- Polley, M.J., Winkler, D.A. and Burden, F.R. (2004) Broad-based quantitative structure–activity relationship modeling of potency and selectivity of farnesyltransferase inhibitors using a Bayesian regularized neural network. *J. Med. Chem.*, **47**, 6230–6238.
- POLLY, Ver. 2.3, Basak, S.C., Harriss, D.K. and Magnuson, V.R., University of Minnesota, MN.
- Pólya, G. (1936) Algebraische Berechnung der Anzahl der Isomeren einiger organischer Verbindungen. *Z. Kristallogr. (German)*, **93**, 415–443.
- Pólya, G. (1937a) Kombinatorische Anzahlbestimmung für Gruppen, Graphen und chemische Verbindungen. *Acta Math.*, **68**, 145–254.
- Pólya, G. (1937b) Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Math.*, **68**, 145–254.
- Pólya, G. and Read, R.C. (eds) (1987) *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds*, Springer, New York.
- Pompe, M. (2005) Variable connectivity index as a tool for solving the anti-connectivity problem. *Chem. Phys. Lett.*, **404**, 296–299.
- Pompe, M., Davis, J.M. and Samuel, C.D. (2004) Prediction of thermodynamic parameters in gas chromatography from molecular structure: hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **44**, 399–409.
- Pompe, M. and Novič, M. (1999) Prediction of gas-chromatographic retention indices using topological descriptors. *J. Chem. Inf. Comput. Sci.*, **39**, 59–67.
- Pompe, M. and Randić, M. (2006) “Anticonnectivity”: a challenge for structure–property–activity studies. *J. Chem. Inf. Model.*, **46**, 2–8.
- Pompe, M. and Randić, M. (2007) Variable connectivity model for determination of pK_a values for selected organic acids. *Acta Chim. Sloven.*, **54**, 605–610.
- Pompe, M., Razinger, M., Novič, M. and Veber, M. (1997) Modelling of gas chromatographic retention indices using counterpropagation neural networks. *Anal. Chim. Acta*, **348**, 215–221.
- Pompe, M., Veber, M., Randić, M. and Balaban, A.T. (2004) Using variable and fixed topological indices for the prediction of reaction rate constants of volatile unsaturated hydrocarbons with OH radicals. *Molecules*, **9**, 1160–1176.
- Ponce, A.M., Blanco, S.E., Molina, A.S., García-Domenech, R. and Gálvez, J. (2000) Study of the action of flavonoids on xanthine-oxidase by molecular topology. *J. Chem. Inf. Comput. Sci.*, **40**, 1039–1045.
- Ponec, R., Amat, L. and Carbó-Dorca, R. (1999) Molecular basis of quantitative structure–properties relationships (QSPR): a quantum similarity approach. *J. Comput. Aid. Mol. Des.*, **13**, 259–270.
- Ponec, R. and Cooper, D.L. (2005) Anatomy of bond formation. Bond length dependence of the extent of electron sharing in chemical bonds. *J. Mol. Struct. (Theochem)*, **727**, 133–138.
- Ponec, R., Gironés, X. and Carbó-Dorca, R. (2002) Molecular basis of linear free energy relationships. The nature of inductive effect in aliphatic series. *J. Chem. Inf. Comput. Sci.*, **42**, 564–570.
- Ponec, R. and Strand, M. (1990) A novel approach to the characterization of molecular similarity. The 2nd order similarity index. *Collect. Czech. Chem. Comm.*, **55**, 896–902.
- Popelier, P.L.A. (1999) Quantum molecular similarity. 1. BCP space. *J. Phys. Chem. A*, **103**, 2883–2890.
- Popelier, P.L.A., Smith, P.J. and Chaudry, U.A. (2004) Quantitative structure–activity relationships of mutagenic activity from quantum topological descriptors: triazenes and halogenated hydroxyfuranones (mutagen-X) derivatives. *J. Comput. Aid. Mol. Des.*, **18**, 709–718.
- Pople, J.A. (1953) Electron interaction in unsaturated hydrocarbons. *Trans. Faraday Soc.*, **49**, 1375–1385.
- Popoviciu, V., Holban, S., Badilescu, I.I. and Simon, Z. (1978) Steric mapping of estrogenic receptor sites by a minimal steric difference method. *Studia Biophys.*, **69**, 75–76.
- Poroikov, V.V. and Filimonov, D. (2001) Computer-aided prediction of biological activity spectra. Application for finding and optimization of new leads, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 403–407.
- Poroikov, V.V., Filimonov, D.A., Borodina, Yu.V., Lagunin, A.A. and Kos, A. (2000) Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **40**, 1349–1355.
- Poroikov, V.V., Filimonov, D.A., Ihlenfeldt, W.D., Glorizova, T.A., Lagunin, A.A., Borodina, Yu.V., Stepanchikova, A.V. and Nicklaus, M.C. (2003) PASS biological activity spectrum predictions in the enhanced open NCI database browser. *J. Chem. Inf. Comput. Sci.*, **43**, 228–236.
- Poshusta, R. and McHughes, M.C. (1989) Embedding frequencies of trees. *J. Math. Chem.*, **3**, 193–215.
- Poso, A., Tuppurainen, K. and Gynther, J. (1994) Modeling of molecular mutagenicity with comparative molecular field analysis (CoMFA):

- structural and electronic properties of MX compounds related to TA100 mutagenicity. *J. Mol. Struct. (Theochem)*, **304**, 255–260.
- Poso, A., Tuppurainen, K., Ruuskanen, J. and Gynther, J. (1993) Binding of some dioxins and dibenzofurans to the Ah receptor. A QSAR model based on comparative molecular field analysis (CoMFA). *J. Mol. Struct. (Theochem)*, **282**, 259–264.
- Potts, R.O. and Guy, R.H. (1995) A predictive algorithm for skin permeability: the effects of molecular size and hydrogen bond activity. *Pharm. Res.*, **12**, 1628–1663.
- Pozzan, A. (2006) Molecular descriptors and methods for ligand based virtual high throughput screening in drug discovery. *Curr. Pharm. Design*, **12**, 2099–2110.
- Pozzan, A., Feriani, A., Tedesco, G. and Capelli, A.M. (2001) 3D pharmacophoric hashed fingerprints, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science Barcelona (Spain) pp. 224–228.
- Prabhakar, Y.S., Gupta, M.K., Roy, N. and Venkateswarlu, Y. (2005) A high dimensional QSAR study on the aldose reductase inhibitory activity of some flavones: topological descriptors in modeling the activity. *J. Chem. Inf. Model.*, **46**, 86–92.
- Prabhakar, Y.S., Rawal, R.K., Gupta, S., Solomon, V.R. and Katti, S.B. (2005) Topological descriptors in modeling the HIV inhibitory activity of 2-aryl-3-pyridyl-thiazolidin-4-ones. *Comb. Chem. High T. Scr.*, **8**, 431–437.
- Prasanna, S., Manivannan, E. and Chaturvedi, S.C. (2004) QSAR analysis of 2,3-diaryl benzopyrans/pyrans as selective COX-2 inhibitors based on semiempirical AM1 calculations. *QSAR Comb. Sci.*, **23**, 621–628.
- Prathipati, P. and Saxena, A.K. (2006) Evaluation of binary QSAR models derived from LUDI and MOE scoring functions for structure based virtual screening. *J. Chem. Inf. Model.*, **46**, 39–51.
- Pratt, A.D. (1977) A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, **28**, 285–292.
- Primas, H. (1981) *Chemistry, Quantum Mechanics and Reductionism*, Springer-Verlag, Berlin, Germany, p. 452.
- Pritchard, H.O. (1963) Equalization of electronegativity. *J. Am. Chem. Soc.*, **85**, 1876.
- Pritchard, H.O. and Skinner, H.A. (1955) The concept of electronegativity. *Chem. Rev.*, **55**, 745.
- Pungpo, P., Wolschann, P. and Hannongbua, S. (2001) Quantitative structure–activity relationships of HIV-1 reverse transcriptase inhibitors, using hologram QSAR, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science Barcelona (Spain) pp. 206–210.
- Purcell, W.P., Bass, G.E. and Clayton, J.M. (1973) *Strategy of Drug Design. A Molecular Guide to Biological Activity*, John Wiley & Sons, Inc., New York.
- Purdy, R. (1996) A mechanism mediated model for carcinogenicity model content and prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 25 organic chemicals. *Environ. Health Persp.*, **104**, 1085–1094.
- Puri, R.D., Mirgal, S.V., Ramaa, C.S. and Kulkarni, V. M. (1996) Chromatographically derived hydrophobicity parameters in QSAR analysis of diarylsulphone analogs. *Indian J. Chem.*, **35B**, 1271–1274.
- Puri, S., Chickos, J.S. and Welsh, W.J. (2002a) Three-dimensional quantitative structure–property relationship (3D-QSPR) models for prediction of thermodynamic properties of polychlorinated biphenyls (PCBs): enthalpy of sublimation. *J. Chem. Inf. Comput. Sci.*, **42**, 109–116.
- Puri, S., Chickos, J.S. and Welsh, W.J. (2002b) Three-dimensional quantitative structure–property relationship (3D-QSPR) models for prediction of thermodynamic properties of polychlorinated biphenyls (PCBs): enthalpy of vaporization. *J. Chem. Inf. Comput. Sci.*, **42**, 299–304.
- Puri, S., Chickos, J.S. and Welsh, W.J. (2003) Three-dimensional quantitative structure–property relationship (3D-QSPR) models for prediction of thermodynamic properties of polychlorinated biphenyls (PCBs): enthalpies of fusion and their application to estimates of enthalpies of sublimation and aqueous solubilities. *J. Chem. Inf. Comput. Sci.*, **43**, 55–62.
- Pussemier, L., De Borger, R., Cloos, P. and Van Bladel, R. (1989) Relation between the molecular structure and the adsorption of arylcarbamate, phenylurea and anilide pesticides in soil and model organic adsorbents. *Chemosphere*, **18**, 1871–1882.
- Put, R., Perrin, C., Questier, F., Coomans, D., Massart, D.L. and Vander Heyden, Y. (2003) Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure–retention relationship studies. *J. Chromat.*, **988**, 261–276.
- Put, R., Xu, Q.-S., Massart, D.L. and Vander Heyden, Y. (2004) Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies. *J. Chromat.*, **1055**, 11–19.
- Putta, S., Lemmen, C., Beroza, P. and Greene, J. (2002) A novel shape-feature based approach to

- virtual library screening. *J. Chem. Inf. Comput. Sci.*, **42**, 1230–1240.
- Pyka, A. and Bober, K. (2003) Selected structural descriptors and R_M values for calculation and prediction of molar volume of homologous series of saturated fatty acids. *Indian J. Chem.*, **42**, 1360–1367.
- Pyka, A., Kepczynska, E. and Bojarski, J. (2003) Application of selected traditional structural descriptors to QSRR and QSAR analysis of barbiturates. *Indian J. Chem.*, **42**, 1405–1413.
- Qi, Y.-H., Zhang, Q.-Y. and Xu, L. (2002) Correlation analysis of the structures and stability constants of gadolinium(III) complexes. *J. Chem. Inf. Comput. Sci.*, **42**, 1471–1475.
- Qian, L., Hirono, S., Matsushita, Y. and Moriguchi, I. (1992) QSARs based on fuzzy adaptive least squares analysis for the aquatic toxicity of organic chemicals. *Environ. Toxicol. Chem.*, **11**, 953–959.
- QikProp, Schrödinger, LLC, New York.
- QuaSAR, Chemical Computing Group, Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada, <http://www.chemcomp.com/journal/descr.htm>.
- Quayle, O.R. (1953) The parachors of organic compounds. An interpretation and catalogue. *Chem. Rev.*, **53**, 439–589.
- Quigley, J.M. and Nauhton, S.M. (2002) The interrelation of physico-chemical parameters and topological descriptors for a series of β -blocking agents. *J. Chem. Inf. Comput. Sci.*, **42**, 976–982.
- Quiñones, C., Caceres, J., Stud, M. and Martinez, A. (2000) Prediction of drug half-life values of antihistamines based on the CODES/neural network model. *Quant. Struct. -Act. Relat.*, **19**, 448–454.
- Quintas, L.V. and Slater, P.J. (1981) Pairs of non-isomorphic graphs having the same path degree sequence. *MATCH Commun. Math. Comput. Chem.*, **12**, 75–86.
- Rabinowitz, J.R. and Little, S.B. (1991) Prediction of the reactivities of cyclopenta-polynuclear aromatic hydrocarbons by quantum mechanical methods. *Xenobiotica*, **21**, 263–275.
- Rada, J., Araujo, O. and Gutman, I. (2001) Randić index of benzenoid systems and phenylenes. *Croat. Chem. Acta*, **74**, 225–235.
- Radecki, A., Lamparczyk, H. and Kaliszan, R. (1979) A relationship between the retention indices on nematic and isotropic phases and the shape of polycyclic aromatic hydrocarbons. *Chromatographia*, **12**, 595–599.
- Raevsky, O.A. (1987) in *QSAR in Drug Design and Toxicology* (eds D. Hadzi and B. Jerman-Blazic), Elsevier, Amsterdam, The Netherlands, pp. 31–36.
- Raevsky, O.A. (1997a) Hydrogen bond strength estimation by means of the HYBOT program package, in *Computer-Assisted Lead Finding and Optimization* (eds H. van de Waterbeemd, B. Testa and G. Folkers), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 367–378.
- Raevsky, O.A. (1997b) Quantification of non-covalent interactions on the basis of the thermodynamic hydrogen bond parameters. *J. Phys. Org. Chem.*, **10**, 405–413.
- Raevsky, O.A. (1999) Molecular structure descriptors in the computer-aided design of biologically active compounds. *Russ. Chem. Rev.*, **68**, 505–524.
- Raevsky, O.A. (2008) H-bonding parametrization in quantitative structure–activity relationships and drug design, in *Molecular Drug Properties*, Vol. 37 (ed. R. Mannhold), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 127–154.
- Raevsky, O.A., Dolmatova, L., Grigor'ev, V.J., Lisiansky, I. and Bondarev, S. (1995) Molecular recognition descriptors in QSAR, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, pp. 241–245.
- Raevsky, O.A., Fetisov, V.I., Trepalina, E.P., McFarland, J.W. and Schaper, K.-J. (2000) Quantitative estimation of drug absorption in humans for passively transported compounds on the basis of their physico-chemical parameters. *Quant. Struct. -Act. Relat.*, **19**, 366–374.
- Raevsky, O.A., Grigor'ev, V.J., Kireev, D.B. and Zefirov, N.S. (1992a) Complete thermodynamic description of H-bonding in the framework of multiplicative approach. *Quant. Struct. -Act. Relat.*, **11**, 49–63.
- Raevsky, O.A., Grigor'ev, V.J., Kireev, D.B. and Zefirov, N.S. (1992b) Correlation analysis and H bond ability in framework of QSAR. *J. Chim. Phys. Phys-Chim. Biol.*, **89**, 1747–1753.
- Raevsky, O.A., Grigor'ev, V.J. and Mednikova, E. (1993) QSAR H-bonding descriptions, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 116–119.
- Raevsky, O.A., Raevskaja, O.E. and Schaper, K.-J. (2007) Physico-chemical properties/descriptors governing the solubility and partitioning of chemicals in water–solvent–gas systems. Vapor pressure and concentration of chemicals above saturated aqueous solutions. *QSAR Comb. Sci.*, **26**, 1060–1064.
- Raevsky, O.A., Raevskaja, O.E. and Shaper, K.-J. (2004) Analysis of water solubility data on the basis of HYBOT descriptors. Part 3. Solubility of solid

- neutral chemicals and drugs. *QSAR Comb. Sci.*, **23**, 327–343.
- Raevsky, O.A., Sapegin, A. and Zefirov, N. (1994) The QSAR discriminant-regression model. *Quant. Struct. -Act. Relat.*, **13**, 412–418.
- Raevsky, O.A., Schaper, K.-J., Artursson, P. and McFarland, J.W. (2001) A novel approach for prediction of intestinal absorption of drugs in humans based on hydrogen bond descriptors and structural similarity. *Quant. Struct. -Act. Relat.*, **20**, 402–413.
- Raevsky, O.A. and Shaper, K.-J. (2003) Analysis of water solubility data on the basis of HYBOT descriptors. Part 1. Partitioning of volatile chemicals in the water–gas phase system. *QSAR Comb. Sci.*, **22**, 926–942.
- Raevsky, O.A. and Skvortsov, V.S. (2002) 3D hydrogen bond thermodynamics (HYBOT) potentials in molecular modelling. *J. Comput. Aid. Mol. Des.*, **16**, 1–10.
- Raevsky, O.A., Trepalin, S.V., Gerasimenko, V.A. and Raevskaja, O.E. (2002) SLIPPER-2001 – software for predicting molecular properties on the basis of physico-chemical descriptors and structural similarity. *J. Chem. Inf. Comput. Sci.*, **42**, 540–549.
- Raevsky, O.A., Trepalin, S.V. and Razdol'skii, A.N. (2000) New QSAR descriptors calculated from interatomic interaction spectra. *Pharm. Chem. J.*, **34**, 646–649.
- Raichurkar, A.V. and Kulkarni, V.M. (2003) Understanding the antitumor activity of novel hydroxysemicarbazide derivatives as ribonucleotide reductase inhibitors using CoMFA and CoMSIA. *J. Med. Chem.*, **46**, 4419–4427.
- Rajkó, R. and Héberger, K. (2001) Conditional Fisher's exact test as a selection criterion for pair-correlation method. Type I and type II errors. *Chemom. Intell. Lab. Syst.*, **57**, 1–14.
- Rajkó, R., Körtvélyesi, T., Sebök-Nagy, K. and Görgényi, M. (2005) Theoretical characterization of McReynolds' constants. *Anal. Chim. Acta*, **554**, 163–171.
- Ralev, N., Karabunarliev, S., Mekenyan, O., Bonchev, D. and Balaban, A.T. (1985) Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). VIII. General principles for computer implementation. *J. Comput. Chem.*, **6**, 587–591.
- Ramos de Armas, R., González Díaz, H., Molina, R., Pérez González, M. and Uriarte, E. (2004) Stochastic-based descriptors studying peptides biological properties: modeling the bitter tasting threshold of dipeptides. *Bioorg. Med. Chem.*, **12**, 4815–4822.
- Ramos de Armas, R., González Díaz, H., Molina, R. and Uriarte, E. (2005) Stochastic-based descriptors studying biopolymers biological properties: extended MARCH-INSIDE methodology describing antibacterial activity of lactoferricin derivatives. *Biopolymers*, **77**, 247–256.
- Ramsden, C.A. (ed.) (1990) *Quantitative Drug Design*, Vol. 4, Pergamon Press, Oxford, UK, p. 766.
- Ran, Y., Jain, N. and Yalkowsky, S.H. (2001) Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.*, **41**, 1208–1217.
- Ran, Y. and Yalkowsky, S.H. (2001) Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.*, **41**, 354–357.
- Randić, M. (1974) On the recognition of identical graphs representing molecular topology. *J. Chim. Phys.*, **60**, 3920–3928.
- Randić, M. (1975a) Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. *Tetrahedron*, **31**, 1477–1481.
- Randić, M. (1975b) On characterization of molecular branching. *J. Am. Chem. Soc.*, **97**, 6609–6615.
- Randić, M. (1975c) On rearrangement of the connectivity matrix of a graph. *J. Chim. Phys.*, **62**, 309–310.
- Randić, M. (1975d) On unique numbering of atoms and unique codes for molecular graphs. *J. Chem. Inf. Comput. Sci.*, **15**, 105–108.
- Randić, M. (1977a) A graph-theoretical approach to conjugation and resonance energies of hydrocarbons. *Tetrahedron*, **33**, 1905–1920.
- Randić, M. (1977b) Aromaticity and conjugation. *J. Am. Chem. Soc.*, **99**, 444–450.
- Randić, M. (1977c) On canonical numbering of atoms in a molecule and graph isomorphism. *J. Chem. Inf. Comput. Sci.*, **17**, 171–180.
- Randić, M. (1978a) Fragment search in acyclic structures. *J. Chem. Inf. Comput. Sci.*, **18**, 101–107.
- Randić, M. (1978b) The structural origin of chromatographic retention data. *J. Chromat.*, **161**, 1–14.
- Randić, M. (1979) Characterization of atoms, molecules, and classes of molecules based on paths enumeration. *MATCH Commun. Math. Comput. Chem.*, **7**, 5–64.
- Randić, M. (1980a) Chemical shift sums. *J. Magn. Reson.*, **39**, 431–436.
- Randić, M. (1980b) Graphical enumeration of conformations of chains. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **7**, 187–197.
- Randić, M. (1980c) Random walks and their diagnostic value for characterization of atomic environment. *J. Comput. Chem.*, **1**, 386–399.

- Randić, M. (1982) On evaluation of the characteristic polynomial for large molecules. *J. Comput. Chem.*, **3**, 421–435.
- Randić, M. (1983) On alternative form of the characteristic polynomial and the problem of graph recognition. *Theor. Chim. Acta*, **62**, 485–498.
- Randić, M. (1984a) Nonempirical approach to structure–activity studies. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **11**, 137–153.
- Randić, M. (1984b) On molecular identification numbers. *J. Chem. Inf. Comput. Sci.*, **24**, 164–175.
- Randić, M. (1986a) Compact molecular codes. *J. Chem. Inf. Comput. Sci.*, **26**, 136–148.
- Randić, M. (1986b) Molecular ID numbers: by design. *J. Chem. Inf. Comput. Sci.*, **26**, 134–136.
- Randić, M. (1988a) Molecular topographic descriptors. *Stud. Phys. Theor. Chem.*, **54**, 101–108.
- Randić, M. (1988b) On characterization of three-dimensional structures. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **15**, 201–208.
- Randić, M. (1988c) Ring ID numbers. *J. Chem. Inf. Comput. Sci.*, **28**, 142–147.
- Randić, M. (1989) Aromaticity in polycyclic conjugated hydrocarbons dianions. *J. Mol. Struct. (Theochem)*, **185**, 249–274.
- Randić, M. (1990a) Design of molecules with desired properties. A molecular similarity approach to property optimization, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 77–145.
- Randić, M. (1990b) The nature of the chemical structure. *J. Math. Chem.*, **4**, 157–184.
- Randić, M. (1991a) Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (Theochem)*, **233**, 45–59.
- Randić, M. (1991b) Generalized molecular descriptors. *J. Math. Chem.*, **7**, 155–168.
- Randić, M. (1991c) Novel graph theoretical approach to heteroatoms in quantitative structure–activity relationships. *Chemom. Intell. Lab. Syst.*, **10**, 213–227.
- Randić, M. (1991d) On computation of optimal parameters for multivariate analysis of structure–property relationship. *J. Comput. Chem.*, **12**, 970–980.
- Randić, M. (1991e) Orthogonal molecular descriptors. *New J. Chem.*, **15**, 517–525.
- Randić, M. (1991f) Resolution of ambiguities in structure–property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.*, **31**, 311–320.
- Randić, M. (1991g) Search for optimal molecular descriptors. *Croat. Chem. Acta*, **64**, 43–54.
- Randić, M. (1992a) Chemical structure. What is “she”? *J. Chem. Educ.*, **69**, 713–718.
- Randić, M. (1992b) In search of structural invariants. *J. Math. Chem.*, **9**, 97–146.
- Randić, M. (1992c) Representation of molecular graphs by basic graphs. *J. Chem. Inf. Comput. Sci.*, **32**, 57–69.
- Randić, M. (1992d) Similarity based on extended basis descriptors. *J. Chem. Inf. Comput. Sci.*, **32**, 686–692.
- Randić, M. (1993a) Comparative regression analysis. Regressions based on a single descriptor. *Croat. Chem. Acta*, **66**, 289–312.
- Randić, M. (1993b) Fitting of nonlinear regressions by orthogonalized power series. *J. Comput. Chem.*, **14**, 363–370.
- Randić, M. (1993c) Novel molecular descriptor for structure–property studies. *Chem. Phys. Lett.*, **211**, 478–483.
- Randić, M. (1994a) Curve-fitting paradox. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **21**, 215–225.
- Randić, M. (1994b) Hosoya matrix – a source of new molecular descriptors. *Croat. Chem. Acta*, **67**, 415–429.
- Randić, M. (1995a) Molecular profiles. Novel geometry-dependent molecular descriptors. *New J. Chem.*, **19**, 781–791.
- Randić, M. (1995b) Molecular shape profiles. *J. Chem. Inf. Comput. Sci.*, **35**, 373–382.
- Randić, M. (1995c) Restricted random walks on graphs. *Theor. Chim. Acta*, **92**, 97–106.
- Randić, M. (1996a) Molecular bonding profiles. *J. Math. Chem.*, **19**, 375–392.
- Randić, M. (1996b) Orthosimilarity. *J. Chem. Inf. Comput. Sci.*, **36**, 1092–1097.
- Randić, M. (1996c) Quantitative structure–property relationship – boiling points of planar benzenoids. *New J. Chem.*, **20**, 1001–1009.
- Randić, M. (1997a) Linear combinations of path numbers as molecular descriptors. *New J. Chem.*, **21**, 945–951.
- Randić, M. (1997b) On characterization of chemical structure. *J. Chem. Inf. Comput. Sci.*, **37**, 672–687.
- Randić, M. (1997c) On characterization of cyclic structures. *J. Chem. Inf. Comput. Sci.*, **37**, 1063–1071.
- Randić, M. (1997d) On molecular branching. *Acta Chim. Sloven.*, **44**, 57–77.
- Randić, M. (1997e) Resonance in catacondensed benzenoid hydrocarbons. *Int. J. Quant. Chem.*, **63**, 585–600.
- Randić, M. (1998a) On characterization of molecular attributes. *Acta Chim. Sloven.*, **45**, 239–252.

- Randić, M. (1998b) On structural ordering and branching of acyclic saturated hydrocarbons. *J. Math. Chem.*, **24**, 345–358.
- Randić, M. (1998c) Topological indices, in *Encyclopedia of Computational Chemistry* (ed. P.R. von Schleyer), John Wiley & Sons, Ltd, London, UK, pp. 3018–3032.
- Randić, M. (2000a) Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.*, **40**, 50–56.
- Randić, M. (2000b) On characterization of DNA primary sequences by a condensed matrix. *Chem. Phys. Lett.*, **317**, 29–34.
- Randić, M. (2000c) On characterization of pharmacophore. *Acta Chim. Sloven.*, **47**, 143–151.
- Randić, M. (2001a) Graph theoretical descriptors of two-dimensional chirality with possible extension to three-dimensional chirality. *J. Chem. Inf. Comput. Sci.*, **41**, 639–649.
- Randić, M. (2001b) Graph valence shells as molecular descriptors. *J. Chem. Inf. Comput. Sci.*, **41**, 627–630.
- Randić, M. (2001c) Novel shape descriptors for molecular graphs. *J. Chem. Inf. Comput. Sci.*, **41**, 607–613.
- Randić, M. (2001d) On complexity of transitive graphs representing degenerate rearrangements. *Croat. Chem. Acta*, **74**, 683–705.
- Randić, M. (2001e) On graphical and numerical characterization of proteomics maps. *J. Chem. Inf. Comput. Sci.*, **41**, 1330–1338.
- Randić, M. (2001f) Retro-regressions – another important multivariate regression improvement. *J. Chem. Inf. Comput. Sci.*, **41**, 602–606.
- Randić, M. (2001g) The connectivity index 25 years after. *J. Mol. Graph. Model.*, **20**, 19–35.
- Randić, M. (2002a) A graph-theoretical characterization of proteomics maps. *Int. J. Quant. Chem.*, **90**, 848–858.
- Randić, M. (2002b) On generalization of Wiener index for cyclic structures. *Acta Chim. Sloven.*, **49**, 483–496.
- Randić, M. (2003a) Aromaticity of polycyclic conjugated hydrocarbons. *Chem. Rev.*, **103**, 3449–3605.
- Randić, M. (2003b) Chemical graph theory – facts and fiction. *Indian J. Chem.*, **42**, 1207–1218.
- Randić, M. (2004a) Algebraic Kekulé formulas for benzenoid hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **44**, 365–372.
- Randić, M. (2004b) Wiener–Hosoya index – a novel graph theoretical molecular descriptor. *J. Chem. Inf. Comput. Sci.*, **44**, 373–377.
- Randić, M. (2007) Conjugated circuits and resonance energies of benzenoid hydrocarbons. *Chem. Phys. Lett.*, **38**, 68–70.
- Randić, M. (2008) On history of the Randić index and emerging hostility toward chemical graph theory. *MATCH Commun. Math. Comput. Chem.*, **59**, 5–124.
- Randić, M. and Balaban, A.T. (2003) On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.*, **43**, 532–539.
- Randić, M., Balaban, A.T. and Basak, S.C. (2001) On structural interpretation of several distance related topological indices. *J. Chem. Inf. Comput. Sci.*, **41**, 593–601.
- Randić, M., Basak, N. and Plavšić, D. (2004) Novel graphical matrix and distance-based molecular descriptors. *Croat. Chem. Acta*, **77**, 251–257.
- Randić, M. and Basak, S.C. (1999) Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.*, **39**, 261–266.
- Randić, M. and Basak, S.C. (2000a) Construction of high-quality structure–property–activity regressions: the boiling points of sulfides. *J. Chem. Inf. Comput. Sci.*, **40**, 899–905.
- Randić, M. and Basak, S.C. (2000b) Multiple regression analysis with optimal molecular descriptors. *SAR & QSAR Environ. Res.*, **11**, 1–23.
- Randić, M. and Basak, S.C. (2001a) A new descriptor for structure–property and structure–activity correlations. *J. Chem. Inf. Comput. Sci.*, **41**, 650–656.
- Randić, M. and Basak, S.C. (2001b) Characterization of DNA primary sequences based on the average distances between bases. *J. Chem. Inf. Comput. Sci.*, **41**, 561–568.
- Randić, M. and Basak, S.C. (2001c) On use of the variable connectivity index ${}^1\chi^f$ in QSAR: toxicity of aliphatic ethers. *J. Chem. Inf. Comput. Sci.*, **41**, 614–618.
- Randić, M. and Basak, S.C. (2002) A comparative study of proteomics maps using graph theoretical biodescriptors. *J. Chem. Inf. Comput. Sci.*, **42**, 983–992.
- Randić, M. and Basak, S.C. (2004) On similarity of proteome maps. *Med. Chem. Res.*, **13**, 800–811.
- Randić, M., Basak, S.C., Pompe, M. and Novič, M. (2001) Prediction of gas chromatographic retention indices using variable connectivity index. *Acta Chim. Sloven.*, **48**, 169–180.
- Randić, M., Brisse, G.M., Spencer, R.B. and Wilkins, C.L. (1979) Search for all self-avoiding paths for molecular graphs. *Computers Chem.*, **3**, 5–13.
- Randić, M., Brisse, G.M., Spencer, R.B. and Wilkins, C.L. (1980) Use of self-avoiding paths for characterization of molecular graphs with multiple bonds. *Computers Chem.*, **4**, 27–43.
- Randić, M., Brisse, G.M. and Wilkins, C.L. (1981) Computer perception of topological symmetry via

- canonical numbering of atoms. *J. Chem. Inf. Comput. Sci.*, **21**, 52–59.
- Randić, M. and DeAlba, L.M. (1997) Dense graphs and sparse matrices. *J. Chem. Inf. Comput. Sci.*, **37**, 1078–1081.
- Randić, M., DeAlba, L.M. and Harris, F.E. (1998) Graphs with the same Detour matrix. *Croat. Chem. Acta*, **71**, 53–68.
- Randić, M. and Dobrowolski, J.Cz. (1998) Optimal molecular connectivity descriptors for nitrogen-containing molecules. *Int. J. Quant. Chem.*, **70**, 1209–1215.
- Randić, M., El-Basil, S., Nikolić, S. and Trinajstić, N. (1998) Clar polynomials of large benzenoid systems. *J. Chem. Inf. Comput. Sci.*, **38**, 563–574.
- Randić, M. and Guo, X. (1999) Giant benzenoid hydrocarbons. Superphenalene resonance energy. *New J. Chem.*, **23**, 251–260.
- Randić, M., Guo, X. and Basak, S.C. (2001) On the characterization of DNA primary sequences by triplet of nucleic acid bases. *J. Chem. Inf. Comput. Sci.*, **41**, 619–626.
- Randić, M., Guo, X. and Calkins, P. (2000) Graph dissection revisited. Application to smaller alkanes. *Acta Chim. Sloven.*, **47**, 489–506.
- Randić, M., Guo, X., Oxley, T., and Krishnapriyan, H. (1993) Wiener matrix: source of novel graph invariants. *J. Chem. Inf. Comput. Sci.*, **33**, 709–716.
- Randić, M., Guo, X., Oxley, T., Krishnapriyan, H., and Naylor, L. (1994) Wiener matrix invariants. *J. Chem. Inf. Comput. Sci.*, **34**, 361–367.
- Randić, M., Hansen, P.J. and Jurs, P.C. (1988) Search for useful graph theoretical invariants of molecular structure. *J. Chem. Inf. Comput. Sci.*, **28**, 60–68.
- Randić, M., Jericevic, Z., Sabljic, A. and Trinajstić, N. (1988) On the molecular connectivity and π -electronic energy in polycyclic hydrocarbons. *Acta Phys. Pol.*, **74**, 317–330.
- Randić, M., Jerman-Blazic, B., Rouvray, D.H., Seybold, P.G. and Grossman, S.C. (1987) The search for active substructures in structure–activity studies. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **14**, 245–260.
- Randić, M., Jerman-Blazic, B. and Trinajstić, N. (1990) Development of 3-dimensional molecular descriptors. *Computers Chem.*, **14**, 237–246.
- Randić, M. and Jurs, P.C. (1989) On a fragment approach to structure–activity correlations. *Quant. Struct. -Act. Relat.*, **8**, 39–48.
- Randić, M., Klein, D.J., El-Basil, S. and Calkins, P. (1996) Resonance in large benzenoid hydrocarbons. *Croat. Chem. Acta*, **69**, 1639–1660.
- Randić, M., Kleiner, A.F. and DeAlba, L.M. (1994) Distance/distance matrices. *J. Chem. Inf. Comput. Sci.*, **34**, 277–286.
- Randić, M., Kraus, G.A. and Jerman-Blazic Dzonova, B. (1983) Ordering of graphs as an approach to structure–activity studies, in *Chemical Applications of Topology and Graph Theory* (ed. R.B. King), Elsevier, Amsterdam, The Netherlands, pp. 18–22.
- Randić, M. and Krilov, G. (1996) Bond profiles for cuboctahedron and twist cuboctahedron. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **23**, 127–139.
- Randić, M. and Krilov, G. (1997a) Characterization of 3-D sequences of proteins. *Chem. Phys. Lett.*, **272**, 115–119.
- Randić, M. and Krilov, G. (1997b) On characterization of molecular surfaces. *Int. J. Quant. Chem.*, **65**, 1065–1076.
- Randić, M. and Krilov, G. (1999) On a characterization of the folding of proteins. *Int. J. Quant. Chem.*, **75**, 1017–1026.
- Randić, M., Lerš, N., Plavšić, D. and Basak, S.C. (2004a) Characterization of 2-D proteome maps based on the nearest neighborhoods of spots. *Croat. Chem. Acta*, **77**, 345–351.
- Randić, M., Lerš, N., Plavšić, D. and Basak, S.C. (2004b) On invariants of a 2-D proteome map derived from neighborhood graphs. *J. Proteome Res.*, **3**, 778–785.
- Randić, M., Lerš, N., Plavšić, D., Basak, S.C. and Balaban, A.T. (2005a) Four-color map representation of DNA or RNA and their numerical characterization. *Chem. Phys. Lett.*, **407**, 205–208.
- Randić, M., Lerš, N., Vukicević, D., Plavšić, D., Gute, B.D. and Basak, S.C. (2005b) Canonical labeling of proteome maps. *J. Proteome Res.*, **4**, 1347–1352.
- Randić, M. and Mezey, P.G. (1996) Palindromic perimeter codes and chirality properties of polyhexes. *J. Chem. Inf. Comput. Sci.*, **36**, 1183–1186.
- Randić, M., Mihalić, Z., Nikolić, S. and Trinajstić, N. (1993) Graph-theoretical correlations – artifacts or facts? *Croat. Chem. Acta*, **66**, 411–434.
- Randić, M., Mihalić, Z., Nikolić, S. and Trinajstić, N. (1994) Graphical bond orders: novel structural descriptors. *J. Chem. Inf. Comput. Sci.*, **34**, 403–409.
- Randić, M., Mills, D. and Basak, S.C. (2000) On characterization of physical properties of amino acids. *Int. J. Quant. Chem.*, **80**, 1199–1209.
- Randić, M., Morales, D.A. and Araujo, O. (1996) Higher-order Fibonacci numbers. *J. Math. Chem.*, **20**, 79–94.
- Randić, M., Müller, W.R., von Knop, J. and Trinajstić, N. (1997) The characteristic polynomial as a structure discriminator. *J. Chem. Inf. Comput. Sci.*, **37**, 1072–1077.

- Randić, M., Novič, M., Vikić-Topić, D. and Plavšić, D. (2006) Novel numerical and graphical representation of DNA sequences and proteins. *SAR & QSAR Environ. Res.*, **17**, 583–595.
- Randić, M., Novič, M. and Vračko, M. (2002) On characterization of dose variations of 2-D proteomics maps by matrix invariants. *J. Proteome Res.*, **1**, 217–226.
- Randić, M., Novič, M. and Vračko, M. (2005) Novel characterization of proteomics maps by sequential neighborhoods of protein spots. *J. Chem. Inf. Model.*, **45**, 1205–1213.
- Randić, M., Oakland, D.O. and Klein, D.J. (1986) Symmetry properties of chemical graphs. IX. The valence tautomerism in the P_7^{3-} skeleton. *J. Comput. Chem.*, **7**, 35–54.
- Randić, M. and Plavšić, D. (2002) On the concept of molecular complexity. *Croat. Chem. Acta*, **75**, 107–116.
- Randić, M. and Plavšić, D. (2003) Characterization of molecular complexity. *Int. J. Quant. Chem.*, **91**, 20–31.
- Randić, M., Plavšić, D. and Lerš, N. (2001) Variable connectivity index for cycle-containing structures. *J. Chem. Inf. Comput. Sci.*, **41**, 657–662.
- Randić, M., Plavšić, D. and Razinger, M. (1997) Double invariants. *MATCH Commun. Math. Comput. Chem.*, **35**, 243–259.
- Randić, M. and Pompe, M. (1999) On characterization of the CC double bond in alkenes. *SAR & QSAR Environ. Res.*, **10**, 451–471.
- Randić, M. and Pompe, M. (2001a) The variable connectivity index $^1\chi^f$ versus the traditional molecular descriptors: a comparative study of $^1\chi^f$ against descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.*, **41**, 631–638.
- Randić, M. and Pompe, M. (2001b) The variable molecular descriptors based on distance related matrices. *J. Chem. Inf. Comput. Sci.*, **41**, 575–581.
- Randić, M., Pompe, M., Mills, D. and Basak, S.C. (2004) Variable connectivity index as a tool for modeling structure–property relationships. *Molecules*, **9**, 1177–1193.
- Randić, M. and Razinger, M. (1995a) Molecular topographic indices. *J. Chem. Inf. Comput. Sci.*, **35**, 140–147.
- Randić, M. and Razinger, M. (1995b) On characterization of molecular shapes. *J. Chem. Inf. Comput. Sci.*, **35**, 594–606.
- Randić, M. and Razinger, M. (1996) Molecular shapes and chirality. *J. Chem. Inf. Comput. Sci.*, **36**, 429–441.
- Randić, M. and Razinger, M. (1997) On characterization of 3D molecular structure, in *From Chemical Topology to Three-Dimensional Geometry* (ed. A.T. Balaban), Plenum Press, New York, pp. 159–236.
- Randić, M., Sabljic, A., Nikolić, S. and Trinajstić, N. (1988) A rational selection of graph-theoretical indices in the QSAR. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **15**, 267–285.
- Randić, M. and Seybold, P.G. (1993) Molecular shape as a critical factor in structure–property–activity studies. *SAR & QSAR Environ. Res.*, **1**, 77–85.
- Randić, M. and Trinajstić, N. (1988) Composition as a method for data reduction: application to carbon-13 NMR chemical shifts. *Theor. Chim. Acta*, **73**, 233–246.
- Randić, M. and Trinajstić, N. (1993a) In search for graph invariants of chemical interest. *J. Mol. Struct.*, **300**, 551–571.
- Randić, M. and Trinajstić, N. (1993b) Viewpoint 4 – Comparative structure–property studies: the connectivity basis. *J. Mol. Struct. (Theochem)*, **284**, 209–221.
- Randić, M. and Trinajstić, N. (1994) Isomeric variations in alkanes: boiling points of nonanes. *New J. Chem.*, **18**, 179–189.
- Randić, M., Trinajstić, N. and Živković, T. (1976) Molecular graphs having identical spectra. *J. Chem. Soc. Faraday Trans II*, **72**, 244–256.
- Randić, M. and Vračko, M. (2000) On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.*, **40**, 599–606.
- Randić, M., Vračko, M., Lerš, N. and Plavšić, D. (2003) Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.*, **368**, 1–6.
- Randić, M., Vračko, M., Nandy, A. and Basak, S.C. (2000) On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.*, **40**, 1235–1244.
- Randić, M., Vračko, M. and Novič, M. (2001) Eigenvalues as molecular descriptors, in *QSPR/QSAR Studies by Molecular Descriptors* (ed. M.V. Diudea), Nova Science Publishers, Huntington, NY, pp. 147–211.
- Randić, M. and Wilkins, C.L. (1979a) Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. *J. Phys. Chem.*, **83**, 1525–1540.
- Randić, M. and Wilkins, C.L. (1979b) Graph theoretical study of structural similarity in benzomorphans. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **6**, 55–71.
- Randić, M. and Wilkins, C.L. (1979c) On a graph theoretical basis for ordering of structures. *Chem. Phys. Lett.*, **63**, 332–336.

- Randić, M. and Wilkins, C.L. (1980) A procedure for characterization of the rings of molecule. *J. Chem. Inf. Comput. Sci.*, **20**, 36–46.
- Randić, M., Witzmann, F., Vračko, M. and Basak, S.C. (2001) On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: application to peroxisome proliferations. *Med. Chem. Res.*, **10**, 456–479.
- Randić, M., Witzmann, F.A., Kodali, V. and Basak, S. C. (2006) On the dependence of a characterization of proteomics maps on the number of protein spots considered. *J. Chem. Inf. Model.*, **46**, 116–122.
- Randić, M., Woodworth, W.L. and Graovac, A. (1983) Unusual random walks. *Int. J. Quant. Chem.*, **24**, 435–452.
- Randić, M. and Zupan, J. (2001) On interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci.*, **41**, 550–560.
- Randić, M., Zupan, J. and Balaban, A.T. (2004) Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.*, **397**, 247–252.
- Randić, M., Zupan, J. and Novič, M. (2001) On 3-D graphical representation of proteomics maps and their numerical characterization. *J. Chem. Inf. Comput. Sci.*, **41**, 1339–1344.
- Randić, M., Zupan, J., Novič, M., Gute, B.D. and Basak, S.C. (2002) Novel matrix invariants for characterization of changes of proteomics maps. *SAR & QSAR Environ. Res.*, **13**, 689–703.
- Rao, K.R. and Lakshminarayanan, S. (2007) Partial correlation based variable selection approach for multivariate data classification methods. *Chemom. Intell. Lab. Syst.*, **86**, 68–81.
- Raos, N. (2002) Suitability of the topological index $W^{1/3}$ for estimation of the stability constants of coordination compounds. *Croat. Chem. Acta*, **75**, 117–120.
- Raos, N. (2003) Mean molecular radius and the Wiener number: a quest for meaning. *Croat. Chem. Acta*, **76**, 81–85.
- Rarey, M. and Dixon, J.S. (1998) Feature trees: a new molecular similarity measure based on tree matching. *J. Comput. Aid. Mol. Des.*, **12**, 471–490.
- Rarey, M. and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aid. Mol. Des.*, **15**, 497–520.
- Rashevsky, N. (1955) Life, information theory and topology. *Bull. Math. Biophys.*, **17**, 229–235.
- Rashevsky, N. (1960) Life, information theory, probability, and physics. *Bull. Math. Biophys.*, **22**, 351–364.
- Rastelli, G., Costantino, L. and Albasini, A. (1995) Theoretical and experimental study of flavones as inhibitors of xanthine oxidase. *Eur. J. Med. Chem.*, **30**, 141–146.
- Rasulev, B.F., Abdullaev, N.D., Syrov, V.N. and Leszczynski, J. (2005) A quantitative structure–activity relationship (QSAR) study of the antioxidant activity of flavonoids. *QSAR Comb. Sci.*, **24**, 1056–1065.
- Ravanel, P., Taillander, G., Tissut, M. and Benoit-Guyod, J.L. (1985) Effects of chlorophenols on isolated plant mitochondria activities: a QSAR study. *Ecotox. Environ. Safety*, **9**, 300–320.
- Ravi, M., Hopfinger, A.J., Hormann, R.E. and Dinan, L. (2001) 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA modeling. *J. Chem. Inf. Comput. Sci.*, **41**, 1587–1604.
- Rawlings, J.O. (1988) *Applied Regression Analysis*, Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Ray, A., Raychaudhury, C. and Nandy, A. (1998) Novel techniques of graphical representation and analysis of DNA sequences – A review. *J. Biosciences*, **23**, 55–71.
- Ray, S.K., Basak, S.C., Raychaudhury, C., Roy, A.B. and Ghosh, J.J. (1981) Quantitative structure–activity relationship studies of bioactive molecules using structural information indices. *Indian J. Chem.*, **20**, 894–897.
- Ray, S.K., Basak, S.C., Raychaudhury, C., Roy, A.B. and Ghosh, J.J. (1982) A quantitative structure–activity relationship study of *N*-alkylmorphetobemidones and triazinones using structural information content. *Arzneim. Forsch. (German)*, **32**, 322–325.
- Ray, S.K., Basak, S.C., Raychaudhury, C., Roy, A.B. and Ghosh, J.J. (1983) The utility of information content, structural information content. Hydrophobicity and van der Waals volume in the design of barbiturates and tumor inhibitory triazines. *Arzneim. Forsch. (German)*, **33**, 352–356.
- Ray, S.K., Gupta, D.K., Basak, S.C., Raychaudhury, C., Roy, A.B. and Ghosh, J.J. (1985) Weighted information indices & anxiolytic drug design. *Indian J. Chem.*, **24**, 1149–1153.
- Raychaudhury, C., Banerjee, A., Bag, P. and Roy, S. (1999) Topological shape and size of peptides: identification of potential allele specific helper T cell antigenic sites. *J. Chem. Inf. Comput. Sci.*, **39**, 248–254.
- Raychaudhury, C. and Klopman, G. (1990) New vertex indexes and their applications in evaluating antileukemic activity of 9-anilinoacridines and the activity of 2',3'-dideoxy-nucleosides against HIV. *Bull. Soc. Chim. Belg.*, **99**, 255–264.
- Raychaudhury, C. and Nandy, A. (1999) Indexing scheme and similarity measures for

- macromolecular sequences. *J. Chem. Inf. Comput. Sci.*, **39**, 243–247.
- Raychaudhury, C., Ray, S.K., Ghosh, J.J., Roy, A.B. and Basak, S.C. (1984) Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.*, **5**, 581–588.
- Raymond, J.W., Blankley, C.J. and Willett, P. (2003) Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *J. Mol. Graph. Model.*, **21**, 421–433.
- Raymond, J.W. and Willett, P. (2002) Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J. Comput. Aid. Mol. Des.*, **16**, 59–71.
- Raymond, J.W. and Willett, P. (2003) Similarity searching in databases of flexible 3D structures using smoothed bounded distance matrices. *J. Chem. Inf. Comput. Sci.*, **43**, 908–916.
- Rayne, S. and Ikonomou, M.G. (2003) Predicting gas chromatographic retention times for the 209 polybrominated diphenyl ether congeners. *J. Chromat.*, **1016**, 235–248.
- Razdol'skii, A.N., Trepalin, S.V. and Raevsky, O.A. (2000) QSAR modeling based on interatomic interaction spectra. *Pharm. Chem. J.*, **34**, 654–657.
- Razinger, M. (1982) Extended connectivity in chemical graphs. *Theor. Chim. Acta*, **61**, 581–586.
- Razinger, M. (1986) Discrimination and ordering of chemical structures by the number of walks. *Theor. Chim. Acta*, **70**, 365–378.
- Razinger, M., Chrétien, J.R. and Dubois, J.-E. (1985) Structural selectivity of topological indexes in alkane series. *J. Chem. Inf. Comput. Sci.*, **25**, 23–27.
- Read, R.C. and Corneil, D.G. (1977) The graph isomorphism disease. *J. Serb. Chem. Soc.*, **1**, 339–363.
- Recanatini, M., Cavalli, A., Belluti, F., Piazzi, L., Rampa, A., Bisi, A., Gobbi, S., Valenti, P., Andrisano, V., Bartolini, M. and Cavrini, V. (2000) SAR of 9-amino-1,2,3,4-tetrahydroacridine-based acetylcholinesterase inhibitors: synthesis, enzyme inhibitory activity, QSAR, and structure-based CoMFA of tacrine analogues. *J. Med. Chem.*, **43**, 2007–2018.
- Reddy, K.N., Dayan, F.E. and Duke, S.O. (1998) QSAR analysis of protoporphyrinogen oxidase inhibitors, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 197–233.
- Reddy, K.N. and Locke, M.A. (1994a) Prediction of soil sorption of herbicides using semi-empirical molecular properties. *Weed Sci.*, **42**, 453–461.
- Reddy, K.N. and Locke, M.A. (1994b) Relationships between molecular properties and log *P* and soil sorption (K_{oc}) of substituted phenylureas: QSAR models. *Chemosphere*, **28**, 1929–1941.
- Reddy, K.N. and Locke, M.A. (1996) Molecular properties as descriptors of octanol/water partition coefficients of herbicides. *Water, Air and Soil Pollution*, **86**, 389–405.
- Reed, A.E., Curtiss, L.A. and Weinhold, F. (1988) Intermolecular interaction from a natural bond orbital. Donor–acceptor viewpoint. *Chem. Rev.*, **88**, 899–926.
- Reed, A.E., Weinstock, R.B. and Weinhold, F. (1985) Natural population analysis. *J. Chim. Phys.*, **83**, 735–746.
- Reed, J.L. (1997) Electronegativity: chemical hardness II. *J. Phys. Chem. A*, **101**, 7401–7407.
- Reichardt, C. (1965) Empirical parameters of the polarity of solvents. *Angew. Chem. Int. Ed. Engl.*, **4**, 29–39.
- Reichardt, C. (1990) *Solvents and Solvent Effects in Organic Chemistry*, VCH Publishers, New York.
- Reichardt, C. and Dimroth, K. (1968) Solvents and empirical parameters for characterization of their polarity. *Fortschr. Chem. Forsch.*, **11**, 1–73.
- Reid, R.C., Prausnitz, J.M. and Poling, B.E. (1988) *The Properties of Gases and Liquids*, McGraw-Hill, New York.
- Reijmers, T.H., Wehrens, R. and Buydens, L. (2001) The influence of different structure representations on the clustering of an RNA nucleotides data set. *J. Chem. Inf. Comput. Sci.*, **41**, 1388–1394.
- Reinhard, M. and Drefahl, A. (1999) *Handbook for Estimating Physico-chemical Properties of Organic Compounds*, John Wiley & Sons, Inc., New York, p. 228.
- Rekker, R.F. (1977a) *The Hydrophobic Fragment Constant*, Elsevier, Amsterdam, The Netherlands.
- Rekker, R.F. (1977b) *The Hydrophobic Fragmental Constant. Its Derivation and Applications. A Means of Characterizing Membrane Systems*, Elsevier, Amsterdam, The Netherlands, p. 390.
- Rekker, R.F. (1992) The history of drug research from Overton to Hansch. *Quant. Struct. -Act. Relat.*, **11**, 195–199.
- Rekker, R.F. and De Kort, H.M. (1979) The hydrophobic fragmental constant: an extension to a 1000 data point set. *Eur. J. Med. Chem.*, **14**, 479–488.
- Rekker, R.F. and de Vries, G. (1993) A basic confrontation of Rekker's revised Σf -system with HPLC retention data obtained on a mixed series of aliphatic and aromatic hydrocarbons, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 132–136.

- Rekker, R.F. and Mannhold, R. (1992) *Calculation of Drug Lipophilicity. The Hydrophobic Fragmental Constant Approach*, VCH Publishers, Weinheim, Germany.
- Rekker, R.F., Mannhold, R., Bijloo, G.J., de Vries, G. and Dross, K. (1998) The lipophilic behaviour of organic compounds. 2. The development of an aliphatic hydrocarbon/water fragmental system via interconnection with octanol/water partitioning data. *Quant. Struct. -Act. Relat.*, **17**, 537–548.
- Rekker, R.F., ter Laak, A.M. and Mannhold, R. (1993) On the reliability of calculated log *P*-values: Rekker, Hansch/Leo and Suzuki approach. *Quant. Struct. -Act. Relat.*, **12**, 152–157.
- Ren, B. (1999) A new topological index for QSPR of alkanes. *J. Chem. Inf. Comput. Sci.*, **39**, 139–143.
- Ren, B. (2002a) Application of novel atom-type AI topological indices to QSPR studies of alkanes. *Computers Chem.*, **26**, 357–369.
- Ren, B. (2002b) Novel atom-type AI indices for QSPR studies of alcohols. *Computers Chem.*, **26**, 223–235.
- Ren, B. (2002c) Novel atomic-level-based AI topological descriptors: application to QSPR/QSAR modeling. *J. Chem. Inf. Comput. Sci.*, **42**, 858–868.
- Ren, B. (2003a) Atom-type-based AI topological descriptors for quantitative structure–retention index correlations of aldehydes and ketones. *Chemom. Intell. Lab. Syst.*, **66**, 29–39.
- Ren, B. (2003b) Atom-type-based AI topological descriptors: application in structure–boiling point correlations of oxo organic compounds. *J. Chem. Inf. Comput. Sci.*, **43**, 1121–1131.
- Ren, B. (2003c) Atomic-level-based AI topological descriptors for structure–property correlations. *J. Chem. Inf. Comput. Sci.*, **43**, 161–169.
- Ren, B. (2003d) New atom-type-based AI topological indices: application to QSPR studies of aldehydes and ketones. *J. Comput. Aid. Mol. Des.*, **17**, 607–620.
- Ren, S. (2002d) Use of molecular descriptors in separating phenols by three mechanisms of toxic action. *Quant. Struct. -Act. Relat.*, **21**, 486–492.
- Ren, S. (2003e) Ecotoxicity prediction using mechanism- and non-mechanism-based QSARs: a preliminary study. *Chemosphere*, **53**, 1053–1065.
- Ren, S. (2003f) Two-step multivariate classification of the mechanisms of toxic action of phenols. *QSAR Comb. Sci.*, **22**, 596–603.
- Ren, Y., Chen, G., Hu, Z., Chen, X. and Yan, B. (2008) Applying novel three-dimensional holographic vector of molecular interaction field to QSAR studies of artemisinin derivatives. *QSAR Comb. Sci.*, **27**, 196–207.
- Renner, S., Ludwig, V., Boden, O., Scheffer, U., Göbel, M. and Schneider, G. (2005) New inhibitors of the Tat–TAR RNA interaction found with a “fuzzy” pharmacophore model. *ChemBioChem*, **6**, 1119–1125.
- Renner, S., Noeske, T., Parsons, C.G., Schneider, P., Weil, T. and Schneider, G. (2005) New allosteric modulators of metabotropic glutamate receptor 5 (mGluR5) found by ligand-based virtual screening. *ChemBioChem*, **6**, 620–625.
- Renner, S. and Schneider, G. (2006) Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem*, **1**, 181–185.
- Restrepo, G., Mesa, H. and Villaveces, J.L. (2006) On the topological sense of chemical sets. *J. Math. Chem.*, **39**, 363–376.
- Restrepo, G. and Villaveces, J.L. (2005) From trees (dendrograms and consensus trees) to topology. *Croat. Chem. Acta*, **78**, 275–281.
- Retzeka, E., Voutsas, E., Magoulas, K. and Tassios, D. (2002) Prediction of physical properties of hydrocarbons, petroleum, and coal liquid fractions. *Ind. Eng. Chem. Res.*, **41**, 1695–1702.
- Reynolds, C.A., Burt, C. and Richards, W.G. (1992) A linear molecular similarity index. *Quant. Struct. -Act. Relat.*, **11**, 34–35.
- Reynolds, C.A., Essex, J.W. and Richards, W.G. (1992) Atomic charges for variable molecular conformations. *J. Am. Chem. Soc.*, **114**, 9075–9079.
- Reynolds, C.H. (1995) Estimating lipophilicity using the GB/SA continuum solvation model: a direct method for computing partition coefficients. *J. Chem. Inf. Comput. Sci.*, **35**, 738–742.
- Reynolds, C.H., Druker, R. and Pfahler, L.B. (1998) Lead discovery using stochastic cluster analysis (SCA): a new method for clustering structurally similar compounds. *J. Chem. Inf. Comput. Sci.*, **38**, 305–312.
- Reynolds, W.F. (1983) Polar substituent effects. *Prog. Phys. Org. Chem.*, **14**, 165–203.
- Reynolds, W.F. and Topsom, R.D. (1984) Field and resonance substituent constants for aromatic derivatives: limitations of Swain’s revised *F* and *R* constants for predicting aromatic substituent effects. *J. Org. Chem.*, **49**, 1989–1992.
- Rhyu, K.-B., Patel, H.C. and Hopfinger, A.J. (1995) A 3D-QSAR study of anticoccidial triazines using molecular shape analysis. *J. Chem. Inf. Comput. Sci.*, **35**, 771–778.
- Richard, A.J. and Kier, L.B. (1980) SAR analysis of hydrazide monoamine oxidase inhibitors using molecular connectivity. *J. Pharm. Sci.*, **69**, 124.
- Richard, A.M. (1991) Quantitative comparison of molecular electrostatic potentials for structure–activity studies. *J. Comput. Chem.*, **12**, 959–969.

- Richard, A.M. and Benigni, R. (2002) AI and SAR approaches for predicting chemical carcinogenicity: survey and status report. *SAR & QSAR Environ. Res.*, **13**, 1–19.
- Richard, A.M. and Hunter, E.S. (1996) Quantitative structure–activity relationships for the developmental toxicity of haloacetic acids in mammalian whole embryo culture. *Teratology*, **53**, 352–360.
- Richards, F.M. (1977) Areas, volumes, packing, and protein structures. *Ann. Rev. Biophys. Bioeng.*, **6**, 151–176.
- Richards, W.G. (1993) Molecular similarity, in *Trends in QSAR and Molecular Modelling* 92 (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 203–206.
- Richards, W.G. (1995) Molecular similarity and dissimilarity, in *Modelling of Biomolecular Structures and Mechanisms* (eds A. Pullman, J. Jortner and B. Pullman), Kluwer, Dordrecht, The Netherlands, pp. 365–369.
- Richards, W.G. and Hodgkin, E.E. (1988) Molecular similarity. *Chem. Brit.*, **24**, 1141–1144.
- Richet, M.C. (1893) Noté sur la Rapport entre la Toxicité et les Propriétés Physiques des Corps. *Compt. Rend. Soc. Biol. (Paris, French)*, **45**, 775–776.
- Ridings, J.E., Manallack, D.T., Saunders, M.R., Baldwin, J.A. and Livingstone, D.J. (1992) Multivariate quantitative structure–toxicity relationships in a series of dopamine mimetics. *Toxicology*, **76**, 209–217.
- Rios-Santamarina, I., García-Domenech, R., Gálvez, J., Cortijo, J., Santamaria, P. and Marcillo, E. (1998) New bronchodilators selected by molecular topology. *Bioorg. Med. Chem. Lett.*, **8**, 477–482.
- Rishton, G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today*, **8**, 86–96.
- Riviere, J.E. and Brooks, J.D. (2005) Predicting skin permeability from complex chemical mixtures. *Toxicol. Appl. Pharm.*, **208**, 99–110.
- Robbat, A., Jr, Corso, N.P., Doherty, P.J. and Marshall, D. (1986a) Multivariate relationships between gas chromatographic retention index and molecular connectivity of mononitrated polycyclic aromatic hydrocarbons. *Anal. Chem.*, **58**, 2072–2077.
- Robbat, A., Jr, Corso, N.P., Doherty, P.J. and Wolf, M. H. (1986b) Gas chromatographic chemiluminescent detection and evaluation of predictive models for identifying nitrated polycyclic aromatic hydrocarbons in a diesel fuel particulate extract. *Anal. Chem.*, **58**, 2078–2084.
- Robert, D., Amat, L. and Carbó-Dorca, R. (1999) Three-dimensional quantitative structure–activity relationships from tuned molecular quantum similarity measures: prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.*, **39**, 333–344.
- Robert, D. and Carbó-Dorca, R. (1998a) A formal comparison between molecular quantum similarity measures and indices. *J. Chem. Inf. Comput. Sci.*, **38**, 469–475.
- Robert, D. and Carbó-Dorca, R. (1998b) Analyzing the triple density molecular quantum similarity measures with the INDSCAL model. *J. Chem. Inf. Comput. Sci.*, **38**, 620–623.
- Robert, D., Gironés, X. and Carbó-Dorca, R. (1999) Facet diagrams for quantum similarity data. *J. Comput. Aid. Mol. Des.*, **13**, 597–610.
- Roberts, D.W. (1995) Linear free energy relationships for reactions of electrophilic halo- and pseudohalobenzenes, and their application in prediction of skin sensitization potential for S_NAr electrophiles. *Chem. Res. Toxicol.*, **8**, 545–551.
- Roberts, D.W., Friginals, R., Lepoittevin, J.-P. and Benezra, C. (1991) Refinement of the relative alkylation index (RAI) model for skin sensitization and application to mouse and guinea-pig test data for alkyl alkanesulphonates. *Archives of Dermatological Research*, **283**, 387–394.
- Roberts, D.W. and Williams, D.L. (1982) The derivation of quantitative correlations between skin sensitisation and physico-chemical parameters for alkylating agents and their application to experimental data for sultones. *J. Theor. Biol.*, **99**, 807–825.
- Roberts, J.D. and Moreland, W.T. (1953) Electrical effects of substituent groups in saturated systems. Reactivities of 4-substituted bicyclo[2.2.2]octane-1-carboxylic acids. *J. Am. Chem. Soc.*, **75**, 2167–2173.
- Robinson, D.D., Barlow, T.W. and Richards, W.G. (1997a) Reduced dimensional representations of molecular structure. *J. Chem. Inf. Comput. Sci.*, **37**, 939–942.
- Robinson, D.D., Barlow, T.W. and Richards, W.G. (1997b) The utilization of reduced dimensional representations of molecular structure for rapid molecular similarity calculations. *J. Chem. Inf. Comput. Sci.*, **37**, 943–950.
- Robinson, D.D., Lyne, P.D. and Richards, W.G. (1999) Alignment of 3D-structures by the method of 2D-projections. *J. Chem. Inf. Comput. Sci.*, **39**, 594–600.
- Robinson, D.D., Winn, P.J., Lyne, P.D. and Richards, W.G. (1999) Self-organizing molecular field analysis: a tool for structure–activity studies. *J. Med. Chem.*, **42**, 573–583.
- Roche, O., Schneider, P., Zuegge, J., Guba, W., Kansy, M., Alanine, A., Bleicher, K., Danel, F., Gutknecht,

- E.-M., Rogers-Evans, M., Neidhart, W., Stalder, H., Dillon, M., Sjögren, E., Fotuhi, N., Gillespie, P., Goodnow, R., Harris, W., Jones, P., Taniguchi, M., Tsujii, S., von der Saal, W., Zimmermann, G. and Schneider, G. (2002) Development of a virtual screening method for identification of "frequent hitters" in compound libraries. *J. Med. Chem.*, **45**, 137–142.
- Rodrigues, R.deF., Lopes, J.C.D. and Montanari, C.A. (2000) A QSAR study on *Pneumocystis carinii* topoisomerases of bis-benzimidazoles. *Quant. Struct. -Act. Relat.*, **19**, 173–175.
- Rodríguez Delgado, M.A., Sánchez, M.J., González, V. and García-Montelongo, F. (1993) Correlations between retention data of polycyclic aromatic hydrocarbons in micellar liquid chromatography and several molecular descriptors. *Fresen. J. Anal. Chem.*, **345**, 748–752.
- Rodríguez Delgado, M.A., Sánchez, M.J., González, V. and García-Montelongo, F. (1995) Prediction of retention for substituted and unsubstituted polycyclic aromatic hydrocarbons in micellar liquid chromatography in the presence of organic modifiers. *J. Chromat.*, **697**, 71–80.
- Rodríguez, A., Tomas, M.S., Perez, J.J. and Rubio-Martínez, J. (2005) Assessment of the performance of cluster analysis grouping using pharmacophores as molecular descriptors. *J. Mol. Struct. (Theochem)*, **727**, 81–87.
- Rodríguez, J.A. (2005) On the Wiener index and the eccentric distance sum of hypergraphs. *MATCH Commun. Math. Comput. Chem.*, **54**, 209–220.
- Rodríguez, J.A. and Sigarreta, J.M. (2005) On the Randić index and conditional parameters of a graph. *MATCH Commun. Math. Comput. Chem.*, **54**, 403–416.
- Rogers, D. (1991) G/SPLINES: a hybrid of Friedman's multivariate adaptive regression splines (MARS) algorithm with Holland's genetic algorithm, in *The Proceedings of the Fourth International Conference on Genetic Algorithms* (eds R.K. Belew and L.B. Booker), Morgan Kaufmann Publishers, San Francisco, CA.
- Rogers, D. (1992) Data analysis using G/SPLINES, in *Advances in Neural Processing Systems*, Vol. 4 (eds J.E. Moody, S.J. Hanson and R. P. Lippmann), Morgan Kaufmann Publisher, San Mateo, CA.
- Rogers, D. (1995) Genetic function approximation: a genetic approach to building quantitative structure–activity relationship models, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, Barcelona (Spain), pp. 420–426.
- Rogers, D., Brown, R.D. and Hahn, M. (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high throughput screening follow-up. *Journal of Biomolecular Screening*, **10**, 682–686.
- Rogers, D. and Hopfinger, A.J. (1994) Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.*, **34**, 854–866.
- Rohde, B. (2003) Representation and manipulation of stereochemistry, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 206–230.
- Rohrbaugh, R.H. and Jurs, P.C. (1985) Prediction of gas chromatographic retention indexes of selected olefins. *Anal. Chem.*, **57**, 2770–2773.
- Rohrbaugh, R.H. and Jurs, P.C. (1986) Prediction of gas chromatographic retention indexes of polycyclic aromatic compounds and nitrated polycyclic aromatic compounds. *Anal. Chem.*, **58**, 1210–1212.
- Rohrbaugh, R.H. and Jurs, P.C. (1987a) Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Anal. Chim. Acta*, **199**, 99–109.
- Rohrbaugh, R.H. and Jurs, P.C. (1987b) Molecular shape and the prediction of high-performance liquid chromatographic retention indexes of polycyclic aromatic hydrocarbons. *Anal. Chem.*, **59**, 1048–1054.
- Rohrbaugh, R.H., Jurs, P.C., Ashman, W.P., Davis, E. G. and Lewis, J.H. (1988) A structure–activity relationship study of organophosphorus compounds. *Chem. Res. Toxicol.*, **1**, 123–127.
- Romanowska, K. (1992) The application of the graph theoretical method in the QSAR scheme possibilities and limits. *Int. J. Quant. Chem.*, **43**, 175–195.
- Romeiro, N.C., Albuquerque, M.G., de Alencastro, R. B., Ravi, M. and Hopfinger, A.J. (2005) Construction of 4D-QSAR models for use in the design of novel p38-MAPK inhibitors. *J. Comput. Aid. Mol. Des.*, **19**, 385–400.
- Rorije, E. and Peijnenburg, W.J.G.M. (1996) QSARs for oxidation of phenols in the aqueous environment, suitable for risk assessment. *J. Chemom.*, **10**, 79–93.
- Rorije, E., Van Wezel, M.C. and Peijnenburg, W.J.G. M. (1995) On the use of backpropagation neural networks in modeling environmental degradation. *SAR & QSAR Environ. Res.*, **4**, 219–235.
- Ros, F., Pintore, M. and Chrétien, J.R. (2002) Molecular descriptor selection combining genetic

- algorithms and fuzzy logic: application to database mining procedures. *Chemom. Intell. Lab. Syst.*, **63**, 15–26.
- Ros, F., Taboureau, O., Pintore, M. and Chrétien, J.R. (2003) Development of predictive models by adaptive fuzzy partitioning. Application to compounds active on the central nervous system. *Chemom. Intell. Lab. Syst.*, **67**, 29–50.
- Rose, J.R. (2003) Machine learning techniques in chemistry, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1082–1097.
- Rose, K. and Hall, L.H. (2002) Modeling blood–brain barrier partitioning using the electrotopological state. *J. Chem. Inf. Comput. Sci.*, **42**, 651–666.
- Rose, V.S. and Wood, J. (1998) Generalized cluster significance analysis and stepwise cluster significance analysis with conditional probabilities. *Quant. Struct. -Act. Relat.*, **17**, 348–356.
- Rose, V.S., Wood, J. and MacFie, H.J.H. (1991) Single class discrimination using principal component analysis (Scd PCA). *Quant. Struct. -Act. Relat.*, **10**, 359–368.
- Rose, V.S., Wood, J. and MacFie, H.J.H. (1992) Generalized single class discrimination (GSCD). A new method for the analysis of embedded structure–activity relationships. *Quant. Struct. -Act. Relat.*, **11**, 492–504.
- Rosen, R. (1990) An approach to molecular similarity, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 369–382.
- Rosenfeld, V.R. and Gutman, I. (1989) A novel approach to graph polynomials. *MATCH Commun. Math. Comput. Chem.*, **24**, 191–199.
- Rosines, E., Bersuker, I.B. and Boggs, J.E. (2001) Pharmacophore identification and bioactivity prediction for group I metabotropic glutamate receptor agonists by the electron-conformational QSAR method. *Quant. Struct. -Act. Relat.*, **20**, 327–334.
- Rosselli, F.P., Albuquerque, C.N. and da Silva, A.B.F. (2005) Quantum chemical and statistical study of megazol-derived compounds with trypanocidal activity. *Int. J. Quant. Chem.*, **103**, 738–748.
- Rost, B., Liu, J., Przybylski, D., Nair, R., Wrzeszczynski, K.O., Bigelow, H. and Ofra, Y. (2003) Prediction of protein structure through evolution, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1789–1811.
- Roussel, C., Piras, P. and Heitmann, I. (1997) An approach to discriminating 25 commercial chiral stationary phases from structural data sets extracted from a molecular database. *Biomed. Chromatogr.*, **11**, 311–316.
- Rouvray, D.E. (1986a) The role of the topological distance matrix in chemistry, in *Mathematical and Computational Concepts in Chemistry* (ed. N. Trinajstić), Ellis Horwood, Chichester, UK, pp. 295–306.
- Rouvray, D.E. (1988a) Novel applications of topological indices. *J. Mol. Struct. (Theochem)*, **165**, 9–20.
- Rouvray, D.H. (1971) Graph theory in chemistry. *R. Inst. Chem. Rev.*, **4**, 173–195.
- Rouvray, D.H. (1973) The search for useful topological indices in chemistry. *Am. Sci.*, **61**, 729–735.
- Rouvray, D.H. (1975) The value of topological indices in chemistry. *MATCH Commun. Math. Comput. Chem.*, **1**, 125–134.
- Rouvray, D.H. (1976) The topological matrix in quantum chemistry, in *Chemical Applications of Graph Theory* (ed. A.T. Balaban), Academic Press, New York, pp. 175–222.
- Rouvray, D.H. (1983) Should we have designs on topological indices?, in *Chemical Applications of Topology and Graph Theory. Studies in Physical and Theoretical Chemistry* (ed. R.B. King), Elsevier, Amsterdam, The Netherlands, pp. 159–177.
- Rouvray, D.H. (1986b) Predicting chemistry from topology. *Sci. Am.*, **255**, 40–47.
- Rouvray, D.H. (1986c) The prediction of biological activity using molecular connectivity indices. *Acta Pharm. Jugosl.*, **36**, 239–252.
- Rouvray, D.H. (1987) The modeling of chemical phenomena using topological indices. *J. Comput. Chem.*, **8**, 470–480.
- Rouvray, D.H. (1988b) The challenge of characterizing branching in molecular species. *Disc. Appl. Math.*, **19**, 317–338.
- Rouvray, D.H. (1989a) The limits of applicability of topological indices. *J. Mol. Struct. (Theochem)*, **185**, 187–201.
- Rouvray, D.H. (1989b) The pioneering contributions of Cayley and Sylvester to the mathematical description of chemical structure. *J. Mol. Struct. (Theochem)*, **185**, 1–14.
- Rouvray, D.H. (ed.) (1990a) *Computational Chemical Graph Theory*, Nova Press, New York.
- Rouvray, D.H. (1990b) The evolution of the concept of molecular similarity, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 15–42.
- Rouvray, D.H. (1991) The origins of chemical graph theory, in *Chemical Graph Theory. Introduction and*

- Fundamentals* (eds D. Bonchev and D.H. Rouvray), Abacus Press/Gordon and Breach Science Publishers, New York, pp. 1–39.
- Rouvray, D.H. (1992) Definition and role of similarity concepts in the chemical and physical sciences. *J. Chem. Inf. Comput. Sci.*, **32**, 580–586.
- Rouvray, D.H. (1995) A rationale for the topological approach to chemistry. *J. Mol. Struct. (Theochem)*, **336**, 101–114.
- Rouvray, D.H. (ed.) (1997) *Fuzzy Logic in Chemistry*, Academic Press, New York, p. 364.
- Rouvray, D.H. and Balaban, A.T. (1979) Chemical applications of graph theory, in *Applications of Graph Theory* (eds R.J. Wilson and L.W. Beineke), Academic Press, London, UK, pp. 177–221.
- Rouvray, D.H. and El-Basil, S. (1988) Novel applications of topological indices. Part 4. Correlation of arene absorption spectra with the Randić molecular connectivity index. *J. Mol. Struct. (Theochem)*, **165**, 9–20.
- Rouvray, D.H. and Kumazaki, H. (1991) Prediction of molecular flexibility in halogenated alkanes via fractal dimensionality. *J. Math. Chem.*, **7**, 169–185.
- Rouvray, D.H. and Pandey, R.B. (1986) The fractal nature, graph invariants, and physico-chemical properties of normal alkanes. *J. Chim. Phys.*, **85**, 2286–2290.
- Rovero, P., Riganelli, D., Fruci, D., Vigano, S., Pegoraro, S., Revoltella, R., Greco, G., Butler, R., Clementi, S. and Tanigaki, N. (1994) The importance of secondary anchor residue motifs of HLA class I proteins: a chemometric approach. *Mol. Immunol.*, **31**, 549–554.
- Rowberg, K.A., Even, M., Martin, E. and Hopfinger, A.J. (1994) QSAR and molecular shape analyses of three series of 1-(phenylcarbonyl)-2-pyrazoline insecticides. *J. Agr. Food Chem.*, **42**, 374–380.
- Roy, A.B., Basak, S.C., Harriss, D.K. and Magnuson, V.R. (1984) Neighborhood complexities and symmetry of chemical graphs and their biological applications, in *Mathematical Modelling in Science and Technology* (eds X.J.R. Avula, R.E. Kalman, A.I. Liapis and E.Y. Rodin), Pergamon Press, New York, pp. 745–750.
- Roy, A.B., Raychaudhury, C., Ghosh, A., Ray, S.K. and Basak, S.C. (1983) Information-theoretic topological indices of a molecule and their applications in QSAR, in *Quantitative Approaches to Drug Design* (ed. J.C. Dearden), Elsevier, Amsterdam, The Netherlands, pp. 75–76.
- Roy, K. (2004a) Topological descriptors in drug design and modeling studies. *Mol. Div.*, **8**, 321–323.
- Roy, K., De, A.U. and Sengupta, C. (2005) QSAR of antimalarial cyclic peroxy ketals. II. Exploration of pharmacophoric site using AM1 calculations. *Quant. Struct. -Act. Relat.*, **20**, 319–326.
- Roy, K. and Ghosh, G. (2003) Introduction of extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies. *Internet Electron. J. Mol. Des.*, **2**, 599–620.
- Roy, K. and Ghosh, G. (2004a) QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. *J. Chem. Inf. Comput. Sci.*, **44**, 559–567.
- Roy, K. and Ghosh, G. (2004b) QSTR with extended topochemical atom indices. 3. Toxicity of nitrobenzenes to *Tetrahymena pyriformis*. *QSAR Comb. Sci.*, **23**, 99–108.
- Roy, K. and Ghosh, G. (2004c) QSTR with extended topochemical atom indices. 4. Modeling of the acute toxicity of phenylsulfonyl carboxylates to *Vibrio fischeri* using principal component factor analysis and principal component regression analysis. *QSAR Comb. Sci.*, **23**, 526–535.
- Roy, K. and Ghosh, G. (2005) QSTR with extended topochemical atom indices. Part 5. Modeling of the acute toxicity of phenylsulfonyl carboxylates to *Vibrio fischeri* using genetic function approximation. *Bioorg. Med. Chem.*, **13**, 1185–1194.
- Roy, K. and Ghosh, G. (2006a) QSTR with extended topochemical atom (ETA) indices. 8. QSAR for the inhibition of substituted phenols on germination rate of *Cucumis sativus* using chemometric tools. *QSAR Comb. Sci.*, **10**, 846–859.
- Roy, K. and Ghosh, G. (2006b) QSTR with extended topochemical atom (ETA) indices. VI. Acute toxicity of benzene derivatives to tadpoles (*Rana japonica*). *J. Mol. Model.*, **12**, 306–316.
- Roy, K. and Ghosh, G. (2007) QSTR with extended topochemical atom (ETA) indices. 9. Comparative QSAR for the toxicity of diverse functional organic compounds to *Chlorella vulgaris* using chemometric tools. *Chemosphere*, **70**, 1–12.
- Roy, K. and Leonard, J.T. (2005) Classical QSAR modeling of anti-HIV 2,3-diaryl-1,3-thiazolidin-4-ones. *QSAR Comb. Sci.*, **24**, 579–592.
- Roy, K., Pal, D.K., De, A.U. and Sengupta, C. (1999) Comparative QSAR studies with molecular negentropy, molecular connectivity, STIMS and TAU indices. Part I. Tadpole narcosis of diverse functional acyclic compounds. *Indian J. Chem.*, **38**, 664–671.
- Roy, K., Pal, D.K., De, A.U. and Sengupta, C. (2001) Comparative QSAR studies with molecular negentropy, molecular connectivity, STIMS and TAU indices. Part II. General anaesthetic activity of

- aliphatic hydrocarbons, halocarbons and ethers. *Indian J. Chem.*, **40**, 129–135.
- Roy, K. and Saha, A. (2003a) Comparative QSPR studies with molecular connectivity. Molecular negentropy and TAU indices. Part 2. Lipid–water partition coefficient of diverse functional acyclic compounds. *Internet Electron. J. Mol. Des.*, **2**, 288–305.
- Roy, K. and Saha, A. (2003b) Comparative QSPR studies with molecular connectivity, molecular negentropy and TAU indices. Part I. Molecular thermochemical properties of diverse functional acyclic compounds. *J. Mol. Model.*, **9**, 259–270.
- Roy, K. and Saha, A. (2004) QSPR with TAU indices: boiling points of sulfides and thiols. *Indian J. Chem.*, **43**, 1369–1376.
- Roy, K. and Sanyal, I. (2006) QSTR with extended topochemical atom indices. 7. QSAR of substituted benzenes to *Saccharomyces cerevisiae*. *QSAR Comb. Sci.*, **25**, 359–371.
- Roy, K., Sanyal, I. and Ghosh, G. (2007) QSPR of *n*-octanol/water partition coefficient of nonionic organic compounds using extended topochemical atom (ETA) indices. *QSAR Comb. Sci.*, **26**, 629–646.
- Roy, K., Sanyal, I. and Roy, P.P. (2006) QSPR of the bioconcentration factors of non-ionic organic compounds in fish using extended topochemical atom (ETA) indices. *SAR & QSAR Environ. Res.*, **17**, 563–582.
- Roy, K. and Toropov, A.A. (2005) QSPR modeling of the water solubility of diverse functional aliphatic compounds by optimization of correlation weights of local graph invariants. *J. Mol. Model.*, **11**, 89–96.
- Roy, N.K., Nidiry, E.S.J., Vasu, K., Bedi, S., Lalljee, B. and Singh, B. (1996) Quantitative structure–activity relationship studies of O,O-bisaryl alkyl phosphonate fungicides by Hansch approach and principal component analysis. *J. Agr. Food Chem.*, **44**, 3971–3976.
- Roy, R.K. (2004b) On the reliability of global and local electrophilicity descriptors. *J. Phys. Chem. A*, **108**, 4934–4939.
- Roy, R.K., Usha, V., Paulović, J. and Hirao, K. (2005) Are the local electrophilicity descriptors reliable indicators of global electrophilicity trends? *J. Phys. Chem. A*, **109**, 4601–4606.
- Roy, T.A., Krueger, A.J., Mackerer, C.R., Neil, W., Arroyo, A.M. and Yang, J.J. (1998) SAR models for estimating the percutaneous absorption of polynuclear aromatic hydrocarbons. *SAR & QSAR Environ. Res.*, **9**, 171–185.
- Ruch, E. (1972) Algebraic aspects of the chirality phenomenon in chemistry. *Acc. Chem. Res.*, **5**, 49–56.
- Rücker, C., Meringer, M. and Kerber, A. (2004) QSPR using MOLGEN-QSPR: the example of haloalkane boiling points. *J. Chem. Inf. Comput. Sci.*, **44**, 2070–2076.
- Rücker, C., Meringer, M. and Kerber, A. (2005) QSPR using MOLGEN-QSPR: the challenge of fluoroalkane boiling points. *J. Chem. Inf. Model.*, **45**, 74–80.
- Rücker, C. and Rücker, G. (1992) Understanding the properties of isospectral points and pairs in graphs: the concept of orthogonal relation. *J. Math. Chem.*, **9**, 207–238.
- Rücker, C. and Rücker, G. (1994) Mathematical relation between extended connectivity and eigenvector coefficients. *J. Chem. Inf. Comput. Sci.*, **34**, 534–538.
- Rücker, C., Rücker, G. and Bertz, S.H. (2004) Organic synthesis – art or science? *J. Chem. Inf. Comput. Sci.*, **44**, 378–386.
- Rücker, C., Rücker, G. and Meringer, M. (2002) Exploring the limits of graph invariant- and spectrum-based discrimination of (sub)structures. *J. Chem. Inf. Comput. Sci.*, **42**, 640–650.
- Rücker, C., Rücker, G. and Meringer, M. (2007) γ -randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.*, **47**, 2345–2357.
- Rücker, G. and Rücker, C. (1990) Computer perception of constitutional (topological) symmetry: TOPSYM, a fast algorithm for partitioning atoms and pairwise relations among atoms into equivalence classes. *J. Chem. Inf. Comput. Sci.*, **30**, 187–191.
- Rücker, G. and Rücker, C. (1991a) Isocodal and isospectral points, edges, and pairs in graphs and how to cope with them in computerized symmetry recognition. *J. Chem. Inf. Comput. Sci.*, **31**, 422–427.
- Rücker, G. and Rücker, C. (1991b) On using the adjacency matrix power method for perception of symmetry and for isomorphism testing of highly intricate graphs. *J. Chem. Inf. Comput. Sci.*, **31**, 123–126.
- Rücker, G. and Rücker, C. (1993) Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.*, **33**, 683–695.
- Rücker, G. and Rücker, C. (1998) Symmetry-aided computation of the Detour matrix and the Detour index. *J. Chem. Inf. Comput. Sci.*, **38**, 710–714.
- Rücker, G. and Rücker, C. (1999) On topological indices, boiling points, and cycloalkanes. *J. Chem. Inf. Comput. Sci.*, **39**, 788–802.
- Rücker, G. and Rücker, C. (2000) Walk counts, labyrinthicity, and complexity of acyclic and cyclic graphs and molecules. *J. Chem. Inf. Comput. Sci.*, **40**, 99–106.

- Rücker, G. and Rücker, C. (2001) Substructure, subgraph, and walk counts as measures of the complexity of graphs and molecules. *J. Chem. Inf. Comput. Sci.*, **41**, 1457–1462.
- Rücker, G. and Rücker, C. (2003) Walking backward: walk counts of negative order. *J. Chem. Inf. Comput. Sci.*, **43**, 1115–1120.
- Ruedenberg, K. (1958) Theorem on the mobile bond orders of alternant conjugated systems. *J. Chim. Phys.*, **29**, 1232–1233.
- Rughooputh, S.D.D.V. and Rughooputh, H.C.S. (2001) Neural network based chemical structure indexing. *J. Chem. Inf. Comput. Sci.*, **41**, 713–717.
- Rugutt, J.K., Rugutt, K.J. and Berner, D.K. (2001) Limonoids from Nigerian *Harrisonia abyssinica* and their stimulatory activity against *Striga hermonthica* seeds. *J. Nat. Prod.*, **64**, 1434–1438.
- Rui Alves, M. and Oliveira, M.B. (2004) Predictive and interpolative biplots applied to canonical variate analysis in the discrimination of vegetable oils by their fatty acid composition. *J. Chemom.*, **18**, 393–401.
- Ruiz, J. and Pouplana, R. (2002) Theoretical prediction of the phenoxyl radical formation capacity and cyclooxygenase inhibition relationships by phenolic compounds. *Quant. Struct. -Act. Relat.*, **21**, 605–612.
- Rum, G. and Herndon, W.C. (1991) Molecular similarity concepts. 5. Analysis of steroid–protein binding constants. *J. Am. Chem. Soc.*, **113**, 9055–9060.
- Rupp, M., Proschak, E. and Schneider, G. (2007) Kernel approach to molecular similarity based on iterative graph similarity. *J. Chem. Inf. Model.*, **47**, 2280–2286.
- Russell, C.J., Dixon, S.L. and Jurs, P.C. (1992) Computer-assisted study of the relationship between molecular structure and Henry's law constant. *Anal. Chem.*, **64**, 1350–1355.
- Russom, C.L., Bradbury, S.P., Broderius, S.J., Hemmermeister, D.E. and Drummond, R.A. (1997) Predicting modes of toxic action from chemical structure: acute toxicity in the fathed minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.*, **16**, 948–967.
- Russom, C.L., Breton, R.L., Walker, J.D. and Bradbury, S.P. (2003) An overview of the use of quantitative structure–activity relationships for ranking and prioritizing large chemical inventories for environmental risk assessments. *Environ. Toxicol. Chem.*, **22**, 1810–1821.
- Ryan, T.P. (1997) *Modern Regression Methods*, John Wiley & Sons, Inc., New York, p. 516.
- Sabin, J.R., Trickey, S.B., Apell, S.P. and Oddershede, J. (2000) Molecular shape, capacitance, and chemical hardness. *Int. J. Quant. Chem.*, **77**, 358–366.
- Sabljić, A. (1983) Quantitative structure–toxicity relationship of chlorinated compounds: a molecular connectivity investigation. *Bull. Environ. Contam. Toxicol.*, **30**, 80–83.
- Sabljić, A. (1984) Predictions of the nature and strength of soil sorption of organic pollutants by molecular topology. *J. Agr. Food Chem.*, **32**, 243–246.
- Sabljić, A. (1985) Calculation of retention indices by molecular topology. Chlorinated benzenes. *J. Chromat.*, **319**, 1–8.
- Sabljić, A. (1987) On the prediction of soil sorption coefficients of organic pollutants from molecular structures: application of molecular connectivity model. *Environ. Sci. Technol.*, **21**, 358–366.
- Sabljić, A. (1988) Application of molecular topology for the estimation of physical data for environmental chemicals, in *Physical Property Prediction in Organic Chemistry* (eds C. Jochum, M. G. Hicks and J. Sunkel), Springer-Verlag, Berlin, Germany, pp. 335–348.
- Sabljić, A. (1989) Quantitative modeling of soil sorption for xenobiotic chemicals. *Environ. Health Persp.*, **83**, 179–190.
- Sabljić, A. (1990) Topological indices and environmental chemistry, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 61–82.
- Sabljić, A. (1991) Chemical topology and ecotoxicology. *Sci. Total Environ.*, **109/110**, 197–220.
- Sabljić, A. (2001) QSAR models for estimating properties of persistent organic pollutants in evaluation of their environmental fate and risk. *Chemosphere*, **43**, 363–375.
- Sabljić, A., Gusten, H., Hermens, J.L.M. and Opperhuizen, A. (1993) Modeling octanol/water partition coefficients by molecular topology chlorinated benzenes and biphenyls. *Environ. Sci. Technol.*, **27**, 1394–1402.
- Sabljić, A., Güsten, H., Schönherr, J. and Riederer, M. (1990) Modeling plant uptake of airborne organic chemicals. 1. Plant cuticle/water partitioning and molecular connectivity. *Environ. Sci. Technol.*, **24**, 1321–1326.
- Sabljić, A., Güsten, H., Verhaar, H.J.M. and Hermens, J.L.M. (1995) QSAR modeling of soil sorption, improvements and systematics of $\log K_{oc}$ vs. $\log K_{ow}$ correlations. *Chemosphere*, **31**, 4489–4514.

- Sabljic, A. and Horvatic, D. (1993) GRAPH III: a computer program for calculating molecular connectivity indices on microcomputers. *J. Chem. Inf. Comput. Sci.*, **33**, 292–295.
- Sabljic, A. and Piver, W.T. (1992) Quantitative modeling of environmental fate and impact of commercial chemicals. *Environ. Toxicol. Chem.*, **11**, 961–972.
- Sabljic, A. and Protic, M. (1982a) Molecular connectivity: a novel method for prediction of bioconcentration factor of hazardous chemicals. *Chem. -Biol. Inter.*, **42**, 301–310.
- Sabljic, A. and Protic, M. (1982b) Relationship between molecular connectivity indices and soil sorption coefficients of polycyclic aromatic hydrocarbons. *Bull. Environ. Contam. Toxicol.*, **28**, 162–165.
- Sabljic, A. and Protic-Sabljic, M. (1983) Quantitative structure–activity study on the mechanism of inhibition of microsomal *p*-hydroxylation of aniline by alcohols. *Mol. Pharm.*, **23**, 213–218.
- Sabljic, A. and Trinajstić, N. (1981) Quantitative structure–activity relationships: the role of topological indices. *Acta Pharm. Jugosl.*, **31**, 189–214.
- Saçan, M.T. and Balcioglu, I.A. (1996) Prediction of soil sorption coefficient of organic pollutants by the characteristic root index model. *Chemosphere*, **32**, 1993–2001.
- Saçan, M.T. and Inel, Y. (1993) Prediction of aqueous solubility of PCBs related to molecular structure. *Turk. J. Chem.*, **17**, 188–195.
- Saçan, M.T. and Inel, Y. (1995) Application of the characteristic root index model to the estimation of *n*-octanol–water partition coefficients. Polychlorinated biphenyls. *Chemosphere*, **30**, 39–50.
- Sachs, H. (1964) Beziehungen zwischen den in einem Graphen enthaltenen Kreisen und seinem charakterischen Polynom. *Publ. Math. (Debrecen)*, **11**, 119–134.
- Sadowski, J. (2003) 3D structure generation, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 231–261.
- Sadowski, J., Gasteiger, J. and Klebe, G. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.*, **34**, 1000–1008.
- Sadowski, J. and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.*, **41**, 3325–3329.
- Sadowski, J., Wagener, M. and Gasteiger, J. (1995) Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem. Int. Ed. Engl.*, **34**, 2674–2677.
- Safa, F. and Hadjmohammadi, M.R. (2006) Use of topological indices of organic sulfur compounds in quantitative structure–retention relationship study. *QSAR Comb. Sci.*, **24**, 1026–1032.
- Safarpour, M.A., Hemmateenejad, B., Min, R. and Jamali, M. (2003) Quantum chemical-QSAR study of some newly synthesized 1,4-dihydropyridine calcium channel blockers. *QSAR Comb. Sci.*, **22**, 997–1005.
- Safe, S.H. (1990) Polychlorinated biphenyls (PCBs), dibenzo-*p*-dioxins (PCDDs), dibenzofurans (PCDFs), and related compounds: environmental and mechanistic considerations which support the development of toxic equivalency factors (TEFs). *Crit. Rev. Toxicol.*, **21**, 51–88.
- Sagan, B.E., Yeh, Y.-N. and Zhang, P. (1996) The Wiener polynomial of a graph. *Int. J. Quant. Chem.*, **60**, 959–969.
- Sagrado, S. and Cronin, M.T.D. (2006) Diagnostic tools to determine the quality of “transparent” regression-based QSARs: the “modelling power” plot. *J. Chem. Inf. Model.*, **46**, 1523–1532.
- Sahu, K.K., Ravichandran, V., Mourya, V.K. and Agrawal, R.K. (2007) QSAR analysis of caffeoyl naphthalene sulfonamide derivatives as HIV-1 integrase inhibitors. *Med. Chem. Res.*, **15**, 418–430.
- Sahu, P.K. and Lee, S.-L. (2004) Novel information theoretic topological index I_k for unsaturated hydrocarbons. *Chem. Phys. Lett.*, **396**, 465–468.
- Saiakhov, R.D., Stefan, L.R. and Klopman, G. (2000) Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs. *Persp. Drug Disc. Des.*, **19**, 133–155.
- Saiz-Urra, L., Pérez González, M., Fall, Y. and Gomez, G. (2007) Quantitative structure–activity relationship studies of HIV-1 integrase inhibition. 1. GETAWAY descriptors. *Eur. J. Med. Chem.*, **42**, 64–70.
- Saiz-Urra, L., Pérez González, M. and Teixeira, M. (2006) QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection. *Bioorg. Med. Chem.*, **14**, 7347–7358.
- Sak, K., Järv, J. and Karelson, M. (2002) ‘Strain effect’ descriptors for ATP and ADP derivatives with modified phosphate groups. *Computers Chem.*, **26**, 341–346.
- Sakaeda, T., Okamura, N., Nagata, S., Yagami, T., Horinouchi, M., Okumura, K., Yamashita, F. and Hashida, M. (2001) Molecular and

- pharmacokinetic properties of 222 commercially available oral drugs in humans. *Biol. Pharm. Bull.*, **24**, 935–940.
- Sakhartova, O.V. and Shatz, V.D. (1984) Vybor usloviy elyuirovaniya v obrashchenno-phazovoi khromatographii. Priblizhennaya apriornaya otsenka uderzhivaniya poliphunktsionalnykh kislorodsoderzhashchikh soedinenii. *Zhur. Anal. Khim.* (Russian), **39**, 1496.
- Salem, L. (1966) *Molecular Orbital Theory of Conjugated Systems*, Benjamin, New York.
- Salim, N., Holliday, J.D. and Willett, P. (2003) Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.*, **43**, 435–442.
- Salo, M., Sarna, S. and Vuorela, H. (1994) Statistical evaluation of molecular descriptors and quantitative structure–property relationship studies of retinoids. *J. Pharm. Biomed. Anal.*, **12**, 867–874.
- Salo, M., Siren, H., Volin, P., Wiedmer, S. and Vuorela, H. (1996) Structure–retention relationships of steroid hormones in reversed phase liquid chromatography and micellar electrokinetic capillary chromatography. *J. Chromat.*, **728**, 83–88.
- Salt, D.W., Ajmani, S., Chrichton, R. and Livingstone, D.J. (2007a) An improved approximation to the estimation of the critical *F* values in best subset regression. *J. Chem. Inf. Model.*, **47**, 143–149.
- Salt, D.W., Ajmani, S., Chrichton, R. and Livingstone, D.J. (2007b) An improved approximation of the critical *F* values in best subset regression. *J. Chem. Inf. Model.*, **47**, 143–149.
- Salt, D.W., Maccari, L., Botta, M. and Ford, M.G. (2004) Variable selection and specification of robust QSAR models from multicollinear data: arylpiperazinyl derivatives with affinity and selectivity for α_2 -adrenoceptors. *J. Comput. Aid. Mol. Des.*, **18**, 495–509.
- Salt, D.W., Yildiz, N., Livingstone, D.J. and Tinsley, C. J. (1992) The use of artificial neural networks in QSAR. *Pestic. Sci.*, **36**, 161–170.
- Salter, G.J. and Kell, D.B. (1995) Solvent selection for whole cell biotransformations in organic media. *Crit. Rev. Biotechnol.*, **15**, 139–177.
- Salvador, J.M., Hernandez, A., Beltram, A., Duran, R. and Mactutis, A. (1998) Fast partial-differential synthesis of the matching polynomial of C_{72} –100. *J. Chem. Inf. Comput. Sci.*, **38**, 1105–1110.
- Samata, A.K., Ray, S.K., Basak, S.C. and Bose, S.K. (1982) Molecular connectivity and antifungal activity. A quantitative structure–activity relationship study of substituted phenols against skin pathogens. *Arzneim. Forsch.* (German), **32**, 1515–1517.
- Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S. (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.*, **41**, 2481–2491.
- Sanderson, R.T. (1951) An interpretation of bond lengths and a classification of bonds. *Science*, **114**, 670–672.
- Sanderson, R.T. (1952) Electronegativity. I. Orbital electronegativity of neutral atoms. *J. Chem. Educ.*, **29**, 540–546.
- Sanderson, R.T. (1954) Electronegativities in inorganic chemistry. III. *J. Chem. Educ.*, **31**, 238–245.
- Sanderson, R.T. (1955) Relation of stability to Pauling electronegativities. *J. Chim. Phys.*, **23**, 2467–2468.
- Sanderson, R.T. (1971) *Chemical Bonds and Bond Energy*, Academic Press, New York.
- Sanderson, R.T. (1983) *Polar Covalence*, Academic Press, New York.
- Sanderson, R.T. (1988) Principles of electronegativity. Part I. General nature. *J. Chem. Educ.*, **65**, 112–118.
- Sanghvi, T., Ni, N., Mayersohn, M. and Yalkowsky, S. H. (2003) Predicting passive intestinal absorption using a single absorption parameter. *Quant. Struct. -Act. Relat.*, **22**, 247–257.
- Sangster, J. (1997) *Octanol–Water Partition Coefficients: Fundamentals and Physical Chemistry*, John Wiley & Sons, Ltd, Chichester, UK, p. 170.
- Santos, J.C., Chamorro, E., Contreras, R. and Fuentealba, P. (2004) Local reactivity index as descriptor of benzene adsorption in cluster models of exchanged zeolite-Y. *Chem. Phys. Lett.*, **383**, 612–616.
- Santos, J.C., Contreras, R., Chamorro, E. and Fuentealba, P. (2002) Local reactivity index defined through the density of states describes the basicity of alkaline-exchanged zeolites. *J. Chim. Phys.*, **116**, 4311–4316.
- Santos-Filho, O.A. and Hopfinger, A.J. (2001) A search for sources of drug resistance by the 4D-QSAR analysis of a set of antimalarial dihydrofolate reductase inhibitors. *J. Comput. Aid. Mol. Des.*, **15**, 1–12.
- Santos-Filho, O.A. and Hopfinger, A.J. (2002) The 4D-QSAR paradigm: application to a novel set of nonpeptidic HIV protease inhibitors. *Quant. Struct. -Act. Relat.*, **21**, 369–381.
- Santos-Filho, O.A., Hopfinger, A.J. and Zheng, T. (2004) Characterization of skin penetration processes of organic molecules using molecular similarity and QSAR analysis. *Mol. Pharm.*, **1**, 466–476.

- Santos-Filho, O.A., Mishra, R.K. and Hopfinger, A.J. (2001) Free energy force field (FEFF) 3D-QSAR analysis of a set of *Plasmodium falciparum* dihydrofolate reductase inhibitors. *J. Comput. Aid. Mol. Des.*, **15**, 787–810.
- Sanz, F., Giraldo, J. and Manaut, F. (eds) (1995) *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*, Prous Science, Barcelona, Spain, p. 688.
- Sardana, S. and Madan, A.K. (2001) Application of graph theory: relationship of molecular connectivity index, Wiener's index and eccentric connectivity index with diuretic activity. *MATCH Commun. Math. Comput. Chem.*, **43**, 85–98.
- Sardana, S. and Madan, A.K. (2002a) Application of graph theory: relationship of antimycobacterial activity of quinolone derivatives with eccentric connectivity index and Zagreb group parameters. *MATCH Commun. Math. Comput. Chem.*, **45**, 35–53.
- Sardana, S. and Madan, A.K. (2002b) Predicting anti-HIV activity of TIBO derivatives: a computational approach using a novel topological descriptor. *J. Mol. Model.*, **8**, 258–265.
- Sardana, S. and Madan, A.K. (2002c) Predicting anticonvulsant activity of benzamides/benzylamines: computational approach using topological descriptors. *J. Comput. Aid. Mol. Des.*, **16**, 545–550.
- Sardana, S. and Madan, A.K. (2003) Topological models for prediction of antihypertensive activity of substituted benzylimidazoles. *J. Mol. Struct. (Theochem)*, **638**, 41–49.
- Sarkar, R., Roy, A.B. and Sarkar, P.K. (1978) Topological information content of genetic molecules. I. *Math. Biosci.*, **39**, 299–312.
- Sasaki, Y., Kubodera, H., Matuszaki, T. and Umeyama, H. (1991) Prediction of octanol/water partition coefficients using parameters derived from molecular structures. *J. Pharmacobiodyn.*, **14**, 207–214.
- Sasaki, Y., Takagi, T. and Kawaki, H. (1992) On the estimation of the quantitative structure–activity relationships descriptor sigma (σ_0) for aliphatic compound. *Chem. Pharm. Bull.*, **40**, 565–569.
- Sasaki, Y., Takagi, T. and Kawaki, H. (1993) Rational estimation of the QSAR (quantitative structure–activity relationships) descriptors sigma (σ_0), and their applications for medicinals now available. *Chem. Pharm. Bull.*, **41**, 415–423.
- Sasaki, Y., Takagi, T., Kawaki, H. and Iwata, A. (1980) Novel substituent entropy constant σ_{s0} represents the molecular connectivity χ and its related indices. *Chem. Pharm. Bull.*, **31**, 330–332.
- Sasaki, Y., Takagi, T., Yamazato, Y., Iwata, A. and Kawaki, H. (1981) Utility of the substituent entropy constants σ_{s0} in the studies of quantitative structure–activity relationships. *Chem. Pharm. Bull.*, **29**, 3073–3075.
- Satoh, H. (2007) Numerical representation of three-dimensional stereochemical environments using FRAU-descriptors. *Croat. Chem. Acta*, **80**, 217–225.
- Satoh, H., Itono, S., Funatsu, K., Takano, K. and Nakata, T. (1999) A novel method for characterization of three-dimensional reaction fields based on electrostatic and steric interactions toward the goal of quantitative analysis and understanding of organic reactions. *J. Chem. Inf. Comput. Sci.*, **39**, 671–678.
- Satoh, H., Koshino, H., Funatsu, K. and Nakata, T. (2000) Novel canonical coding method for representation of three-dimensional structures. *J. Chem. Inf. Comput. Sci.*, **40**, 622–630.
- Satoh, H., Koshino, H., Funatsu, K. and Nakata, T. (2001) Representation of molecular configurations by CASTcoding method. *J. Chem. Inf. Comput. Sci.*, **41**, 1106–1112.
- Satoh, H., Koshino, H. and Nakata, T. (2002) Extended CAST coding method for exact search of stereochemical structures. *J. Comp. Aided Chem.*, **3**, 48–55.
- Satoh, H., Sacher, O., Nakata, T., Chen, L., Gasteiger, J. and Funatsu, K. (1998) Classification of organic reactions: similarity of reactions based on changes in the electronic features of oxygen atoms at the reaction sites. *J. Chem. Inf. Comput. Sci.*, **38**, 210–219.
- Sauer, W.H.B. and Schwarz, M.K. (2003) Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.*, **43**, 987–1003.
- Savin, A. (2005) The electron localization function (ELF) and its relatives: interpretations and difficulties. *J. Mol. Struct. (Theochem)*, **727**, 127–131.
- Sawada, M., Tsuno, Y. and Yukawa, Y. (1972) The substituent effect. II. Normal substituent constants for polynuclear aryls from the hydrolysis of arylcarbonyl benzoates. *Bull. Chem. Soc. Jap.*, **45**, 1206–1209.
- Saxena, A.K. (1995a) Physico-chemical significance of topological parameters. Connectivity indices and information content. Part 1. Correlation studies in the sets with aromatic and aliphatic substituents. *Quant. Struct. -Act. Relat.*, **14**, 31–38.
- Saxena, A.K. (1995b) Physico-chemical significance of topological parameters. Connectivity indices and information content. Part 2. Correlation

- studies with molar refractivity and lipophilicity. *Quant. Struct. -Act. Relat.*, **14**, 142–150.
- Saxty, G., Woodhead, S.J., Berdini, V., Davies, T.G., Verdonk, M.L., Wyatt, P.G., Boyle, R.G., Barford, D., Downham, R., Garrett, M.D. and Carr, R.A. (2007) Identification of inhibitors of protein kinase B using fragment-based lead discovery. *J. Med. Chem.*, **50**, 2293–2296.
- Schaad, L.J. and Hess, B.A., Jr (1972) Hückel molecular orbital π resonance energies. The question of the σ structure. *J. Am. Chem. Soc.*, **94**, 3068–3074.
- Schaad, L.J. and Hess, B.A., Jr (1977) Theory of linear equations as applied to quantitative structure–activity correlations. *J. Med. Chem.*, **20**, 619–625.
- Schaad, L.J., Hess, B.A., Jr, Purcell, W.P., Cammarata, A., Franke, R. and Kubinyi, H. (1981) Compatibility of the Free–Wilson and Hansch quantitative structure–activity relations. *J. Med. Chem.*, **24**, 900–901.
- Schaper, K.-J., Kunz, B. and Raevsky, O.A. (2003) Analysis of water solubility data on the basis of HYBOT descriptors. Part 2. Solubility of liquid chemicals and drugs. *QSAR Comb. Sci.*, **22**, 943–958.
- Schaper, K.-J., Zhang, H. and Raevsky, O.A. (2001) pH-dependent partitioning of acidic and basic drugs into liposomes – a quantitative structure–activity relationship analysis. *Quant. Struct. -Act. Relat.*, **20**, 46–54.
- Schefzik, S. and Bradley, M. (2004) Comparison of commercially available genetic algorithms: GAs as variable selection tool. *J. Comput. Aid. Mol. Des.*, **18**, 511–521.
- Schefzik, S., Kibbey, C. and Bradley, M.P. (2004) Prediction of HPLC conditions using QSPR techniques: an effective tool to improve combinatorial library design. *J. Comb. Chem.*, **6**, 916–927.
- Schelenz, T., Klunker, J., Bernhardt, T., Schäfer, W. and Dost, J. (2001) Relationships between hydrophobicity and algistatic activity of 5-aryl-3H-[1,3,4]oxadiazole-2-thiones. *Quant. Struct. -Act. Relat.*, **20**, 291–297.
- Schleifer, K.-J. and Tot, E. (2002) CoMFA, CoMSIA and GRID/GOLPE studies on calcium entry blocking 1,4-dihydropyridines. *Quant. Struct. -Act. Relat.*, **21**, 239–248.
- Schleyer, P.V.R., Maerker, C., Dransfeld, A., Jiao, H. and van Eikema Hommes, N.J.R. (1996) Nucleus-independent chemical shifts: a simple and efficient aromaticity probe. *J. Am. Chem. Soc.*, **118**, 6317–6318.
- Schmalz, T.G., Klein, D.J. and Sandleback, B.L. (1992) Chemical graph-theoretical cluster expansion and diamagnetic susceptibility. *J. Chem. Inf. Comput. Sci.*, **32**, 54–57.
- Schmalz, T.G., Živković, T. and Klein, D.J. (1987) Cluster expansion of the Hückel molecular energy of acyclic: applications to PI resonance theory. *Stud. Phys. Theor. Chem.*, **54**, 173–190.
- Schmidli, H. (1997) Multivariate prediction for QSAR. *Chemom. Intell. Lab. Syst.*, **37**, 125–134.
- Schmidt, T.J. and Heilmann, J. (2002) Quantitative structure–cytotoxicity relationships of sesquiterpene lactones derived from partial charge (Q)-based fractional accessible surface area descriptors (Q_frASAs). *Quant. Struct. -Act. Relat.*, **21**, 276–287.
- Schmitt, H., Altenburger, R., Jastorff, B. and Schüürmann, G. (2000) Quantitative structure–activity analysis of the algae toxicity of nitroaromatic compounds. *Chem. Res. Toxicol.*, **13**, 441–450.
- Schmuker, M., Givehchi, A. and Schneider, G. (2004) Impact of different software implementations on the performance of the *Maxmin* method for diverse subset selection. *Mol. Div.*, **8**, 421–425.
- Schneider, G. and Fechner, U. (2005) Computer-based *de novo* design of drug-like molecules. *Nature Reviews. Drug Discovery*, **4**, 649–663.
- Schneider, G., Neidhart, W., Giller, T. and Schmid, G. (1999) “Scaffold-hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed. Engl.*, **38**, 2894–2895.
- Schneider, H.-J., Rüdiger, V. and Raevsky, O.A. (1993) The incremental description of host–guest complexes: free energy increments derived from hydrogen bonds applied to crown ethers and cryptands. *J. Org. Chem.*, **58**, 3648–3653.
- Schnikter, J., Gopalswamy, R. and Crippen, G.M. (1997) Objective models for steroid binding sites of human globulins. *J. Comput. Aid. Mol. Des.*, **11**, 93–110.
- Schnur, D. (1999) Design and diversity analysis of large combinatorial libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.*, **39**, 36–45.
- Schomaker, V. and Stevenson, D.P. (1941) Some revisions of the covalent radii and the additivity rule for the lengths of partially ionic single covalent bonds. *J. Am. Chem. Soc.*, **63**, 37–40.
- Schotte, W. (1992) Prediction of the molar volume at the normal boiling point. *Chem. Eng. J.*, **48**, 167–172.
- Schramke, J.A., Murphy, S.F., Doucette, W.J. and Hintze, W.D. (1999) Prediction of aqueous diffusion coefficients for organic compounds at 25 °C. *Chemosphere*, **38**, 2381–2406.

- Schroeter, T., Schwaighofer, A., Mika, S., ter Laak, A. M., Suelzle, D., Ganzer, U., Heinrich, N. and Müller, K.-R. (2007a) Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput. Aid. Mol. Des.*, **21**, 485–498.
- Schroeter, T., Schwaighofer, A., Mika, S., ter Laak, A. M., Suelzle, D., Ganzer, U., Heinrich, N. and Müller, K.-R. (2007b) Machine learning models for lipophilicity and their domain of applicability. *Mol. Pharm.*, **4**, 524–538.
- Schubert, W. and Ugi, I. (1978) Constitutional symmetry and unique descriptors of molecules. *J. Am. Chem. Soc.*, **100**, 37–41.
- Schuffenhauer, A., Brown, N., Selzer, P., Ertl, P. and Jacoby, E. (2006) Relationships between molecular complexity, biological activity, and structural diversity. *J. Chem. Inf. Model.*, **46**, 525–535.
- Schuffenhauer, A., Floersheim, P., Acklin, P. and Jacoby, E. (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.*, **43**, 391–405.
- Schuffenhauer, A., Gillet, V.J. and Willett, P. (2000) Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.*, **40**, 295–307.
- Schultz, H.P. (1989) Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.*, **29**, 227–228.
- Schultz, H.P. (2000) Topological organic chemistry. 13. Transformation of graph adjacency matrixes to distance matrixes. *J. Chem. Inf. Comput. Sci.*, **40**, 1158–1159.
- Schultz, H.P., Schultz, E.B. and Schultz, T.P. (1990) Topological organic chemistry. 2. Graph theory, matrix determinants and eigenvalues, and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.*, **30**, 27–29.
- Schultz, H.P., Schultz, E.B. and Schultz, T.P. (1992) Topological organic chemistry. 4. Graph theory, matrix permanents, and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.*, **32**, 69–72.
- Schultz, H.P., Schultz, E.B. and Schultz, T.P. (1993) Topological organic chemistry. 7. Graph theory and molecular topological indices of unsaturated and aromatic hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **33**, 863–867.
- Schultz, H.P., Schultz, E.B. and Schultz, T.P. (1994) Topological organic chemistry. 8. Graph theory and topological indices of heteronuclear systems. *J. Chem. Inf. Comput. Sci.*, **34**, 1151–1157.
- Schultz, H.P., Schultz, E.B. and Schultz, T.P. (1995) Topological organic chemistry. 9. Graph theory and molecular topological indices of stereoisomeric organic compounds. *J. Chem. Inf. Comput. Sci.*, **35**, 864–870.
- Schultz, H.P., Schultz, E.B. and Schultz, T.P. (1996) Topological organic chemistry. 10. Graph theory and topological indices of conformational isomers. *J. Chem. Inf. Comput. Sci.*, **36**, 996–1000.
- Schultz, H.P. and Schultz, T.P. (1991) Topological organic chemistry. 3. Graph theory, binary and decimal adjacency matrices, and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.*, **31**, 144–147.
- Schultz, H.P. and Schultz, T.P. (1992) Topological organic chemistry. 5. Graph theory, matrix hafnians and pfaffians, and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.*, **32**, 364–368.
- Schultz, H.P. and Schultz, T.P. (1993) Topological organic chemistry. 6. Graph theory and molecular topological indices of cycloalkanes. *J. Chem. Inf. Comput. Sci.*, **33**, 240–244.
- Schultz, H.P. and Schultz, T.P. (1998) Topological organic chemistry. 11. Graph theory and reciprocal Schultz-type molecular topological indices of alkanes and cycloalkanes. *J. Chem. Inf. Comput. Sci.*, **38**, 853–857.
- Schultz, H.P. and Schultz, T.P. (2000) Topological organic chemistry. 12. Whole-molecule Schultz topological indices of alkanes. *J. Chem. Inf. Comput. Sci.*, **40**, 107–112.
- Schultz, T.W. (1987) The use of the ionization constant (pK_a) in selecting models of toxicity in phenols. *Ecotox. Environ. Safety*, **14**, 178–183.
- Schultz, T.W. (1999) Structure–toxicity relationships for benzenes evaluated with *Tetrahymena pyriformis*. *Chem. Res. Toxicol.*, **12**, 1262–1267.
- Schultz, T.W. and Cajina-Quezada, M. (1987) Structure–activity relationships for monoalkylated or halogenated phenols. *Toxicol. Lett.*, **37**, 121–130.
- Schultz, T.W. and Cronin, M.T.D. (1999) Response–surface analyses for toxicity to *Tetrahymena pyriformis*: reactive carbonyl-containing aliphatic chemicals. *J. Chem. Inf. Comput. Sci.*, **39**, 304–309.
- Schultz, T.W., Cronin, M.T.D. and Netzeva, T.I. (2003a) The present status of QSAR in toxicology. *J. Mol. Struct. (Theochem)*, **622**, 23–28.
- Schultz, T.W., Cronin, M.T.D., Netzeva, T.I. and Aptula, A.O. (2002) Structure–toxicity relationship for aliphatic chemicals evaluated with *Tetrahymena pyriformis*. *Chem. Res. Toxicol.*, **15**, 1602–1609.
- Schultz, T.W., Cronin, M.T.D., Walker, J.D. and Aptula, A.O. (2003b) Quantitative structure–activity relationships (QSARs) in toxicology: a

- historical perspective. *J. Mol. Struct. (Theochem)*, **622**, 1–22.
- Schultz, T.W., Kier, L.B. and Hall, L.H. (1982) Structure–toxicity relationships of selected nitrogenous heterocyclic compounds. III. Relations using molecular connectivity. *Bull. Environ. Contam. Toxicol.*, **28**, 373.
- Schultz, T.W., Lin, D.T., Wilke, T.S. and Arnold, L.M. (1990) Quantitative structure–activity relationships for the *Tetrahymena* Pyriformis population growth endpoint: a mechanism of action approach, in *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (eds W. Karcher and J. Devillers), Kluwer, Dordrecht, The Netherlands, pp. 241–262.
- Schultz, T.W. and Moulton, B.A. (1985) Structure–activity relationships of selected pyridines. I. Substituent constant analysis. *Ecotox. Environ. Safety*, **10**, 97–111.
- Schultz, T.W., Netzeva, T.I., Roberts, D.W. and Cronin, M.T.D. (2005) Structure–toxicity relationships for the effects to *Tetrahymena pyriformis* of aliphatic, carbonyl-containing, α , β -unsaturated chemicals. *Chem. Res. Toxicol.*, **18**, 330–341.
- Schultz, T.W. and Seward, J.R. (2000) Dimyristoyl phosphatidylcholine/water partitioning-dependent modeling of narcotic toxicity to *Tetrahymena pyriformis*. *Quant. Struct. -Act. Relat.*, **19**, 339–344.
- Schultz, T.W., Sinks, G.D. and Bearden, A.P. (1998) QSAR in aquatic toxicology: a mechanism of action approach comparing toxic potency to *Pimephales promelas*, *Tetrahymena pyriformis*, and *Vibrio fischeri*, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 51–109.
- Schuur, J. and Gasteiger, J. (1996) 3D-MoRSE code – a new method for coding the 3D structure of molecules, in *Software Development in Chemistry*, Vol. 10 (ed. J. Gasteiger), Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main, Germany, pp. 67–80.
- Schuur, J. and Gasteiger, J. (1997) Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D structure representation. *Anal. Chem.*, **69**, 2398–2405.
- Schuur, J., Selzer, P. and Gasteiger, J. (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.*, **36**, 334–344.
- Schüürmann, G. (1990) Quantitative structure–property relationships for the polarizability, solvatochromic parameters and lipophilicity. *Quant. Struct. -Act. Relat.*, **9**, 326–333.
- Schüürmann, G. (1995) Quantum chemical approach to estimate physico-chemical compound properties application to substituted benzenes. *Environ. Toxicol. Chem.*, **14**, 2067–2076.
- Schüürmann, G. (1996) Modelling pK_a of carboxylic acids and chlorinated phenols. *Quant. Struct. -Act. Relat.*, **15**, 121–132.
- Schüürmann, G., Aptula, A.O., Kühne, R. and Ebert, R.-U. (2003) Stepwise discrimination between four modes of toxic action of phenols in the *Tetrahymena pyriformis* assay. *Chem. Res. Toxicol.*, **16**, 974–987.
- Schüürmann, G., Flemmig, B. and Dearden, J.C. (1997) CoMFA study of acute toxicity of nitrobenzenes to *Tetrahymena pyriformis*, in *Quantitative Structure–Activity Relationships in Environmental Sciences – VII* (eds F. Chen and G. Schüürmann), SETAC Press, Pensacola, FL, pp. 315–327.
- Schüürmann, G. and Funar-Timofei, S. (2003) Multilinear regression and comparative molecular field analysis (CoMFA) of azo dye–fiber affinities. 2. Inclusion of solution-phase molecular orbital descriptors. *J. Chem. Inf. Comput. Sci.*, **43**, 1502–1512.
- Schüürmann, G., Segner, H. and Jung, K. (1997) Multivariate mode-of-action analysis of acute toxicity of phenols. *Aquat. Toxicol.*, **38**, 277–296.
- Schüürmann, G., Somashekar, R.K. and Kristen, U. (1996) Structure–activity relationships for chlorophenol and nitrophenol toxicity in the pollen tube growth test. *Environ. Toxicol. Chem.*, **15**, 1702–1708.
- Schwaighofer, A., Schroeter, T., Mika, S., Laub, J., ter Laak, A.M., Sülzle, D., Ganzer, U., Heinrich, N. and Müller, K.-R. (2007) Accurate solubility prediction with error bars for electrolytes: a machine learning approach. *J. Chem. Inf. Model.*, **47**, 407–424.
- Schweitzer, R.C. and Morris, J.B. (1999) The development of a quantitative structure–property relationship (QSPR) for the prediction of dielectric constants using neural networks. *Anal. Chim. Acta*, **384**, 285–303.
- Sciabola, S., Alex, A., Higginson, P.D., Mitchell, J.C., Snowden, M.J. and Morao, I. (2005) Theoretical prediction of the enantiomeric excess in asymmetric catalysis. An alignment-independent molecular interaction field based approach. *J. Org. Chem.*, **70**, 9025–9027.
- Sciabola, S., Morao, I. and De Groot, M.J. (2007) Pharmacophoric fingerprint method (TOPP) for

- 3D-QSAR modeling: application to CYP2D6 metabolic stability. *J. Chem. Inf. Model.*, **47**, 76–84.
- Scsibraný, H., Karlovits, M., Demuth, W., Müller, F. and Varmuza, K. (2003) Clustering and similarity of chemical structures represented by binary substructure descriptors. *Chemom. Intell. Lab. Syst.*, **67**, 95–108.
- Scsibraný, H. and Varmuza, K. (1992a) Common substructures in groups of compounds exhibiting similar mass spectra. *Fresen. J. Anal. Chem.*, **344**, 220–222.
- Scsibraný, H. and Varmuza, K. (1992b) Topological similarity of molecules based on maximum common substructures, in *Software Development in Chemistry – Proceedings of the 7th CIC-Workshop “Computers in Chemistry”*, (ed. D. Ziessow) Berlin/Gosen, Germany.
- Sedlar, J., Andelic, I., Gutman, I. and Vukicević, D. (2006) Vindicating the Pauling-bond-order concept. *Chem. Phys. Lett.*, **427**, 418–420.
- Seel, M., Turner, D.B. and Willett, P. (1999) Effect of parameter variations on the effectiveness of HQSAR analyses. *Quant. Struct. -Act. Relat.*, **18**, 245–252.
- Segala, M. and Takahata, Y. (2003) Conformational analyses and SAR studies of antispermatogenic hexahydroindenopyridines. *J. Mol. Struct. (Theochem)*, **633**, 93–104.
- Seibel, G.L. and Kollman, P.A. (1990) Molecular mechanics and the modeling of drug structures, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 125–138.
- Seiler, P. (1974) Interconversion of lipophilicities from hydrocarbon/water systems into octanol/water system. *Eur. J. Med. Chem.*, **9**, 473–479.
- Sekusak, S. and Sabljic, A. (1992) Soil sorption and chemical topology. *J. Math. Chem.*, **11**, 271–280.
- Sekusak, S. and Sabljic, A. (1993) Calculation of retention indices by molecular topology. III. Chlorinated dibenzodioxins. *J. Chromat.*, **628**, 69–79.
- Selassie, C.D. and Klein, T.E. (1998) Comparative quantitative structure–activity relationships (QSAR) of the inhibition of dihydrofolate reductase, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 235–284.
- Sello, G. (1992) A new definition of functional groups and a general procedure for their identification in organic structures. *J. Am. Chem. Soc.*, **114**, 3306–3311.
- Sello, G. (1998) Similarity measures: is it possible to compare dissimilar structures? *J. Chem. Inf. Comput. Sci.*, **38**, 691–701.
- Selzer, P. (2003) Correlation between chemical structure and infrared spectra, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1349–1367.
- Selzer, P. and Ertl, P. (2005) Identification and classification of GPCR ligands using self-organizing neural networks. *QSAR Comb. Sci.*, **24**, 270–276.
- Selzer, P., Gasteiger, J., Thomas, H. and Salzer, R. (2000) Rapid access to infrared reference spectra of arbitrary organic compounds: scope and limitations of an approach to the simulation of infrared spectra by neural networks. *Chem. Eur. J.*, **6**, 920–927.
- Selzer, P., Schuur, J. and Gasteiger, J. (1996) Simulation of IR spectra with neural networks using the 3D-MoRSE code, in *Software Development in Chemistry*, Vol. 10 (ed. J. Gasteiger), Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main, Germany, pp. 293–302.
- Senese, C.L., Duca, J.S., Pan, D., Hopfinger, A.J. and Tseng, Y.J. (2004) 4D-fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci.*, **44**, 1526–1539.
- Senn, P. (1988) The computation of the distance matrix and the Wiener index for graphs of arbitrary complexity with weighted vertices and edges. *Computers Chem.*, **12**, 219–227.
- Senthilkumar, L. and Kolandaivel, P. (2005) Study of effective hardness and condensed Fukui functions using AIM, *ab initio*, and DFT methods. *Mol. Phys.*, **103**, 547–556.
- Seri-Levy, A., Salter, R., West, S. and Richards, W.G. (1994) Shape similarity as a single independent variable in QSAR. *Eur. J. Med. Chem.*, **29**, 687–694.
- Seri-Levy, A., West, S. and Richards, W.G. (1994) Molecular similarity, quantitative chirality, and QSAR for chiral drugs. *J. Med. Chem.*, **37**, 1727–1732.
- Serra, J.R., Jurs, P.C. and Kaiser, K.L.E. (2001) Linear regression and computational neural network prediction of *Tetrahymena* acute toxicity for aromatic compounds from molecular structure. *Chem. Res. Toxicol.*, **14**, 1535–1545.
- Serra, J.R., Thompson, E.D. and Jurs, P.C. (2003) Development of binary classification of structural chromosome aberrations for a diverse set of organic compounds from molecular structure. *Chem. Res. Toxicol.*, **16**, 153–163.
- Serrano, J.L., Marcos, M., Melendez, E., Albano, C., Wold, S. and Elguero, J. (1985) Classification of mesogenic benzalazines by multivariate data analysis. *Acta Chem. Scand.*, **39**, 329–341.

- Seybold, P.G. (1983a) Topological influences on the carcinogenicity of aromatic hydrocarbons. I. The bay region geometry. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **10**, 95–101.
- Seybold, P.G. (1983b) Topological influences on the carcinogenicity of aromatic hydrocarbons. II. Substituent effects. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **10**, 103–108.
- Seybold, P.G., May, M. and Bagal, U.A. (1987) Molecular structure–property relationships. *J. Chem. Educ.*, **64**, 575–581.
- Seydel, J.K. (ed.) (1985) *QSAR and Strategies in the Design of Bioactive Compounds*, Wiley-VCH Verlag GmbH, Weinheim, Germany.
- Shacham, M. and Brauner, N. (1999) Considering precision of experimental data in construction of optimal regression models. *Chem. Eng. Process.*, **38**, 477–486.
- Shacham, M. and Brauner, N. (2003) The SROV program for data analysis and regression model identification. *Computers Chem. Eng.*, **27**, 701–714.
- Shacham, M., Brauner, N., Cholakov, G.S. and Stateva, R.P. (2004) Property prediction by correlations based on similarity of molecular structures. *AIChE*, **50**, 2481–2492.
- Shalabi, A.S. (1991) Random walks: computations and applications to chemistry. *J. Chem. Inf. Comput. Sci.*, **31**, 483–491.
- Shamsipur, M., Ghavami, R., Hemmateenejad, B. and Sharghi, H. (2004) Highly correlating distance-connectivity-based topological indices. 2. Prediction of 15 properties of a large set of alkanes using a stepwise factor selection-based PCR analysis. *QSAR Comb. Sci.*, **23**, 734–753.
- Shamsipur, M., Hemmateenejad, B. and Akhond, M. (2004) Highly correlating distance/connectivity-based topological indices. 1. QSPR studies of alkanes. *Bull. Kor. Chem. Soc.*, **25**, 253–259.
- Shankar Raman, V. and Maranas, C.D. (1998) Optimization in product design with properties correlated with topological indices. *Computers Chem. Eng.*, **22**, 747–763.
- Shannon, C. (1948a) A mathematical theory of communication. Part I. *Bell Syst. Tech. J.*, **27**, 379–423.
- Shannon, C. (1948b) A mathematical theory of communication. Part II. *Bell Syst. Tech. J.*, **27**, 623–656.
- Shannon, C. and Weaver, W. (1949) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL.
- Shao, J. (1993) Linear model selection by cross-validation. *J. Am. Stat. Ass.*, **88**, 486–494.
- Shao, X.-G., Leung, A.K.-M. and Chau, F.-T. (2003) Wavelet: a new trend in chemistry. *Acc. Chem. Res.*, **36**, 276–283.
- Shapiro, B.A. and Zhang, K.Z. (1990) Comparing multiple RNA secondary structure using tree comparisons. *Comput. Appl. Biosci.*, **6**, 309–318.
- Shapiro, S. and Guggenheim, B. (1998a) Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. *Quant. Struct. -Act. Relat.*, **17**, 327–337.
- Shapiro, S. and Guggenheim, B. (1998b) Inhibition of oral bacteria by phenolic compounds. Part 2. Correlations with molecular descriptors. *Quant. Struct. -Act. Relat.*, **17**, 338–347.
- Sharma, V., Goswami, R. and Madan, A.K. (1997) Eccentric connectivity index: a novel highly discriminating topological descriptor for structure–property and structure–activity studies. *J. Chem. Inf. Comput. Sci.*, **37**, 273–282.
- Shatz, V.D., Sakhartova, O.V., Brivkalne, L.A. and Belikov, V.A. (1984) Vybor uslovii elyuirovaniya v obrashchenno-phazovoi khromatographii. Indeks svyazyvaemosti i uderzhivanie uglevodorodov i prosteishikh kislorodsoderzhashchikh soedinemii. *Zhur. Anal. Khim. (Russian)*, **39**, 94.
- Shelley, C.A. and Munk, M.E. (1977) Computer perception of topological symmetry. *J. Chem. Inf. Comput. Sci.*, **17**, 110–113.
- Shemetulskis, N.E., Dunbar, J.B., Jr, Dunbar, B.W., Moreland, D.W. and Humblet, C. (1995) Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput. Aid. Mol. Des.*, **9**, 407–416.
- Shemetulskis, N.E., Weininger, D., Blankley, C.J., Yang, J.J. and Humblet, C. (1996) Stigmata: an algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.*, **36**, 862–871.
- Shen, M., LeTiran, A., Xiao, Y.-D., Golbraikh, A., Kohn, H. and Tropsha, A. (2002) Quantitative structure–activity relationship analysis of functionalized amino acid anticonvulsant agents using *k* nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.*, **45**, 2811–2823.
- Shen, Q., Jiang, J.-H., Tao, J.-C., Shen, G.-L. and Yu, R.-Q. (2005) Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. *J. Chem. Inf. Comput. Sci.*, **45**, 1024–1029.
- Sheridan, R.P. (2000) The centroid approximation for mixtures: calculating similarity and deriving structure–activity relationships. *J. Chem. Inf. Comput. Sci.*, **40**, 1456–1469.

- Sheridan, R.P. (2002) The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.*, **42**, 103–108.
- Sheridan, R.P. (2003) Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci.*, **43**, 1037–1050.
- Sheridan, R.P., Hunt, P. and Culberson, J.C. (2006) Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.*, **46**, 180–192.
- Sheridan, R.P. and Miller, M.D. (1998) A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci.*, **38**, 915–924.
- Sheridan, R.P., Miller, M.D., Underwood, D.J. and Kearsley, S.K. (1996) Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.*, **36**, 128–136.
- Sheridan, R.P., Nilakantan, R., Dixon, J.S. and Venkataraghavan, R. (1986) The ensemble approach to distance geometry: application to the nicotinic pharmacophore. *J. Med. Chem.*, **29**, 899–906.
- Sheridan, R.P., Nilakantan, R., Rusinko, A., III, Bauman, N., Haraki, K. and Venkataraghavan, R. (1989) 3DSEARCH: a system for three-dimensional structure searching. *J. Chem. Inf. Comput. Sci.*, **29**, 255–260.
- Sheridan, R.P., Singh, S.B., Fluder, E.M. and Kearsley, S.K. (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.*, **41**, 1395–1406.
- Sheridan, R.P. and Venkataraghavan, R. (1987) New methods in computer-aided drug design. *Acc. Chem. Res.*, **20**, 322–329.
- Shevade, A.V., Homer, M.L., Taylor, C.J., Zhou, H., Jewell, A.D., Manatt, K.S., Kisor, A.K., Yen, S.-P.S. and Ryan, M.A. (2006) Correlating polymer-carbon composite sensor response with molecular descriptors. *Journal of The Electrochemical Society*, **153**, H209–H216.
- Shi, L.M., Fan, Y., Myers, T.G., O'Connor, P.M., Paull, K.D., Friend, S.H. and Weinstein, J.N. (1998) Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.*, **38**, 189–199.
- Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C. L. and Sheehan, D.M. (2001) QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.*, **41**, 186–195.
- Shi, W., Qian, X., Zhang, R. and Song, G. (2001) Synthesis and quantitative structure–activity relationships of new 2,5-disubstituted-1,3,4-oxadiazoles. *J. Agr. Food Chem.*, **49**, 124–130.
- Shorter, J. (1978) Multiparameter extensions of the Hammett equation, in *Correlation Analysis in Chemistry* (eds N.B. Chapman and J. Shorter), Plenum Press, New York, pp. 119–173.
- Shpilkin, S.A., Smolenskii, E.A. and Zefirov, N.S. (1996) Topological structure of the configuration space and the separation of spin and spatial variables for *N*-electron systems. *J. Chem. Inf. Comput. Sci.*, **36**, 409–412.
- Shusterman, A.J. (1992) Predicting chemical mutagenicity by using quantitative structure–activity relationships. *ACS Symp. Ser.*, **484**, 181–190.
- Shvets, N., Terletskaia, A., Dimoglo, A. and Chumakov, Y. (1999) Study of the electronic and structural features characteristic of 4,5-dihydro-1-phenyl-1*H*-2,4-benzodiazepines demonstrating antiarrhythmic activity. *J. Mol. Struct. (Theochem)*, **463**, 105–110.
- Siegel, S. and Komarmy, J.M. (1960) Quantitative relationships in the reactions of *trans*-4-X-cyclohexanecarboxylic acids and their methyl esters. *J. Am. Chem. Soc.*, **82**, 2547–2553.
- Sild, S. and Karelson, M. (2002) A general QSPR treatment for dielectric constants of organic compounds. *J. Chem. Inf. Comput. Sci.*, **42**, 360–367.
- Silipo, C. and Hansch, C. (1975) Correlation analysis. Its application to the structure–activity relationship of triazines inhibiting dihydrofolate reductase. *J. Am. Chem. Soc.*, **97**, 6849–6861.
- Silipo, C. and Vittoria, A. (1990) Three-dimensional structure of drugs, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 153–204.
- Silipo, C. and Vittoria, A. (eds) (1991) *QSAR: Rational Approaches to the Design of Bioactive Compounds*, Elsevier, Amsterdam, The Netherlands, p. 576.
- Silla, E., Tunon, I. and Pascual-Ahuir, J.L. (1991) GEPO: an improved description of molecular surfaces. II. Computing the molecular area and volume. *J. Comput. Chem.*, **12**, 1077–1088.
- Silverman, B.D. (2000a) The thirty-one benchmark steroids revisited: comparative molecular moment analysis (CoMMA) with principal component regression. *Quant. Struct. -Act. Relat.*, **19**, 237–246.
- Silverman, B.D. (2000b) Three-dimensional moments of molecular property fields. *J. Chem. Inf. Comput. Sci.*, **40**, 1470–1476.
- Silverman, B.D., Pitman, M.C., Platt, D.E. and Rigoutsos, I. (1998) Molecular moment similarity between clozapine and substituted [(4-phenylpiperazinyl)-methyl] benzamides: selective

- dopamine D4 agonist. *J. Comput. Aid. Mol. Des.*, **12**, 525–532.
- Silverman, B.D., Pitman, M.C., Platt, D.E. and Rigoutsos, I. (1999) Molecular moment similarity between several nucleoside analogs of thymidine and thymidine. *J. Biomol. Struct. Dyn.*, **16**, 1169–1175.
- Silverman, B.D. and Platt, D.E. (1996) Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.*, **39**, 2129–2140.
- Silverman, B.D., Platt, D.E., Pitman, M. and Rigoutsos, I. (1998) Comparative molecular moment analysis (CoMMA), in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin.), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 183–198.
- Simmons, K.A. (1999) A simple structure-based calculator for estimating vapor pressure. *J. Agr. Food Chem.*, **47**, 1711–1716.
- Simmons, K.A., Dixon, J.A., Halling, B.P., Plummer, E.L., Plummer, M.J., Tymonko, J.M., Schmidt, R.J., Wyle, M.J., Webster, C.A., Bayer, W.A., Witkowski, D.A., Peters, G.R. and Gravelle, W.D. (1992) Synthesis and activity optimization of herbicidal substituted 4-aryl-1,2,4-triazole-5(1H)-thiones. *J. Agr. Food Chem.*, **40**, 297–305.
- Simon, V., Gasteiger, J. and Zupan, J. (1993) A combined application of two different neural network types for the prediction of chemical reactivity. *J. Am. Chem. Soc.*, **115**, 9148–9159.
- Simon, Z. (1974) Specific interactions. Intermolecular forces, steric requirements, and molecular size. *Angew. Chem. Int. Ed. Engl.*, **13**, 719–727.
- Simon, Z. (1993) MTD and hyperstructure approaches, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 307–319.
- Simon, Z., Badilescu, I.I. and Racovitan, T. (1977) Mapping of dihydrofolate reductase receptor site by correlations with minimal topological (steric) difference. *J. Theor. Biol.*, **66**, 485–495.
- Simon, Z., Balaban, A.T., Ciubotariu, D. and Balaban, T.-S. (1985) QSAR for carcinogenesis by polycyclic aromatic hydrocarbons and derivatives in terms of delocalization energy, minimal sterical differences and topological indices. *Rev. Roum. Chim.*, **30**, 985–1000.
- Simon, Z. and Bohl, M. (1992) Structure–activity relations in gestagenic steroids by the MTD method. The case of hard molecules and soft receptors. *Quant. Struct. -Act. Relat.*, **11**, 23–28.
- Simon, Z., Chiriac, A., Holban, S., Ciubotariu, D. and Mihalas, G.I. (1984) *Minimum Steric Difference. The MTD Method for QSAR Studies*, Research Studies Press, Letchworth, UK, p. 174.
- Simon, Z., Chiriac, A., Motoc, I., Holban, S., Ciubotariu, D. and Szabadaï, Z. (1976) Receptor site mapping. Search strategy of standard for correlations with minimal steric differences. *Studia Biophys.*, **55**, 217–226.
- Simon, Z., Ciubotariu, D. and Balaban, A.T. (1985) Reactivity and stereochemical parameters in QSAR for carcinogenic polycyclic hydrocarbon derivatives, in *QSAR and Strategies in the Design in Bioactive Compounds* (ed. J.K. Seydel), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 370–373.
- Simon, Z., Dragomir, N., Plauchitiu, M.G., Holban, S., Glatt, H. and Kerek, F. (1973) Receptor site mapping for cardiotoxic aglicones by the minimal steric difference method. *Eur. J. Med. Chem.*, **15**, 521–527.
- Simon, Z., Holban, S., Motoc, I., Mracec, M., Chiriac, A., Kerek, F., Ciubotariu, D., Szabadaï, Z., Pop, R. D. and Schwartz, I. (1976) Minimal steric difference in structure–biological activity correlations for α -chymotrypsin-catalyzed hydrolyses. *Studia Biophys.*, **59**, 181–197.
- Simon, Z. and Szabadaï, Z. (1973a) Minimal steric difference parameter and the importance of steric fit for structure–biological activity correlations. *Studia Biophys.*, **39**, 239–252.
- Simon, Z. and Szabadaï, Z. (1973b) Minimal steric difference parameter and the importance of steric fit for quantitative structure–activity correlations. *Studia Biophys.*, **39**, 123–132.
- Simón-Manso, Y. (2005) Linear free-energy relationships and the density functional theory: an analog of the Hammett equation. *J. Phys. Chem. A*, **109**, 2006–2011.
- Singer, J.A. and Purcell, W.P. (1967) Relationships among current quantitative structure–activity models. *J. Med. Chem.*, **10**, 1000–1002.
- Singh, N., Gupta, R.L. and Roy, N.K. (1996) Synthesis and quantitative structure–activity relationships of aryl-2-chloroethyl-methyl phosphate fungicides. *Indian J. Chem.*, **35**, 697–702.
- Singh, P., Ojha, T.N., Sharma, R.C. and Tiwari, K.S. (1993) Quantitative structure–activity relationship study of benzodiazepine receptor ligands. 3. *Indian J. Chem.*, **32**, 555–561.
- Singh, P., Ojha, T.N., Tiwari, S. and Sharma, R.C. (1996) Fujita–Ban and Hansch analyses of A1-adenosine receptor binding and A2-adenosine receptor binding affinities of some 4-amino[1,2,4]triazolo[4,3- α]-quinoxalines. *Indian J. Chem.*, **35B**, 929–934.

- Singh, V.K., Tewari, V.P., Gupta, D.K. and Srivastava, A.K. (1984) Calculation of heat of formation: molecular connectivity and IOC- ω technique. A comparative study. *Tetrahedron*, **40**, 2859–2863.
- Sippl, W. (2002) Binding affinity prediction of novel estrogen receptor ligands using receptor-based 3-D QSAR methods. *Bioorg. Med. Chem.*, **10**, 3741–3755.
- Sippl, W. (2006) 3D-QSAR using the GRID/GOLPE approach, in *Molecular Interaction Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 145–170.
- Sivaraman, N., Srinivasan, T.G. and Vasudeva Rao, P.R. (2001) QSPR modeling for solubility of fullerene (C60) in organic solvents. *J. Chem. Inf. Comput. Sci.*, **41**, 1067–1074.
- Sixt, S., Altschuh, J. and Brüggemann, R. (1995) Quantitative structure–toxicity relationships for 80 chlorinated compounds using quantum chemical descriptors. *Chemosphere*, **30**, 2397–2414.
- Sixt, S., Altschuh, J. and Brüggemann, R. (1996) Estimation of pK_a for organic oxyacids using semiempirical quantum chemical methods, in *Software Development in Chemistry*, Vol. 10 (ed. J. Gasteiger), Fachgruppe Chemie-Information-Computer (CIC), Frankfurt am Main, Germany, pp. 147–153.
- Sjöberg, P. (1997) MOLSURF – a generator of chemical descriptors for QSAR, in *Computer-Assisted Lead Finding and Optimization* (eds H. van de Waterbeemd, B. Testa and G. Folkers), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 81–92.
- Sjöberg, P., Murray, J.S. and Brinck, T. (1990) Average local ionization energies on the molecular surfaces of aromatic systems as guides to chemical reactivity. *Can. J. Chem.*, **68**, 1440–1443.
- Sjögren, M., Li, H., Banner, C., Rafter, J., Westerholm, R. and Rannug, U. (1996) Influence of physical and chemical characteristics of diesel fuels and exhaust emissions on biological effects of particle extracts: a multivariate statistical analysis of ten diesel fuels. *Chem. Res. Toxicol.*, **9**, 197–207.
- Sjöström, M. and Eriksson, L. (1995) Experimental design in synthesis planning and structure–property correlations, applications of statistical experimental design and PLS modeling in QSAR, in *Chemometrics Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 63–90.
- Sjöström, M., Rännar, S. and Wieslander, Å. (1995) Polypeptide sequence property relationships in *Escherichia coli* based on auto cross covariances. *Chemom. Intell. Lab. Syst.*, **29**, 295–305.
- Sjöström, M. and Wold, S. (1985) A multivariate study of the relationship between the genetic code and the physical–chemical properties of amino acids. *J. Mol. Evol.*, **22**, 272–277.
- Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G. and Ebert, C. (1989) Principal properties for aromatic substituents. A multivariate approach for design in QSAR. *Quant. Struct. -Act. Relat.*, **8**, 32–38.
- Skagerberg, B., Sjöström, M. and Wold, S. (1987) Multivariate characterization of amino acids by reversed phase high pressure liquid chromatography. *Quant. Struct. -Act. Relat.*, **6**, 158–164.
- Skorobogatov, V.A. and Dobrynin, A.A. (1988) Metric analysis of graphs. *MATCH Commun. Math. Comput. Chem.*, **23**, 105–151.
- Skorobogatov, V.A., Konstantinova, E.V., Nekrasov, Yu.S., Sukharev, Yu.N. and Tepfer, E.E. (1991) On the correlation between the molecular information topological and mass spectra indices of organometallic compounds. *MATCH Commun. Math. Comput. Chem.*, **26**, 215–228.
- Skvortsova, M.I., Baskin, I.I., Skvortsov, L.A., Palyulin, V.A., Zefirov, N. and Stankevitch, I.V. (1999) Chemical graphs and their basis invariants. *J. Mol. Struct. (Theochem)*, **466**, 211–217.
- Skvortsova, M.I., Baskin, I.I., Slovokhotova, O.L., Palyulin, V.A. and Zefirov, N.S. (1993) Inverse problem in QSAR/QSPR studies for the case of topological indices characterizing molecular shape (Kier indices). *J. Chem. Inf. Comput. Sci.*, **33**, 630–634.
- Skvortsova, M.I., Baskin, I.I., Stankevitch, I.V., Palyulin, V.A. and Zefirov, N.S. (1998) Molecular similarity. 1. Analytical description of the set of graph similarity measures. *J. Chem. Inf. Comput. Sci.*, **38**, 785–790.
- Slater, J.C. and Kirkwood, J.G. (1931) The van der Waals forces in gases. *Phys. Rev.*, **37**, 682–697.
- Slater, J.M. and Paynter, J. (1994) Prediction of gas sensor response using basic molecular parameters. *The Analyst*, **119**, 191–195.
- Small, P.A. (1953) Factors affecting the solubility of polymers. *J. Appl. Chem.*, **3**, 71–80.
- SMARTS Tutorial, Daylight Chemical Information Systems, Santa Fe, NM, <http://www.daylight.com/>.
- Smeeks, F.C. and Jurs, P.C. (1990) Prediction of boiling points of alcohols from molecular structure. *Anal. Chim. Acta*, **233**, 111–119.
- Smellie, A. (2007) General purpose interactive physico-chemical property exploration. *J. Chem. Inf. Model.*, **47**, 1182–1187.
- Smith, C., Payne, V., Doolittle, D.J., Debnath, A.K., Lawlor, T. and Hansch, C. (1992) Mutagenic activity

- of a series of synthetic and naturally occurring heterocyclic amines in *Salmonella*. *Mut. Res.*, **279**, 61–73.
- Smith, E.G. and Baker, P.A. (1975) *The Wiswesser Line-Formula Chemical Notation (WLN)*, Chemical Information Management, Cherry Hill, NJ.
- Smith, P.A., Sorich, M.J., McKinnon, R.A. and Miners, J.O. (2003) Pharmacophore and quantitative structure–activity relationship modeling: complementary approaches for the rationalization and prediction of UDP-glucuronosyltransferase 1A4 substrate selectivity. *J. Med. Chem.*, **46**, 1617–1626.
- Smith, P.J. and Popelier, P.L.A. (2004) Quantitative structure–activity relationships from optimised *ab initio* bond lengths: steroid binding affinity and antibacterial activity of nitrofurans derivatives. *J. Comput. Aid. Mol. Des.*, **18**, 135–143.
- Smith, P.J. and Popelier, P.L.A. (2005) Quantum chemical topology (QCT) descriptors as substitutes for appropriate Hammett constants. *Org. Biomol. Chem.*, **3**, 3399–3407.
- Smith, R.N., Hansch, C. and Ames, M.M. (1975) Selection of a reference partitioning system for drug design work. *J. Pharm. Sci.*, **64**, 599–606.
- Smolenskii, E.A. (1964) Application of the theory of graphs to calculations of the additive structural properties of hydrocarbons. *Russ. J. Phys. Chem.*, **38**, 700–702.
- Snarey, M., Terrett, N.K., Willett, P. and Wilton, D.J. (1997) Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.*, **15**, 372–385.
- Sneath, P.H.A. (1966) Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.*, **12**, 157–195.
- Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy*, Freeman, San Francisco, CA.
- Snyder, R., Sangar, R., Wang, J. and Ekins, S. (2002) Three-dimensional quantitative structure–activity relationship for Cyp2d6 substrates. *Quant. Struct. - Act. Relat.*, **21**, 357–368.
- So, S.-S. and Karplus, M. (1996a) Evolutionary optimization in quantitative structure–activity relationship: an application of genetic neural networks. *J. Med. Chem.*, **39**, 1521–1530.
- So, S.-S. and Karplus, M. (1996b) Genetic neural networks for quantitative structure–activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABA_A receptors. *J. Med. Chem.*, **39**, 5246–5256.
- So, S.-S. and Karplus, M. (1997a) Three-dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. *J. Med. Chem.*, **40**, 4347–4359.
- So, S.-S. and Karplus, M. (1997b) Three-dimensional quantitative structure–activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *J. Med. Chem.*, **40**, 4360–4371.
- So, S.-S. and Karplus, M. (1999) A comparative study of ligand–receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Comput. Aid. Mol. Des.*, **13**, 243–258.
- So, S.-S. and Karplus, M. (2001) Evaluation of designed ligands by a multiple screening method: application to glycogen phosphorylase inhibitors constructed with a variety of approaches. *J. Comput. Aid. Mol. Des.*, **15**, 613–647.
- So, S.-S. and Richards, W.G. (1992) Application of neural networks: quantitative structure–activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl) pyrimidines as DHFR inhibitors. *J. Med. Chem.*, **35**, 3201–3207.
- So, S.-S., van Helden, S.P., van Geerestein, V.J. and Karplus, M. (2000) Quantitative structure–activity relationship studies of progesterone receptor binding steroids. *J. Chem. Inf. Comput. Sci.*, **40**, 762–772.
- Sobczyk, L., Grabowski, S.J. and Krygowski, T.M. (2005) Interrelation between H-bond and π -electron delocalization. *Chem. Rev.*, **105**, 3513–3560.
- Solov'ev, V.P. and Varnek, A. (2003) Anti-HIV activity of HEPT, TIBO, and cyclic urea derivatives: structure–property studies, focused combinatorial library generation, and hits selection using substructural molecular fragments method. *J. Chem. Inf. Comput. Sci.*, **43**, 1703–1719.
- Solov'ev, V.P. and Varnek, A. (2004) Structure–property modeling of metal binders using molecular fragments. *Russ. Chem. Bull.*, **53**, 1434–1445.
- Solov'ev, V.P., Varnek, A. and Wipff, G. (2000) Modeling of ion complexation and extraction using substructural molecular fragments. *J. Chem. Inf. Comput. Sci.*, **40**, 847–858.
- Soltzberg, L.J. and Wilkins, C.L. (1976) Computer recognition of activity class from molecular transforms. *J. Am. Chem. Soc.*, **98**, 4006.
- Soltzberg, L.J. and Wilkins, C.L. (1977) Molecular transforms: a potential tool for structure–activity studies. *J. Am. Chem. Soc.*, **99**, 439–443.
- Soltzberg, L.J., Wilkins, C.L., Kaberline, S.L. and Lam, T.F. (1976) Evaluation and comparison of pattern classifiers for chemical applications. *J. Am. Chem. Soc.*, **98**, 7139–7144.

- Son, S.H., Han, C.K., Ahn, S.K., Yoon, J.H. and No, K. T. (1999) Development of three-dimensional descriptors represented by tensors: free energy of hydration density tensor. *J. Chem. Inf. Comput. Sci.*, **39**, 601–609.
- Song, M., Breneman, C.M., Bi, J., Sukumar, N. and Bennett, K.P. (2002) Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *J. Chem. Inf. Comput. Sci.*, **42**, 1347–1357.
- Song, Y., Coupar, I.M. and Iskander, M.N. (2001) Structural predictions of adenosine 2B antagonist affinity using molecular field analysis. *Quant. Struct. -Act. Relat.*, **20**, 23–30.
- Sørensen, P.B., Brüggemann, R., Carlsen, L., Mogensen, B.B., Kreuger, J. and Pudenz, S. (2003) Analysis of monitoring data of pesticide residues in surface waters using partial order ranking theory. *Environ. Toxicol. Chem.*, **22**, 661–670.
- Soriano, E., Cerdán, S. and Ballesteros, P. (2004) Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct. (Theochem)*, **684**, 121–128.
- Sorich, M.J., McKinnon, R.A., Miners, J.O., Winkler, D.A. and Smith, P.A. (2004) Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. *J. Med. Chem.*, **47**, 5311–5317.
- Šoškić, M., Klaić, B., Magnus, V. and Sabljic, A. (1995) Quantitative structure–activity relationships for *N*-(indol-3-ylacetyl)amino acids used as sources of auxin in plant tissue culture. *Plant Growth Regul.*, **16**, 141–152.
- Šoškić, M. and Plavšić, D. (2001) QSAR study of 1,8-naphthyridin-4-ones as inhibitors of photosystem II. *J. Chem. Inf. Comput. Sci.*, **41**, 1316–1321.
- Šoškić, M., Plavšić, D. and Trinajstić, N. (1996a) 2-Difluoromethylthio-4,6-bis-(monoalkylamino)-1,3,5-triazines as inhibitors of Hill reaction: a QSAR study with orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.*, **36**, 146–150.
- Šoškić, M., Plavšić, D. and Trinajstić, N. (1996b) Link between orthogonal and standard multiple linear regression models. *J. Chem. Inf. Comput. Sci.*, **36**, 829–832.
- Šoškić, M., Plavšić, D. and Trinajstić, N. (1997) Inhibition of the Hill reaction by 2-methylthio-4,6-bis(monoalkylamino)-1,3,5-triazines. A QSAR study. *J. Mol. Struct. (Theochem)*, **394**, 57–65.
- Šoškić, M. and Sabljic, A. (1993) Herbicidal selectivity of (*E*)-3-(2,4-dichlorophenoxy)acrylates: QSAR study with molecular connectivity indexes. *Pestic. Sci.*, **39**, 245–250.
- Šoškić, M. and Sabljic, A. (1995) QSAR study of 4-hydroxypyridine derivatives as inhibitors of the Hill reaction. *Pestic. Sci.*, **45**, 133–141.
- Sotomatsu, T. and Fujita, T. (1989) The steric effect of *ortho* substituents on the acidic hydrolysis of benzamides. *J. Org. Chem.*, **54**, 4443–4448.
- Sotomatsu-Niwa, T. and Ogino, A. (1997) Evaluation of the hydrophobic parameters of the amino acids chains of peptides and their application in QSAR and conformational studies. *J. Mol. Struct. (Theochem)*, **392**, 43–54.
- Sottriffer, C., Stahl, M. and Klebe, G. (2003) The docking problem, in *Handbook of Chemoinformatics*, Vol. 4 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1732–1768.
- SPARTAN, Wavefunction, Inc., 18401 Von Karman Avenue, Suite 370, Irvine, CA.
- Spialter, L. (1963) The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP): a new computer-oriented chemical nomenclature. *J. Am. Chem. Soc.*, **85**, 2012–2013.
- Spialter, L. (1964a) The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP). *J. Chem. Doc.*, **4**, 261–269.
- Spialter, L. (1964b) The atom connectivity matrix characteristic polynomial (ACMCP) and its physico-geometric (topological) significance. *J. Chem. Doc.*, **4**, 269–274.
- Spycher, S., Nendza, M. and Gasteiger, J. (2004) Comparison of different classification methods applied to a mode of toxic action data set. *QSAR Comb. Sci.*, **23**, 779–791.
- Spycher, S., Pellegrini, E. and Gasteiger, J. (2005) Use of structure descriptors to discriminate between modes of toxic action of phenols. *J. Chem. Inf. Model.*, **45**, 200–208.
- Sreenivasa, V. and Kulkarni, V.M. (2002) 3D-QSAR CoMFA and CoMSIA on protein tyrosine phosphatase 1B inhibitors. *Bioorg. Med. Chem.*, **10**, 2267–2282.
- Srinivasan, J., Castellino, A., Bradley, E.K., Eksterowicz, J.E., Grootenhuys, P.D.J., Putta, S. and Stanton, R.V. (2002) Evaluation of a novel shape-based computational filter for lead evolution: application to thrombin inhibitors. *J. Med. Chem.*, **45**, 2494–2500.
- Srivastava, S., Richardson, W.W., Bradley, M.P. and Crippen, G.M. (1993) Three-dimensional receptor modeling using distance geometry and Voronoi polyhedra, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 409–430.

- Stærk, D., Skole, B. and Jørgensen, F.S. (2004) Isolation of a library of aromadendranes from *Landolphia dulcis* and its characterization using the VolSurf approach. *J. Nat. Prod.*, **67**, 799–805.
- Stahl, M. and Böhm, M. (1998) Development of filter functions for protein–ligand docking. *J. Mol. Graph. Model.*, **16**, 121–132.
- Stahl, M. and Mauser, H. (2005) Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Model.*, **45**, 542–548.
- Stahura, F.L., Godden, J.W. and Bajorath, J. (2002) Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. *J. Chem. Inf. Comput. Sci.*, **42**, 550–558.
- Stahura, F.L., Godden, J.W., Xue, L. and Bajorath, J. (2000) Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.*, **40**, 1245–1252.
- Štambuk, N. (1999) On circular coding properties of gene and protein sequences. *Croat. Chem. Acta*, **72**, 999–1008.
- Štambuk, N. (2000) Universal metric properties of the genetic code. *Croat. Chem. Acta*, **73**, 1123–1139.
- Stanforth, R.W., Kolossov, E. and Mirkin, B. (2007) A measure of domain of applicability for QSAR modelling based on intelligent *k*-means clustering. *QSAR Comb. Sci.*, **26**, 837–844.
- Stankevitch, I.V., Skvortsova, M.I. and Zefirov, N.S. (1995) On a quantum chemical interpretation of molecular connectivity indices for conjugated hydrocarbons. *J. Mol. Struct. (Theochem)*, **342**, 173–179.
- Stankevitch, M.I., Stankevitch, I.V. and Zefirov, N.S. (1988) Topological indices in organic chemistry. *Russ. Chem. Rev.*, **57**, 191–208.
- Stanley, H.E. and Ostrowsky, N. (eds) (1990) *Correlations and Connectivity: Geometric Aspects of Physics, Chemistry, and Biology*, Kluwer, Dordrecht, The Netherlands.
- Stanton, D.T. (1999) Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.*, **39**, 11–20.
- Stanton, D.T. (2000) Development of a quantitative structure–property relationship model for estimating normal boiling points of small multifunctional organic molecules. *J. Chem. Inf. Comput. Sci.*, **40**, 81–90.
- Stanton, D.T. (2003) On the physical interpretation of QSAR models. *J. Chem. Inf. Comput. Sci.*, **43**, 1423–1433.
- Stanton, D.T., Egolf, L.M., Jurs, P.C. and Hicks, M.G. (1992) Computer-assisted prediction of normal boiling points of pyrans and pyrroles. *J. Chem. Inf. Comput. Sci.*, **32**, 306–316.
- Stanton, D.T. and Jurs, P.C. (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. *Anal. Chem.*, **62**, 2323–2329.
- Stanton, D.T. and Jurs, P.C. (1992) Computer-assisted study of the relationship between molecular structure and surface tension of organic compounds. *J. Chem. Inf. Comput. Sci.*, **32**, 109–115.
- Stanton, D.T., Jurs, P.C. and Hicks, M.G. (1991) Computer-assisted prediction of normal boiling points of furans, tetrahydrofurans, and thiophenes. *J. Chem. Inf. Comput. Sci.*, **31**, 301–310.
- Stanton, D.T., Mattioni, B.E., Knittel, J.J. and Jurs, P.C. (2004) Development and use of hydrophobic surface area (HSA) descriptors for computer-assisted quantitative structure–activity and structure–property relationship studies. *J. Chem. Inf. Comput. Sci.*, **44**, 1010–1023.
- Stanton, D.T., Morris, T.W., Roychoudhury, S. and Parker, C.N. (1999) Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. *J. Chem. Inf. Comput. Sci.*, **39**, 21–27.
- Stanton, D.T., Murray, W.J. and Jurs, P.C. (1993) Comparison of QSAR and molecular similarity approaches for a structure–activity relationship of DHFR inhibitors. *Quant. Struct. -Act. Relat.*, **12**, 239–245.
- Štefanič-Petek, A., Krbavčić, A. and Šolmajer, T. (2002) QSAR of flavonoids. 4. Differential inhibition of aldose reductase and p56^{lck} protein tyrosine kinase. *Croat. Chem. Acta*, **75**, 517–529.
- Stefanis, E., Constantinou, L. and Panayiotou, C. (2004) A group-contribution method for predicting pure component properties of biochemical and safety interest. *Ind. Eng. Chem. Res.*, **43**, 6253–6261.
- Stegeman, M.H., Peijnenburg, W.J.G.M. and Verboom, H.H. (1993) A quantitative structure–activity relationship for the direct photohydrolysis of *meta* substituted halobenzene derivatives in water. *Chemosphere*, **26**, 837–849.
- Stein, S.E., Babushok, V.I., Brown, R.L. and Linstrom, P.J. (2007) Estimation of Kováts retention indices using group contributions. *J. Chem. Inf. Model.*, **47**, 975–980.
- Stein, T.M., Gordon, S.H. and Greene, R.V. (1999) Amino acids as plasticizers. II. Use of quantitative structure–property relationships to predict the behavior of monoammonium monocarboxylate

- plasticizers in starch–glycerol blends. *Carbohydr. Polym.*, **39**, 7–16.
- Steinbeck, C. (2003) Correlation between chemical structures and NMR data, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1368–1377.
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. and Willighagen, E. (2003) The chemistry development kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.
- Stenberg, P., Norinder, U., Luthman, K. and Artursson, P. (2001) Experimental and computational screening models for the prediction of intestinal drug absorption. *J. Med. Chem.*, **44**, 1927–1937.
- Stephenson, K. and Zelen, M. (1989) Rethinking centrality: methods and applications. *Social Networks*, **11**, 1–37.
- Stewart, J.P. (1990) MOPAC: a semiempirical molecular orbital program. *J. Comput. Aid. Mol. Des.*, **4**, 1–105.
- Steyaert, G., Lisa, G., Gaillard, P., Boss, G., Reymond, F., Girault, H.H., Carrupt, P.-A. and Testa, B. (1997) Intermolecular forces expressed in 1,2-dichloroethane–water partition coefficients. A solvatochromic analysis. *J. Chem. Soc. Faraday Trans.*, **93**, 401–406.
- Stiefl, N. and Baumann, K. (2003) Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure–activity relationship technique. *J. Med. Chem.*, **46**, 1390–1407.
- Stiefl, N., Bringmann, G., Rummey, C. and Baumann, K. (2003) Evaluation of extended parameter sets for the 3D-QSAR technique MaP: implications for interpretability and model quality exemplified by antimalarially active naphthylisoquinoline alkaloids. *J. Comput. Aid. Mol. Des.*, **17**, 347–365.
- Still, W.C., Tempczyk, A., Hawley, R.C. and Hendrickson, T. (1990) Semi-analytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **112**, 6127–6129.
- Stoklosa, H.J. (1973) Computer program for calculation of charge distributions in molecules. *J. Chem. Educ.*, **50**, 290.
- Stone, M. (1974) Cross-validatory choice and assessment of statistical predictors. *J. R. Stat. Soc., B*, **36**, 111–147.
- Stouch, T.R. and Jurs, P.C. (1986) A simple method for the representation, quantification, and comparison of the volumes and shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **26**, 4–12.
- Streich, W.J., Dove, S. and Franke, R. (1980) On the rotational selection of test series. 1. Principal component method combined with multidimensional mapping. *J. Med. Chem.*, **23**, 1452–1456.
- Streich, W.J. and Franke, R. (1985) Topological pharmacophores. New methods and their applications to a set of antimalarials. Part 1. The methods LOGANA and LOCON. *Quant. Struct. - Act. Relat.*, **4**, 13–18.
- Streitweiser, A., Jr (1961) *Molecular Orbital Theory for Organic Chemists*, John Wiley & Sons, Inc., New York.
- Stuer-Lauridsen, F. and Pedersen, F. (1997) On the influence of the polarity index of organic matter in predicting environmental sorption of chemicals. *Chemosphere*, **35**, 761–773.
- Stuper, A.J. and Jurs, P.C. (1975) Classification of psychotropic drugs as sedatives or tranquilizers using pattern recognition techniques. *J. Am. Chem. Soc.*, **97**, 182–187.
- Stuper, A.J. and Jurs, P.C. (1978) Structure–activity studies of barbiturates using pattern recognition techniques. *J. Pharm. Sci.*, **67**, 745–751.
- Sudgen, S. (1924) The variation of surface tension with temperature and some related functions. *J. Chem. Soc.*, **125**, 32–41.
- Sulea, T., Kurunczi, L., Oprea, T.I. and Simon, Z. (1998) MTD-ADJ: a multiconformational minimal topologic difference for determining bioactive conformers using adjusted biological activities. *J. Comput. Aid. Mol. Des.*, **12**, 133–146.
- Sulea, T., Kurunczi, L. and Simon, Z. (1995) Dioxin-type activity for polyhalogenated aryl derivatives. A QSAR model based on MTD method. *SAR & QSAR Environ. Res.*, **3**, 37–61.
- Sulea, T., Oprea, T.I., Muresan, S. and Chan, S.L. (1997) A different method for steric field evaluation in CoMFA improves model robustness. *J. Chem. Inf. Comput. Sci.*, **37**, 1162–1170.
- Sulea, T. and Purisima, E.O. (1999) Desolvation free energy field derived from boundary element continuum dielectric calculations. *Quant. Struct. - Act. Relat.*, **18**, 154–158.
- Sullivan, J.J., Jones, A.D. and Tanji, K.K. (2000) QSAR treatment of electronic substituent effects using frontier orbital theory and topological parameters. *J. Chem. Inf. Comput. Sci.*, **40**, 1113–1127.
- Sun, H. (2005) A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.*, **48**, 4031–4039.
- Sun, H., Huang, G. and Dai, S. (1996) Adsorption behaviour and QSPR studies of organotin compounds on estuarine sediment. *Chemosphere*, **33**, 831–838.

- Sun, L., Zhou, Y., Genrong, L. and Li, S.Z. (2004) Molecular electronegativity-distance vector (MEDV-4): a two-dimensional QSAR method for the estimation and prediction of biological activities of estradiol derivatives. *J. Mol. Struct. (Theochem)*, **679**, 107–113.
- Sundaram, A. and Venkatasubramanian, V. (1998) Parametric sensitivity and search-space characterization studies of genetic algorithms for computer-aided polymer design. *J. Chem. Inf. Comput. Sci.*, **38**, 1177–1191.
- Suresh, C.H. and Gadre, S.R. (1998) A novel electrostatic approach to substituent constants: doubly substituted benzenes. *J. Am. Chem. Soc.*, **120**, 7049–7055.
- Susarla, S., Masunaga, S. and Yonezawa, Y. (1996) Kinetics of halogen substituted aniline transformation in anaerobic estuarine sediment. *Water Sci. Technol.*, **34**, 37–43.
- Sutherland, J.J. and Weaver, D.F. (2003) Development of quantitative structure–activity relationships and classification models for anticonvulsant activity of hydantoin analogues. *J. Chem. Inf. Comput. Sci.*, **43**, 1028–1036.
- Sutherland, J.J. and Weaver, D.F. (2004) Three-dimensional quantitative structure–activity and structure–selectivity relationships of dihydrofolate reductase inhibitors. *J. Comput. Aid. Mol. Des.*, **18**, 309–331.
- Sutter, J.M., Dixon, J.S. and Jurs, P.C. (1995) Automated descriptor selection for quantitative structure–activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.*, **35**, 77–84.
- Sutter, J.M. and Jurs, P.C. (1995) Selection of molecular descriptors for quantitative structure–activity relationships, in *Adaption of Simulated Annealing to Chemical Optimization Problems* (ed. J.H. Kalivas), Elsevier, Amsterdam, The Netherlands, pp. 111–132.
- Sutter, J.M. and Jurs, P.C. (1996) Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure–property relationship. *J. Chem. Inf. Comput. Sci.*, **36**, 100–107.
- Sutter, J.M., Peterson, T.A. and Jurs, P.C. (1997) Prediction of gas chromatographic retention indices of alkylbenzene. *Anal. Chim. Acta*, **342**, 113–122.
- Suzuki, T. (1991) Development of an automatic estimation system for both the partition coefficient and aqueous solubility. *J. Comput. Aid. Mol. Des.*, **5**, 149–166.
- Suzuki, T. (2001) A nonlinear group contribution method for predicting the free energies of inclusion complexation of organic molecules with α - and β -cyclodextrins. *J. Chem. Inf. Comput. Sci.*, **41**, 1266–1273.
- Suzuki, T., Ide, K., Ishida, M. and Shapiro, S. (2001) Classification of environmental estrogens by physico-chemical properties using principal component analysis and hierarchical cluster analysis. *J. Chem. Inf. Comput. Sci.*, **41**, 718–726.
- Suzuki, T., Ishida, M. and Fabian, W.M.F. (2000) Classical QSAR and comparative molecular field analyses of the host–guest interaction of organic molecules with cyclodextrins. *J. Comput. Aid. Mol. Des.*, **14**, 669–678.
- Suzuki, T. and Kudo, Y. (1990) Automated log *P* estimation based on combined additive modeling methods. *J. Comput. Aid. Mol. Des.*, **4**, 155–198.
- Suzuki, T., Ohtaguchi, K. and Koide, K. (1992a) Application of principal components analysis to calculate Henry's constant from molecular structure. *Computers Chem.*, **16**, 41–52.
- Suzuki, T., Ohtaguchi, K. and Koide, K. (1992b) Computer-aided prediction of solubilities of organic compounds in water. *J. Chem. Eng. Jpn.*, **25**, 729–734.
- Suzuki, T., Ohtaguchi, K. and Koide, K. (1992c) Correlation between solubilities in water and molecular descriptors of hydrocarbons. *J. Chem. Eng. Jpn.*, **25**, 434–438.
- Suzuki, T., Timofei, S., Iuoras, B.E., Uray, G., Verdino, P. and Fabian, W.M.F. (2001a) Quantitative structure–enantioselective retention relationships for chromatographic separation of arylalkylcarbinols on Pirkle type chiral stationary phases. *J. Chromat.*, **922**, 13–23.
- Suzuki, T., Timofei, S., Kurunczi, L., Dietze, U. and Schüürmann, G. (2001b) Correlation of aerobic biodegradability of sulfonated azo dyes with the chemical structure. *Chemosphere*, **45**, 1–9.
- Svetnik, V., Liaw, A., Tong, C., Culberson, C., Sheridan, R.P. and Feuston, B.P. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.*, **43**, 1947–1958.
- Svozil, D., Sevcik, J.G. and Kvasnicka, V. (1997) Neural network prediction of the solvatochromic polarity/polarizability parameter π^H . *J. Chem. Inf. Comput. Sci.*, **37**, 338–342.
- Swaan, P.W., Koops, B.C., Moret, E.E. and Tukker, J.J. (1998) Mapping the binding site of the small intestinal peptide carrier (PepT1) using comparative molecular field analysis. *Receptor Channel*, **6**, 189.
- Swaan, P.W., Szoka, F.C., Jr and Øie, S. (1997) Molecular modeling of the intestinal bile acid

- carrier: a comparative molecular field analysis study. *J. Comput. Aid. Mol. Des.*, **11**, 581–588.
- Swain, C.G. (1984) Substituent and solvent effects on chemical reactivity. *J. Org. Chem.*, **49**, 2005–2010.
- Swain, C.G. and Lupton, E.C., Jr (1968) Field and resonance components of substituent effects. *J. Am. Chem. Soc.*, **90**, 4328–4337.
- Swain, C.G., Unger, S.H., Rosenquist, N.R. and Swain, M.S. (1983) Substituent effects on chemical reactivity. Improved evaluation of field and resonance components. *J. Am. Chem. Soc.*, **105**, 492–502.
- Swinborne-Sheldrake, R., Herndon, W.C. and Gutman, I. (1975) Kekulé structures and resonance energies of benzenoid hydrocarbons. *Tetrahedron Lett.*, **10**, 755–758.
- SYBYL Force Field, Ver. 6.1, Tripos Associates, Inc., 1699 S Hanley Road, Suite 303, St. Louis, MO.
- Sylvester, J.J. (1877) Chemistry and algebra. *Nature*, **17**, 284–309.
- Sylvester, J.J. (1878) On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics. *Am. J. Math.*, **1**, 64–125.
- Szabo, A. and Ostlund, N.S. (1996) *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Dover Publications, New York, p. 480.
- Szántai-Kis, C., Kövesdi, I., Kéri, G. and Örfi, L. (2003) Validation subset selection for extrapolation oriented QSPAR models. *Mol. Div.*, **7**, 37–43.
- Szász, Gy., Papp, O., Vámos, J., Hankó-Novák, K. and Kier, L.B. (1983) Relationships between molecular connectivity indices, partition coefficients and chromatographic parameters. *J. Chromat.*, **269**, 91–95.
- Szatyłowicz, H., Krygowski, T.M. and Zachara-Horeglad, J.E. (2007) Long-distance structural consequences of H-bonding. How H-bonding affects aromaticity of the ring in variously substituted aniline/anilinium/anilide complexes with bases and acids. *J. Chem. Inf. Model.*, **47**, 875–886.
- Szymanski, K., Müller, W.R., Sabljic, A., Trinajstić, N. and Carter, S. (1987) On the use of the weighted identification numbers in QSAR study of the toxicity of aliphatic ethers. *Int. J. Quantum Chem. Quant. Biol. Symp.*, **14**, 325–330.
- Szymanski, K., Müller, W.R., von Knop, J. and Trinajstić, N. (1985) On Randić's molecular identification numbers. *J. Chem. Inf. Comput. Sci.*, **25**, 413–415.
- Szymanski, K., Müller, W.R., von Knop, J. and Trinajstić, N. (1986a) Molecular ID numbers. *Croat. Chem. Acta*, **59**, 719–723.
- Szymanski, K., Müller, W.R., von Knop, J. and Trinajstić, N. (1986b) On the identification numbers for chemical structures. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **20**, 173–183.
- Tabaraki, R., Khayamian, T. and Ensafi, A.A. (2006) Wavelet neural network modeling in QSPR for prediction of solubility of 25 anthraquinone dyes at different temperatures and pressures in supercritical carbon dioxide. *J. Mol. Graph. Model.*, **25**, 46–54.
- Taft, R.W. (1952) Polar and steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. *J. Am. Chem. Soc.*, **74**, 3120–3128.
- Taft, R.W. (1953a) Linear steric energy relationships. *J. Am. Chem. Soc.*, **75**, 4538–4539.
- Taft, R.W. (1953b) The general nature of the proportionality of polar effects of substituent groups in organic chemistry. *J. Am. Chem. Soc.*, **75**, 4231–4238.
- Taft, R.W. (1953c) The separation of relative free energies of activation to three basic contributing factors and the relationship of these to structure. *J. Am. Chem. Soc.*, **75**, 4534–4537.
- Taft, R.W. (1956) Separation of polar, steric, and resonance effects in reactivity, in *Steric Effects in Organic Chemistry* (ed. M.S. Newman), John Wiley & Sons, Inc., New York, pp. 556–675.
- Taft, R.W. (1960) Sigma values from reactivities. *J. Phys. Chem.*, **64**, 1805–1815.
- Taft, R.W. (1983) Protonic acidities and basicities in the gas phase and in solution: substituent and solvent effects. *Prog. Phys. Org. Chem.*, **14**, 247–350.
- Taft, R.W., Abboud, J.-L.M. and Kamlet, M.J. (1984) Linear solvation energy relationships. 28. An analysis of Swain's solvent "acidity" and "basicity" scales. *J. Org. Chem.*, **49**, 2001–2005.
- Taft, R.W., Abboud, J.-L.M., Kamlet, M.J. and Abraham, M.H. (1985) Linear solvation energy relations. *J. Solut. Chem.*, **14**, 153–186.
- Taft, R.W., Abraham, M.H., Famini, G.R., Doherty, R. M. and Kamlet, M.J. (1985) Solubility properties in polymers and biological media. 5. An analysis of the physico-chemical properties which influence octanol–water partition coefficients of aliphatic and aromatic solutes. *J. Pharm. Sci.*, **74**, 807–814.
- Taft, R.W., Ehrenson, S., Lewis, I.C. and Glick, R.E. (1959) Evaluation of resonance effects on reactivity by application of the linear inductive energy relationship. VI. Concerning the effects of polarization and conjugation on the mesomeric order. *J. Am. Chem. Soc.*, **81**, 5352–5361.
- Taft, R.W. and Grob, C.A. (1974) Concerning the separation of polar and resonance effects in the

- ionization of 4-substituted pyridinium ions. *J. Am. Chem. Soc.*, **96**, 1236–1238.
- Taft, R.W. and Kamlet, M.J. (1976) The solvatochromic comparison method. 2. The α -scale of solvent hydrogen-bond donor (HBD) acidities. *J. Am. Chem. Soc.*, **98**, 2886–2894.
- Taft, R.W. and Kamlet, M.J. (1979) Linear solvation energy relationships. Part 4. Correlations with and limitations of the α scale of solvent hydrogen bond donor acidities. *J. Chem. Soc. Perkin Trans. 2*, 1723–1729.
- Taft, R.W. and Lewis, I.C. (1958) The general applicability of a fixed scale of inductive effects. II. Inductive effects of dipolar substituents in the reactivities of *m*- and *p*-substituted derivatives of benzene. *J. Am. Chem. Soc.*, **80**, 2436–2443.
- Taft, R.W. and Lewis, I.C. (1959) Evaluation of resonance effects on reactivity by application of the linear inductive energy relationship. V. Concerning a σ_R scale of resonance effects. *J. Am. Chem. Soc.*, **81**, 5343–5352.
- Taft, R.W., Price, E., Fox, I.R., Lewis, I.C., Andersen, K.K. and Davis, G.T. (1963a) Fluorine nuclear magnetic resonance shielding in *meta*-substituted fluorobenzenes. The effect of solvent on the inductive order. *J. Am. Chem. Soc.*, **85**, 709–724.
- Taft, R.W., Price, E., Fox, I.R., Lewis, I.C., Andersen, K.K. and Davis, G.T. (1963b) Fluorine nuclear magnetic resonance shielding in *p*-substituted fluorobenzenes. The influence of structure and solvent on resonance effects. *J. Am. Chem. Soc.*, **85**, 3146–3156.
- Taft, R.W. and Topsom, R.D. (1987) The nature and analysis of substituent electronic effects. *Prog. Phys. Org. Chem.*, **16**, 1–84.
- Taillander, G., Domard, M. and Boucherle, A. (1983) QSAR et Séries Aromatiques: Propositions de Paramètres Stériques. *Il Farmaco*, **38**, 473–487.
- Tairi-Kellou, S., Cartier, A., Maouche, B. and Maigret, B. (2001) Electronic descriptors of the 1,4-benzodiazepine derivatives related to the antagonist activity on cholecystokinin receptors. *J. Mol. Struct. (Theochem)*, **571**, 207–223.
- Takahashi, Y., Miashita, Y., Tanaka, Y., Hayasaka, H., Abe, H. and Sasaki, S. (1985) Discriminative structural analysis using pattern recognition techniques in the structure–taste problem of perillartines. *J. Pharm. Sci.*, **73**, 737–741.
- Takahashi, Y., Sukekawa, M. and Sasaki, S.I. (1992) Automatic identification molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci.*, **32**, 639–643.
- Takahata, Y., Andreazza Costa, M.C. and Gaudio, A.C. (2003) Comparison between neural networks (NN) and principal component analysis (PCA): structure–activity relationships of 1,4-dihydropyridine calcium channel antagonists (nifedipine analogues). *J. Chem. Inf. Comput. Sci.*, **43**, 540–544.
- Takane, S.-Y. and Mitchell, J.B.O. (2004) A structure–odour relationship study using EVA descriptors and hierarchical clustering. *Org. Biomol. Chem.*, **2**, 3250–3255.
- Takaoka, Y., Endo, Y., Yamanobe, S., Kakinuma, H., Okubo, T., Shimazaki, Y., Ota, T., Sumiya, S. and Yoshikawa, K. (2003) Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Comput. Sci.*, **43**, 1269–1275.
- Takeuchi, K., Kuroda, C. and Ishida, M. (1990) Prolog-based functional group perception and calculation of 1-octanol/water partition coefficients using Rekker's fragment method. *J. Chem. Inf. Comput. Sci.*, **30**, 22–26.
- Takihi, N., Rosenkranz, H.S., Klopman, G. and Mattison, D.R. (1994) Structural determinants of developmental toxicity. *Risk Anal.*, **14**, 649–657.
- Tame, J.R.H. (2005) Scoring functions – the first 100 years. *J. Comput. Aid. Mol. Des.*, **19**, 445–451.
- Tämm, K., Fara, D.C., Katritzky, A.R., Burk, P. and Karelson, M. (2004) A quantitative structure–property relationship study of lithium cation basicities. *J. Phys. Chem. A*, **108**, 4812–4818.
- Tan, N., Li, J., Li, Z. and Li, X. (2006) Prediction of antitumor activity for epothilone analogues based on 3D molecular descriptors. *Acta Phys-Chim Sin.*, **22**, 397–402.
- Tan, Y. and Siebert, K.J. (2004) Quantitative structure–activity relationship modeling of alcohol, ester, aldehyde, and ketone flavor thresholds in beer from molecular features. *J. Agr. Food Chem.*, **52**, 3057–3064.
- Tanaka, A. and Fujiwara, H. (1996) Quantitative structure–activity relationship study of fibrinogen inhibitors ((4-(4-amidinophenoxy)butanoyl) aspartyl)valine (FK633) derivatives, using a novel hydrophobic descriptor. *J. Med. Chem.*, **39**, 5017–5020.
- Tanaka, A., Nakamura, K., Nakanishi, I. and Fujiwara, H. (1994) A novel and useful descriptor for hydrophobicity, partition coefficient micellar-water, and its application to a QSAR study of antiplatelet agents. *J. Med. Chem.*, **37**, 4563–4566.
- Tanford, C. (1957) The location of electrostatic charges in Kirkwood's model of organic ions. *J. Am. Chem. Soc.*, **79**, 5348–5352.

- Tanford, C. (1961) *Physical Chemistry of Macromolecules*, John Wiley & Sons, Inc., New York.
- Tanford, C. (1973) *The Hydrophobic Effect*, John Wiley & Sons, Inc., New York.
- Tang, K. and Li, T. (2002) Combining PLS with GA-GP for QSAR. *Chemom. Intell. Lab. Syst.*, **64**, 55–64.
- Tang, L.-J., Zhou, Y.-P., Jiang, J.-H., Zou, H.-Y., Wu, H.-L., Shen, G.-L. and Yu, R.-Q. (2007) Radial basis function network-based transform for a nonlinear support vector machine as optimized by a particle swarm optimization algorithm with application to QSAR studies. *J. Chem. Inf. Model.*, **47**, 1438–1445.
- Tang, Y., Jiang, H., Chen, K. and Ji, R. Y. (1996) QSAR study of artemisinin (Qinghaosu) derivatives using neural network method. *Indian J. Chem.*, **35B**, 325–332.
- Tao, P., Wang, R. and Lai, L. (1999) Calculating partition coefficients of peptides by the addition method. *J. Mol. Model.*, **5**, 189–195.
- Tao, S. and Lu, X. (1999) Estimation of organic carbon normalized sorption coefficient (K_{OC}) for soils by topological indices and polarity factors. *Chemosphere*, **39**, 2019–2034.
- Tao, S., Piao, H., Dawson, R., Lu, X. and Hu, H. (1999) Estimation of organic carbon normalized sorption coefficient (K_{OC}) for soils using the fragment constant method. *Environ. Sci. Technol.*, **33**, 2719–2725.
- Tarasov, V.A., Mustafaev, O.N., Abilev, S.K. and Mel'nik, V.A. (2005) Use of compound structural descriptors for increasing the efficiency of QSAR study. *Russ. J. Gen.*, **41**, 814–821.
- Taraviras, S.L., Ivanciuc, O. and Cabrol-Bass, D. (2000) Identification of groupings of graph theoretical molecular descriptors using a hybrid cluster analysis approach. *J. Chem. Inf. Comput. Sci.*, **40**, 1128–1146.
- Tarkhov, A. (2003) Chemistry on Internet, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 794–843.
- Tarko, L. and Ivanciuc, O. (2001) QSAR modeling of the anticonvulsant activity of phylacetanilides with PRECLAV (property evaluation by class variables). *MATCH Commun. Math. Comput. Chem.*, **44**, 201–214.
- Taskinen, J. and Yliruusi, J. (2003) Prediction of physico-chemical properties based on neural network modelling. *Adv. Drug Deliv. Rev.*, **55**, 1163–1183.
- Taylor, P.J. (1990) Hydrophobic properties of drugs, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 241–294.
- te Heesen, H., Schlitter, A.M. and Schlitter, J. (2007) Empirical rules facilitate the search for binding sites on protein surfaces. *J. Mol. Graph. Model.*, **25**, 671–679.
- Teague, S.J., Davis, A.M., Leeson, P.D. and Oprea, T.I. (1999) The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.*, **38**, 3743–3748.
- Tehan, B.G., Lloyd, E.J., Wong, M.G., Pitt, W.R., Gancia, E. and Manallack, D.T. (2002a) Estimation of pK_a using semiempirical molecular orbital methods. Part 2. Application to amines, anilines and various nitrogen containing heterocyclic compounds. *Quant. Struct. -Act. Relat.*, **21**, 473–485.
- Tehan, B.G., Lloyd, E.J., Wong, M.G., Pitt, W.R., Montana, J.G., Manallack, D.T. and Gancia, E. (2002b) Estimation of pK_a using semiempirical molecular orbital methods. Part 1. Application to phenols and carboxylic acids. *Quant. Struct. -Act. Relat.*, **21**, 457–472.
- ter Laak, A.M., Tsai, R.-S., Donné-Op den Kelder, G. M., Carrupt, P.-A., Testa, B. and Timmerman, H. (1994) Lipophilicity and hydrogen-bonding capacity of H_1 -antihistaminic agents in relation to their central sedative side effects. *Eur. J. Pharm. Sci.*, **2**, 373–384.
- Terletskaia, A., Shvets, N., Dimoglo, A. and Chumakov, Y. (1999) Computer-aided investigation of the structure–activity relationships of benzodiazepine derivatives at diazepam-sensitive receptors. *J. Mol. Struct. (Theochem)*, **463**, 99–103.
- Testa, B. and Bojarski, A.J. (2000) Molecules as complex adaptative systems: constrained molecular properties and their biochemical significance. *Eur. J. Pharm. Sci.*, **11**, S3–S14.
- Testa, B., Carrupt, P.-A., Gaillard, P., Billois, F. and Weber, P. (1996) Lipophilicity in molecular modeling. *Pharm. Res.*, **13**, 335–343.
- Testa, B., Crivori, P., Reist, M. and Carrupt, P.-A. (2000) The influence of lipophilicity on the pharmacokinetic behavior of drugs: concepts and examples. *Persp. Drug Disc. Des.*, **19**, 179–211.
- Testa, B., El Tayar, N., Altomare, C., Carrupt, P.-A., Tsai, R.-S. and Carotti, A. (1993) *The Hydrogen Bonding of Drugs: Its Experimental Determination and Role in Pharmacokinetics and Pharmacodynamics*, in *Trends in Chemical Research* (eds P. Angeli, U. Gulini and W. Quaglia), Elsevier, Amsterdam, The Netherlands, pp. 61–72.
- Testa, B. and Kier, L.B. (1991) The concept of molecular structure in structure–activity relationship studies and drug design. *Med. Res. Rev.*, **11**, 35–48.

- Testa, B., Kier, L.B. and Carrupt, P.-A. (1997) A systems approach to molecular structure, intermolecular recognition, and emergence-dissolve in medicinal research. *Med. Res. Rev.*, **17**, 303–326.
- Testa, B. and Purcell, W.P. (1978) A QSAR study of sulfonamide binding to carbonic anhydrase as test of steric models. *Eur. J. Med. Chem.*, **13**, 509–514.
- Testa, B., Raynaud, I. and Kier, L.B. (1999) What differentiates free amino acids and aminoacyl residues? An exploration of conformational and lipophilicity spaces. *Helv. Chim. Acta*, **82**, 657–665.
- Testa, B. and Seiler, P. (1981) Steric and lipophobic components of the hydrophobic fragmental constant. *Arzneim. Forsch. (German)*, **31**, 1053–1058.
- Tetko, I.V. (1998) Application of a pruning algorithm to optimize artificial neural networks for pharmaceutical fingerprinting. *J. Chem. Inf. Comput. Sci.*, **38**, 660–668.
- Tetko, I.V. (2003) The WWW as a tool to obtain molecular parameters. *Mini Rev. Med. Chem.*, **3**, 809–820.
- Tetko, I.V., Bruneau, P., Mewes, H.-W., Rohrer, D.C. and Poda, G.I. (2006) Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov. Today*, **11**, 700–707.
- Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D.J., Ertl, P., Palyulin, V.A., Radchenko, E.V., Zefirov, A.N., Makarenko, A.S., Tanchuk, V.Y. and Prokopenko, V.V. (2005) Virtual computation chemistry laboratory – design and description. *J. Comput. Aid. Mol. Des.*, **19**, 453–463.
- Tetko, I.V. and Livingstone, D.J. (2007) Rule-based systems to predict lipophilicity, in *ADME-Tox Approaches*, Vol. 5 (eds B. Testa and H. van de Waterbeemd), Elsevier, Amsterdam, The Netherlands, pp. 649–668.
- Tetko, I.V., Livingstone, D.J. and Luik, A.I. (1995) Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.*, **35**, 826–833.
- Tetko, I.V. and Poda, G.I. (2004) Application of ALOGPS 2.1 to predict log *D* distribution coefficient for Pfizer proprietary compounds. *J. Med. Chem.*, **47**, 5601–5604.
- Tetko, I.V. and Poda, G.I. (2008) Prediction of log *P* with property-based methods, in *Molecular Drug Properties*, Vol. 37 (ed. R. Mannhold), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 381–406.
- Tetko, I.V. and Tanchuk, V.Y. (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.*, **42**, 1136–1145.
- Tetko, I.V., Tanchuk, V.Y., Kasheva, T.N. and Villa, A. E.P. (2001a) Estimation of aqueous solubility of chemical compounds using *E*-state indices. *J. Chem. Inf. Comput. Sci.*, **41**, 1488–1493.
- Tetko, I.V., Tanchuk, V.Y., Kasheva, T.N. and Villa, A. E.P. (2001b) Internet software for the calculation of the lipophilicity and aqueous solubility of chemical compounds. *J. Chem. Inf. Comput. Sci.*, **41**, 246–252.
- Tetko, I.V., Tanchuk, V.Y. and Villa, A.E.P. (2001c) Prediction of *n*-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and *E*-state indices. *J. Chem. Inf. Comput. Sci.*, **41**, 1407–1421.
- Tetko, I.V., Villa, A.E.P. and Livingstone, D.J. (1996) Neural network studies. 2. Variable selection. *J. Chem. Inf. Comput. Sci.*, **26**, 794–803.
- Tetteh, J., Howells, S.L., Metcalfe, E. and Suzuki, T. (1998) Optimisation of radial basis function neural networks using biharmonic spline interpolation. *Chemom. Intell. Lab. Syst.*, **41**, 17–29.
- Tetteh, J., Metcalfe, E. and Howells, S.L. (1996) Optimization of radial basis and backpropagation neural networks for modelling auto-ignition temperature by quantitative structure–property relationships. *Chemom. Intell. Lab. Syst.*, **32**, 177–191.
- Tetteh, J., Suzuki, T., Metcalfe, E. and Howells, S. (1999) Quantitative structure–property relationships for the estimation of boiling point and flash point using a radial basis function neural network. *J. Chem. Inf. Comput. Sci.*, **39**, 491–507.
- Thakur, A., Thakur, M., Kakani, N., Joshi, A., Thakur, S. and Gupta, A. (2004a) Application of topological and physico-chemical descriptors: QSAR study of phenylamino-acridine descriptors derivatives. *ARKIVOC*, (xiv), 36–43.
- Thakur, M., Thakur, A. and Khadikar, P.V. (2004b) QSAR studies on psychotomimetic phenylalkylamines. *Bioorg. Med. Chem.*, **12**, 825–831.
- Thangavel, P. and Venuvanalingam, P. (1993) Algorithms for the computation of molecular distance matrix and distance polynomial of chemical graphs on parallel computers. *J. Chem. Inf. Comput. Sci.*, **33**, 412–414.
- Thibaut, U. (1993) Applications of CoMFA and related 3D QSAR approaches, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 661–696.
- Thibaut, U., Folkers, G., Klebe, G., Kubinyi, H., Merz, A. and Rognan, D. (1994) Recommendations for CoMFA studies and 3D-QSAR publications. *Quant. Struct. -Act. Relat.*, **13**, 1–3.

- Thinh, T.P. and Trong, T.K. (1976) Estimation of standard heats of formation, ΔH_f° , standard entropies of formation, ΔS_f° , standard free energies of formation, ΔG_f° , and absolute entropies, ΔS° , of hydrocarbons from group contributions: an accurate approach. *Can. J. Chem. Eng.*, **54**, 344–357.
- Thohalaki, S. and Pachter, R. (2005) Prediction of melting points for ionic liquids. *QSAR Comb. Sci.*, **24**, 485–490.
- Thomas, E.R. and Eckert, C.A. (1984) Prediction of limiting activity coefficients by a modified separation of cohesive energy density model and UNIFAC. *I & EC. Process Des. Dev.*, **23**, 194–209.
- Thomas, J., Berkoff, C.E., Flagg, W.B., Gallo, J.J., Haff, R.F. and Pinto, C.A. (1975) Antiviral quinolinehydrazones. A modified Free–Wilson analysis. *J. Med. Chem.*, **18**, 245–250.
- Thomsen, M., Dobel, S., Lassen, P., Carlsen, L., Mogensen, B.B. and Hansen, P.E. (2002) Reverse quantitative structure–activity relationship for modelling the sorption of esfenvalerate to dissolved organic matter. A multivariate approach. *Chemosphere*, **49**, 1317–1325.
- Thomsen, M., Rasmussen, A.G. and Carlsen, L. (1999) SAR/QSAR approaches to solubility, partitioning and sorption of phthalates. *Chemosphere*, **38**, 2613–2624.
- Thorner, D.A., Willett, P., Wright, P.M. and Taylor, R. (1997) Similarity searching in files of three-dimensional chemical structures: representation and searching of molecular electrostatic potentials using field graphs. *J. Comput. Aid. Mol. Des.*, **11**, 163–174.
- Thull, U., Kneubuhler, S., Gaillard, P., Carrupt, P.-A., Testa, B., Altomare, C., Carotti, A., Jenner, P. and McNaught, K.S. (1995) Inhibition of monoamine oxidase by isoquinoline derivatives qualitative and 3D quantitative structure–activity relationships. *Biochem. Pharmacol.*, **50**, 869–877.
- Tian, F., Zhou, P. and Li, Z. (2007) T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *J. Mol. Struct.*, **830**, 106–115.
- Timerbaev, A.R., Semenova, O.P. and Petrukhin, O. M. (2002) Migration behavior of metal complexes in capillary zone electrophoresis. Interpretation in terms of quantitative structure–mobility relationships. *J. Chromat.*, **943**, 263–274.
- Timofei, S. and Fabian, W.M.F. (1998) Comparative molecular field analysis of heterocyclic monoazo dye–fiber affinities. *J. Chem. Inf. Comput. Sci.*, **38**, 1218–1222.
- Timofei, S., Kurunczi, L., Schmidt, W. and Simon, Z. (1995) Structure–affinity binding relationships of some 4-aminoazobenzene derivatives for cellulose fibre. *Dyes & Pigments*, **29**, 251–258.
- Timofei, S., Kurunczi, L., Schmidt, W. and Simon, Z. (1996) Lipophilicity in dye cellulose fiber binding. *Dyes & Pigments*, **32**, 25–42.
- Timofei, S., Kurunczi, L. and Simon, Z. (2001) Structure–affinity relationships by the MTD method for binding to cellulose fibre of some heterocyclic monoazo dyes. *MATCH Commun. Math. Comput. Chem.*, **44**, 349–360.
- Timofei, S., Schmidt, W., Kurunczi, L. and Simon, Z. (2000) A review of QSAR for dye affinity for cellulose fibres. *Dyes & Pigments*, **47**, 5–16.
- Tinker, J. (1981) Relating mutagenicity to chemical structure. *J. Chem. Inf. Comput. Sci.*, **21**, 3–7.
- Tiwari, V. and Pande, R. (2006) Molecular descriptors of N-arylhydroxamic acids: a tool in drug design. *Chem. Biol. Drug Des.*, **68**, 225–228.
- Todeschini, R. (1997) Data correlation, number of significant principal components and shape of molecules. The K correlation index. *Anal. Chim. Acta*, **348**, 419–430.
- Todeschini, R. (2004) Reality and models. Concepts, strategies and tools for QSAR, in *Designing Drugs and Crop Protectants: Processes, Problems and Solutions* (eds M. Ford, D.J. Livingstone, J.C. Dearden and H. van de Waterbeemd), Blackwell, Oxford, UK, pp. 235–242.
- Todeschini, R. (2006) Molecular descriptors and chemometrics. *G. I. T. Laboratory Journal*, **5**, 40–42.
- Todeschini, R., Ballabio, D., Consonni, V. and Mauri, A. (2007) A new similarity/diversity measure for sequential data. *MATCH Commun. Math. Comput. Chem.*, **57**, 51–67.
- Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. and Pavan, M. (2005) CAIMAN (classification and influence matrix analysis): a new classification method based on leverage-scaled functions. *Chemom. Intell. Lab. Syst.*, **87**, 3–17.
- Todeschini, R., Bettiol, C., Giurin, G., Gramatica, P., Miana, P. and Argese, E. (1996) Modeling and prediction by using WHIM descriptors in QSAR studies. Submitochondrial particles (SMP) as toxicity biosensors of chlorophenols. *Chemosphere*, **33**, 71–79.
- Todeschini, R., Cazar, R. and Collina, E. (1992) The chemical meaning of topological indices. *Chemom. Intell. Lab. Syst.*, **15**, 51–59.
- Todeschini, R. and Consonni, V. (2000) *Handbook of Molecular Descriptors*, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 668.
- Todeschini, R. and Consonni, V. (2003) Descriptors from molecular geometry, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger),

- Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1004–1033.
- Todeschini, R., Consonni, V., Galvagni, D. and Gramatica, P. (1999) A new molecular structure representation: spectral weighted molecular (SWM) signals for studies of molecular similarity. *Quim. Anal.*, **18**, 41–47.
- Todeschini, R., Consonni, V. and Gramatica, P. (2009) Chemometrics in QSAR, in *Comprehensive Chemometrics*, vol. 4 (eds S. Brown, B. Walczak, and R. Tauler), Elsevier, Oxford, UK, 129–172.
- Todeschini, R., Consonni, V. and Maiocchi, A. (1998) The K correlation index: theory development and its applications in chemometrics. *Chemom. Intell. Lab. Syst.*, **46**, 13–29.
- Todeschini, R., Consonni, V., Mauri, A. and Ballabio, D. (2006) Characterization of DNA primary sequences by a new similarity/diversity measure based on the partial ordering. *J. Chem. Inf. Model.*, **46**, 1905–1911.
- Todeschini, R., Consonni, V., Mauri, A. and Pavan, M. (2003) MOBYDIGS: software for regression and classification models by genetic algorithms, in *Chemometrics: Genetic Algorithms and Artificial Neural Networks* (ed. R. Leardi), Elsevier, Amsterdam, The Netherlands, pp. 141–167.
- Todeschini, R., Consonni, V., Mauri, A. and Pavan, M. (2004a) Detecting “bad” regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta*, **515**, 199–208.
- Todeschini, R., Consonni, V., Mauri, A. and Pavan, M. (2004b) New fitness functions to avoid bad regression models in variable subset selection by genetic algorithms, in *Designing Drugs and Crop Protectants: Processes, Problems and Solutions* (eds M. Ford, D.J. Livingstone, J.C. Dearden and H. van de Waterbeemd), Blackwell, Oxford, UK, pp. 323–325.
- Todeschini, R., Consonni, V. and Pavan, M. (2004c) A distance measure between models: a tool for similarity/diversity analysis of model populations. *Chemom. Intell. Lab. Syst.*, **70**, 55–61.
- Todeschini, R., Consonni, V. and Pavan, M. (2004d) Distance measure between models: a tool for model similarity/diversity analysis, in *Designing Drugs and Crop Protectants: Processes, Problems and Solutions* (eds M. Ford, D.J. Livingstone, J.C. Dearden and H. van de Waterbeemd), Blackwell, Oxford, UK, pp. 467–469.
- Todeschini, R. and Gramatica, P. (1997a) 3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. *Quant. Struct. -Act. Relat.*, **16**, 113–119.
- Todeschini, R. and Gramatica, P. (1997b) 3D-modelling and prediction by WHIM descriptors. Part 6. Application of whim descriptors in QSAR studies. *Quant. Struct. -Act. Relat.*, **16**, 120–125.
- Todeschini, R. and Gramatica, P. (1997c) The WHIM theory: new 3D molecular descriptors for QSAR in environmental modelling. *SAR & QSAR Environ. Res.*, **7**, 89–115.
- Todeschini, R. and Gramatica, P. (1998) New 3D molecular descriptors: the WHIM theory and QSAR applications, in *3D QSAR in Drug Design*, Vol. 2 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 355–380.
- Todeschini, R., Gramatica, P., Marengo, E. and Provenzani, R. (1995) Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physico-chemical properties of polyaromatic hydrocarbons (PAH). *Chemom. Intell. Lab. Syst.*, **27**, 221–229.
- Todeschini, R., Lasagni, M. and Marengo, E. (1994) New molecular descriptors for 2D- and 3D-structures. Theory. *J. Chemom.*, **8**, 263–273.
- Todeschini, R., Moro, G., Boggia, R., Bonati, L., Cosentino, U., Lasagni, M. and Pitea, D. (1997) Modeling and prediction of molecular properties. Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptors. *Chemom. Intell. Lab. Syst.*, **36**, 65–73.
- Todeschini, R., Vighi, M., Finizio, A. and Gramatica, P. (1997) 3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors. *SAR & QSAR Environ. Res.*, **7**, 173–193.
- Todeschini, R., Vighi, M., Provenzani, R., Finizio, A. and Gramatica, P. (1996) Modeling and prediction by using WHIM descriptors in QSAR studies: toxicity of heterogeneous chemicals on *Daphnia magna*. *Chemosphere*, **32**, 1527–1545.
- Tokarski, J. and Hopfinger, A.J. (1994) Three-dimensional molecular shape analysis – quantitative structure–activity relationship of a series of cholecystokinin-A receptor antagonists. *J. Med. Chem.*, **37**, 3639–3654.
- Tomczak, J. (2003) Data types, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 392–409.
- Tomić, S., Nilsson, L. and Wade, R.C. (2000) Nuclear receptor–DNA binding specificity: a COMBINE and Free–Wilson QSAR analysis. *J. Med. Chem.*, **43**, 1780–1792.
- Tominaga, Y. (1998a) Data structure comparison using box counting analysis. *J. Chem. Inf. Comput. Sci.*, **38**, 867–875.

- Tominaga, Y. (1998b) Novel 3D descriptors using excluded volume. 2. Application to drug classification. *J. Chem. Inf. Comput. Sci.*, **38**, 1157–1160.
- Tominaga, Y. (1999) Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. *Chemom. Intell. Lab. Syst.*, **49**, 105–115.
- Tominaga, Y. and Fujiwara, I. (1997a) Data structure comparison using fractal analysis. *Chemom. Intell. Lab. Syst.*, **39**, 187–193.
- Tominaga, Y. and Fujiwara, I. (1997b) Novel 3D descriptors using excluded volume: application to 3D quantitative structure–activity relationships. *J. Chem. Inf. Comput. Sci.*, **37**, 1158–1161.
- Tomović, Ž. and Gutman, I. (2001a) Modeling boiling points of cycloalkanes by means of iterated line graph sequences. *J. Chem. Inf. Comput. Sci.*, **41**, 1041–1045.
- Tomović, Ž. and Gutman, I. (2001b) Narumi–Katayama index of phenylenes. *J. Serb. Chem. Soc.*, **66**, 243–247.
- Tömpe, P., Clementis, G., Petnehazy, I., Jaszay, Z.M. and Toke, L. (1995) Quantitative structure electrochemistry relationships of alpha, beta unsaturated ketones. *Anal. Chim. Acta*, **305**, 295–303.
- Tong, J., Liu, S., Zhou, P., Bulan, W. and Li, Z. (2008) A novel descriptor of amino acids and its application in peptide QSAR. *J. Theor. Biol.*, **253**, 90–97.
- Tong, W., Collantes, E.R., Chen, Y. and Welsh, W.J. (1996) A comparative molecular field analysis study of *N*-benzylpiperidines as acetylcholinesterase inhibitors. *J. Med. Chem.*, **39**, 380–387.
- Tong, W., Lowis, D.R., Perkins, R., Chen, Y., Welsh, W.J., Goddette, D.W., Heritage, T.W. and Sheehan, D.M. (1998) Evaluation of quantitative structure–activity relationship method for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.*, **38**, 669–677.
- Tonmunphean, S., Kokpol, S., Parasuk, V., Wolschann, P., Winger, R.H., Liedl, K.R. and Rode, B.M. (1998) Comparative molecular field analysis of artemisinin derivatives: *ab initio* versus semiempirical optimized structures. *J. Comput. Aid. Mol. Des.*, **12**, 397–409.
- Topliss, J.G. (ed.) (1983) *Quantitative Structure–Activity Relationships of Drugs*, Academic Press, New York.
- Topliss, J.G. (1993) Some observation on classical QSAR. *Persp. Drug Disc. Des.*, **1**, 253–268.
- Topliss, J.G. and Costello, R.J. (1972) Chance correlations in structure–activity studies using multiple regression analysis. *J. Med. Chem.*, **15**, 1066–1068.
- Topliss, J.G. and Edwards, R.P. (1979) Chance factors in studies of quantitative structure–activity relationships. *J. Med. Chem.*, **22**, 1238–1244.
- Topliss, J.G. and Shapiro, E.L. (1975) Quantitative structure–activity relationships in the Δ^6 -6-substituted progesterone series. A reappraisal. *J. Med. Chem.*, **18**, 621–623.
- Topsom, R.D. (1976) The nature and analysis of substituent electronic effects. *Prog. Phys. Org. Chem.*, **12**, 1–20.
- Topsom, R.D. (1987a) Electronic substituent effects in molecular spectroscopy. *Prog. Phys. Org. Chem.*, **16**, 193–235.
- Topsom, R.D. (1987b) Some theoretical studies of electronic substituent effects in organic chemistry. *Prog. Phys. Org. Chem.*, **16**, 125–191.
- Topsom, R.D. (1987c) Substituent effects on ground-state molecular structures and charge distributions. *Prog. Phys. Org. Chem.*, **16**, 85–124.
- Toropov, A.A. and Benfenati, E. (2004a) QSAR modelling of aldehyde toxicity against a protozoan, *Tetrahymena pyriformis*, by optimization of correlation weights of nearest neighboring codes. *J. Mol. Struct. (Theochem)*, **679**, 225–228.
- Toropov, A.A. and Benfenati, E. (2004b) QSAR modelling of aldehyde toxicity by means of optimisation of correlation weights of nearest neighbouring codes. *J. Mol. Struct. (Theochem)*, **676**, 165–169.
- Toropov, A.A., Duchowicz, P.R. and Castro, E.A. (2003) Structure–toxicity relationships for aliphatic compounds based on correlation weighting of local graph invariants. *Int. J. Mol. Sci.*, **4**, 272–283.
- Toropov, A.A., Gutman, I. and Furtula, B. (2005) Graph of atomic orbitals and molecular structure descriptors based on it. *J. Serb. Chem. Soc.*, **70**, 669–674.
- Toropov, A.A., Nesterov, I.V. and Nabiev, O.M. (2003a) QSAR modeling of dihydrofolate reductase inhibitory activity by correlation weighting of nearest neighboring codes. *J. Mol. Struct. (Theochem)*, **622**, 269–273.
- Toropov, A.A., Nesterov, I.V. and Nabiev, O.M. (2003b) QSPR modeling of cycloalkanes properties by correlation weighting of extended graph valence shells. *J. Mol. Struct. (Theochem)*, **637**, 37–42.
- Toropov, A.A., Rasulev, B.F. and Leszczynski, J. (2007) QSAR modeling of acute toxicity for nitrobenzene derivatives towards rats: comparative analysis by MLRA and optimal descriptors. *QSAR Comb. Sci.*, **26**, 686–693.

- Toropov, A.A. and Roy, K. (2004) QSPR modeling of lipid–water partition coefficient by optimization of correlation weights of local graph invariants. *J. Chem. Inf. Comput. Sci.*, **44**, 179–186.
- Toropov, A.A. and Schultz, T.W. (2003) Prediction of aquatic toxicity: use of optimization of correlation weights of local graph invariants. *J. Chem. Inf. Comput. Sci.*, **43**, 560–567.
- Toropov, A.A. and Toropova, A.P. (2000a) QSPR modeling of the formation constants for complexes using atomic orbitals graphs. *Russ. J. Inorg. Chem.*, **26**, 398–405.
- Toropov, A.A. and Toropova, A.P. (2000b) QSPR modeling of the stability constants of biometal complexes with phosphate derivatives of adenosine. *Russ. J. Coord. Chem.*, **26**, 792–797.
- Toropov, A.A. and Toropova, A.P. (2001a) Modeling of lipophilicity by means of correlation weighting of local graph invariants. *J. Mol. Struct. (Theochem)*, **538**, 197–199.
- Toropov, A.A. and Toropova, A.P. (2001b) Prediction of heteroaromatic amine mutagenicity by means of correlation weighting of atomic orbital graphs of local invariants. *J. Mol. Struct. (Theochem)*, **538**, 287–293.
- Toropov, A.A. and Toropova, A.P. (2001c) QSPR modeling of stability of complexes of adenosine phosphate derivatives with metals absent from the complexes of the teaching access. *Russ. J. Coord. Chem.*, **27**, 574–578.
- Toropov, A.A. and Toropova, A.P. (2002a) Modeling of acyclic carbonyl compounds normal boiling points by correlation weighting of nearest neighboring codes. *J. Mol. Struct. (Theochem)*, **581**, 11–15.
- Toropov, A.A. and Toropova, A.P. (2002b) QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. *J. Mol. Struct. (Theochem)*, **578**, 129–134.
- Toropov, A.A. and Toropova, A.P. (2002c) QSPR modeling of complex stability by optimization of correlation weights of the hydrogen bond index and the local graph invariants. *Russ. J. Coord. Chem.*, **28**, 877–880.
- Toropov, A.A. and Toropova, A.P. (2003) QSPR modeling of alkanes properties based on graph of atomic orbitals. *J. Mol. Struct. (Theochem)*, **637**, 1–10.
- Toropov, A.A. and Toropova, A.P. (2004) Nearest neighboring code and hydrogen bond index in labeled hydrogen-filled graph and graph of atomic orbitals: application to model of normal boiling points of haloalkanes. *J. Mol. Struct. (Theochem)*, **711**, 173–183.
- Toropov, A.A., Toropova, A.P. and Gutman, I. (2005) Comparison of QSPR models based on hydrogen-filled graphs and on graphs of atomic orbitals. *Croat. Chem. Acta*, **78**, 503–509.
- Toropov, A.A., Toropova, A.P., Ismailov, T. and Bonchev, D. (1998) 3D weighting of molecular descriptors for QSAR/QSPR by the method of ideal symmetry (MIS). 1. Application to boiling points of alkanes. *J. Mol. Struct. (Theochem)*, **424**, 237–247.
- Toropov, A.A., Toropova, A.P., Muftahov, R.A., Ismailov, T. and Muftahov, A.G. (1994) Simulation of molecular systems by the ideal symmetry method for revealing quantitative structure–property relations. *Russ. J. Phys. Chem.*, **68**, 577–579.
- Toropov, A.A., Toropova, A.P., Nesterov, I.V. and Nabiev, O.M. (2003) Comparison of QSAR models of anti-HIV-1 potencies based on labeled hydrogen filled graph and graph of atomic orbitals. *J. Mol. Struct. (Theochem)*, **640**, 175–181.
- Toropov, A.A., Toropova, A.P., Netserova, A.I. and Nabiev, O.M. (2004) Prediction of alkane enthalpies by means of correlation weighting of Morgan extended connectivity in molecular graphs. *Chem. Phys. Lett.*, **384**, 357–363.
- Toropova, A.P. and Toropov, A.A. (2000) QSPR modeling of stability constants of coordination compounds by optimization of correlation weights of local graph invariants. *Russ. J. Coord. Chem.*, **45**, 1057–1059.
- Toropova, A.P. and Toropov, A.A. (2001) Using of optimization of local graph invariants correlation weights for QSPR simulation of crystal lattice energies. *Russ. J. Struct. Chem.*, **42**, 1230–1232.
- Torrens, F. (2000) Universal organic solvent–water partition coefficient model. *J. Chem. Inf. Comput. Sci.*, **40**, 236–240.
- Torrens, F. (2001) A new topological index to elucidate apolar hydrocarbons. *J. Comput. Aid. Mol. Des.*, **15**, 709–719.
- Torrens, F. (2002) Fractal hybrid orbitals analysis of the tertiary structure of protein molecules. *Molecules*, **7**, 26–37.
- Torrens, F. (2003a) A new chemical index inspired by biological plastic evolution. *Indian J. Chem.*, **42**, 1258–1263.
- Torrens, F. (2003b) Valence topological charge-transfer indices for dipole moments. *Molecules*, **8**, 169–185.
- Torrens, F. (2004) Valence topological charge-transfer indices for dipole moments: percutaneous enhancers. *Molecules*, **9**, 1222–1235.
- Torrens, F. (2005) Valence topological charge-transfer indices for reflecting polarity: correction for heteromolecules. *Molecules*, **10**, 334–345.
- Tosato, M.L., Chiorboli, C., Eriksson, L. and Jonsson, J. (1991) Multivariate modelling of the rate

- constant of the gas-phase reaction of haloalkanes with the hydroxyl radical. *Sci. Total Environ.*, **190/110**, 307–325.
- Tosato, M.L., Piazza, R., Chiorboli, C., Passerini, L., Pino, A., Cruciani, G. and Clementi, S. (1992) Application of chemometrics to the screening of hazardous chemicals. *Chemom. Intell. Lab. Syst.*, **16**, 155–167.
- Toungue, B.A., Pfahler, L.B. and Reynolds, C.H. (2002) Chemical information based scaling of molecular descriptors: a universal chemical scale for library design and analysis. *J. Chem. Inf. Comput. Sci.*, **42**, 879–884.
- Tovar, A., Eckert, H. and Bajorath, J. (2008) Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem*, **2**, 208–217.
- Trapani, G., Carotti, A., Franco, M., Latrofa, A., Genchi, G. and Liso, G. (1993) Structure–affinity relationships of some alkoxycarbonyl 2H-pyrimido or alkoxycarbonyl-4H-pyrimido-[2,1-b] benzothiazol-2-one or 4-one benzodiazepine receptor ligands. *Eur. J. Med. Chem.*, **28**, 13–21.
- Tratch, S.S., Devdariani, R.O. and Zefirov, N.S. (1990) Combinatorial models and algorithms in chemistry. Topological–configurational analogs of the Wiener index. *Zhur. Org. Khim. (Russian)*, **26**, 921–932.
- Tratch, S.S., Lomova, O.A., Sukhachev, D.V., Palyulin, V.A. and Zefirov, N.S. (1992) Generation of molecular graphs for QSAR studies: an approach based on acyclic fragment combinations. *J. Chem. Inf. Comput. Sci.*, **32**, 130–139.
- Tratch, S.S., Stankevitch, I.V. and Zefirov, N.S. (1990) Combinatorial models and algorithms in chemistry. The expanded Wiener number – a novel topological index. *J. Comput. Chem.*, **11**, 899–908.
- Traube, I. (1904) Theorie der Osmose und Narkose. *Arch. für die ges. Physiol.*, **105**, 541–558.
- Trepalin, S.V., Gerasimenko, V.A., Kozyukov, A.V., Savchuk, N.P. and Ivashchenko, A.A. (2002) New diversity calculations algorithms used for compound selection. *J. Chem. Inf. Comput. Sci.*, **42**, 249–258.
- Trinajstić, N. (1988) The characteristic polynomial of a chemical graph. *J. Math. Chem.*, **2**, 197–215.
- Trinajstić, N. (1991) Graph theory and molecular orbitals, in *Chemical Graph Theory. Introduction and Fundamentals* (eds D. Bonchev and D.H. Rouvray), Abacus Press/Gordon and Breach Science Publishers, New York, pp. 235–279.
- Trinajstić, N. (1992) *Chemical Graph Theory*. CRC Press, Boca Raton, FL, p. 322.
- Trinajstić, N., Babic, D., Nikolić, S., Plavšić, D., Amić, D. and Mihalić, Z. (1994) The Laplacian matrix in chemistry. *J. Chem. Inf. Comput. Sci.*, **34**, 368–376.
- Trinajstić, N. and Gutman, I. (2002) Mathematical chemistry. *Croat. Chem. Acta*, **75**, 329–356.
- Trinajstić, N., Jericevic, Z., von Knop, J., Müller, W.R. and Szymanski, K. (1983) Computer generation of isomeric structures. *Prot. Struct. Funct. Gen.*, **55**, 379–390.
- Trinajstić, N., Klein, D.J. and Randić, M. (1986) On some solved and unsolved problems of chemical graph theory. *Int. J. Quantum Chem. Quant. Chem. Symp.*, **20**, 699–742.
- Trinajstić, N., Nikolić, S., Basak, S.C. and Lukovits, I. (2001) Distance indices and their hyper-counterparts: intercorrelation and use in the structure–property modeling. *SAR & QSAR Environ. Res.*, **12**, 31–54.
- Trinajstić, N., Nikolić, S., Lučić, B. and Amić, D. (1996) On QSAR modeling. *Acta Pharm. Jugosl.*, **46**, 249–263.
- Trinajstić, N., Nikolić, S., Lučić, B., Amić, D. and Mihalić, Z. (1997) The Detour matrix in chemistry. *J. Chem. Inf. Comput. Sci.*, **37**, 631–638.
- Trinajstić, N., Randić, M. and Klein, D.J. (1986) On the quantitative structure–activity relationship in drug research. *Acta Pharm. Jugosl.*, **36**, 267–279.
- Trindle, C. (1969) Bond index description of delocalization. *J. Am. Chem. Soc.*, **91**, 219–220.
- Trohalaki, S., Gifford, E. and Pachter, R. (2000) Improved QSARs for predictive toxicology of halogenated hydrocarbons. *Computers Chem.*, **24**, 421–427.
- Trohalaki, S., Pachter, R., Drake, G. and Hawkins, T. (2005) Quantitative structure–property relationships for melting points and densities of ionic liquids. *Energy & Fuels*, **19**, 279–284.
- Trone, M.D., Leonard, M.S. and Khaleli, M.G. (2000) Congeneric behavior in estimations of octanol–water partition coefficients by micellar electrokinetic chromatography. *Anal. Chem.*, **72**, 1228–1235.
- Tropsha, A. and Cho, S.J. (1998) Cross-validated R^2 guided region selection for CoMFA studies, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 57–69.
- Tropsha, A., Gramatica, P. and Gombar, V.K. (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.*, **22**, 69–77.
- Tropsha, A. and Reynolds, C.H. (2002) Designing focused libraries for drug discovery: hit to lead to drug. *J. Mol. Graph. Model.*, **20**, 427–428.

- Tropsha, A. and Zheng, W. (2002) Rational principles of compound selection for combinatorial library design. *Comb. Chem. High T. Scr.*, **5**, 111–123.
- Trucco, E. (1956a) A note on the information content of graphs. *Bull. Math. Biophys.*, **18**, 129–135.
- Trucco, E. (1956b) On the information content of graphs: compound symbols; different states for each point. *Bull. Math. Biophys.*, **18**, 237–253.
- Tsai, R.-S., Carrupt, P.-A. and Testa, B. (1995) Measurement of partition coefficient using centrifugal partition chromatography. Method development and application to the determination of solute properties, in *Modern Countercurrent Chromatography* (eds W.D. Conway and R.J. Petroski), American Chemical Society, New York, pp. 143–154.
- Tsai, R.-S., Fan, W., El Tayar, N., Carrupt, P.-A., Testa, B. and Kier, L.B. (1993) Solute–water interactions in the organic phase of a biphasic system. 1. Structural influence of organic solutes on the “water-dragging” effect. *J. Am. Chem. Soc.*, **115**, 9632–9639.
- Tsai, R.-S., Testa, B., El Tayar, N. and Carrupt, P.-A. (1991) Structure–lipophilicity relationships of zwitterionic amino acids. *J. Chem. Soc. Perkin Trans. 2*, 1802.
- Tsakovska, I., Lessigiarska, I., Netzeva, T.I. and Worth, A.P. (2008) A mini review of mammalian toxicity (Q)SAR models. *QSAR Comb. Sci.*, **27**, 41–48.
- Tsantili-Kakoulidou, A. and Kier, L.B. (1992) A quantitative structure–activity relationship (QSAR) study of alkylpyrazine odor modalities. *Pharm. Res.*, **9**, 1321–1323.
- Tsantili-Kakoulidou, A., Kier, L.B. and Joshi, N. (1992) The use of electrotopological state indexes in QSAR studies. *J. Chim. Phys. Phys-Chim. Biol.*, **89**, 1729–1733.
- TSAR Reference Manual, Oxford Molecular Ltd, The Magdalen Centre, Oxford Science Park, Sandford-on-Thames, Oxford, UK.
- Tsygankova, I.G. (2004) Combination of fragmental and topological descriptors for QSPR estimations of boiling temperature. *QSAR Comb. Sci.*, **23**, 629–636.
- Tugcu, N., Ladiwala, A., Breneman, C.M. and Cramer, S.M. (2003) Identification of chemically selective displacers using parallel batch screening experiments and quantitative structure efficacy relationship models. *Anal. Chem.*, **75**, 5806–5816.
- Tugcu, N., Song, M., Breneman, C.M., Sukumar, N., Bennett, K.P. and Cramer, S.M. (2003) Prediction of the effect of mobile-phase salt type on protein retention and selectivity in anion exchange systems. *Anal. Chem.*, **75**, 3563–3572.
- Tunkel, J., Mayo, K., Austin, C., Hickerson, A. and Howard, P. (2005) Practical considerations on the use of predictive models for regulatory purposes. *Environ. Sci. Technol.*, **39**, 2188–2199.
- Tuppurainen, K. (1994) QSAR approach to molecular mutagenicity. A survey and a case study: MX compounds. *J. Mol. Struct. (Theochem)*, **306**, 49–56.
- Tuppurainen, K. (1999a) EEVA (electronic eigenvalue): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. *SAR & QSAR Environ. Res.*, **10**, 39–46.
- Tuppurainen, K. (1999b) Frontier orbital energies. Hydrophobicity and steric factors as physical QSAR descriptors of molecular mutagenicity. A review with a case study: MX compounds. *Chemosphere*, **38**, 3015–3030.
- Tuppurainen, K. and Lotjonen, S. (1993) On the mutagenicity of MX compounds. *Mut. Res.*, **287**, 235–241.
- Tuppurainen, K., Lotjonen, S., Laatikainen, R. and Vartiainen, T. (1992) Structural and electronic properties of MX compounds related to ta100 mutagenicity: a semiempirical molecular orbital QSAR study. *Mut. Res.*, **266**, 181–188.
- Tuppurainen, K. and Ruuskanen, J. (2000) Electronic eigenvalue (EEVA): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. A QSAR approach to the Ah receptor binding affinity of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs). *Chemosphere*, **41**, 843–848.
- Tuppurainen, K., Viisas, M., Laatikainen, R. and Peräkylä, M. (2002) Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.*, **42**, 607–613.
- Tuppurainen, K., Viisas, M., Peräkylä, M. and Laatikainen, R. (2004) Ligand intramolecular motions in ligand–protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extended steroid benchmark dataset. *J. Comput. Aid. Mol. Des.*, **18**, 175–187.
- Turabekova, M.A. and Rasulev, B.F. (2004) A QSAR toxicity study of a series of alkaloids with the lycoctonine skeleton. *Molecules*, **9**, 1194–1207.
- TURBOMOLE, TURBOMOLE GmbH, HRB702063, Amtsgericht Mannheim, Germany.
- Türker, L. (2003a) A novel topological index for coding of alternant systems. *Indian J. Chem.*, **42**, 1295–1297.
- Türker, L. (2003b) Contemplation on the Hosoya indices. *J. Mol. Struct. (Theochem)*, **623**, 75–77.

- Türker, L. (2003c) Hosoya indices and a new approach to molecular similarity. *Indian J. Chem.*, **42**, 1442–1445.
- Turner, D.B., Costello, C.L. and Jurs, P.C. (1998) Prediction of critical temperatures and pressures of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **38**, 639–645.
- Turner, D.B. and Willett, P. (2000a) Evaluation of the EVA descriptor for QSAR studies. 3. The use of a genetic algorithm to search for models with enhanced predictive properties (EVA_GA). *J. Comput. Aid. Mol. Des.*, **14**, 1–21.
- Turner, D.B. and Willett, P. (2000b) The EVA spectral descriptor. *Eur. J. Med. Chem.*, **35**, 367–375.
- Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W. (1995) Similarity searching in files of three-dimensional structures: evaluation of similarity coefficients and standardization methods for field-based similarity searching. *SAR & QSAR Environ. Res.*, **3**, 101–130.
- Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W. (1997) Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput. Aid. Mol. Des.*, **11**, 409–422.
- Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W. (1999) Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies. 2. Model validation using a benchmark steroid dataset. *J. Comput. Aid. Mol. Des.*, **13**, 271–296.
- Turowski, M., Kaliszan, R., Lüllmann, C., Genieser, H.G. and Jastorff, B. (1996) New stationary phases for high-performance liquid chromatographic separation of nucleosides and cyclic nucleotides. Synthesis and chemometric analysis of retention data. *J. Chromat.*, **728**, 201–211.
- Turro, N.J. (1986) Geometric and topological thinking in organic chemistry. *Angew. Chem. Int. Ed. Engl.*, **25**, 882–901.
- Tusnády, G.E. and Simon, I. (2001) Topology of membrane proteins. *J. Chem. Inf. Comput. Sci.*, **41**, 364–368.
- Tute, M.S. (1990) History and objectives of quantitative drug design, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 1–31.
- Tvaruzek, P. and Komenda, J. (1991) Geometrical descriptors of molecules. *Collect. Czech. Chem. Comm.*, **56**, 253–257.
- Tversky, A. (1977) Features of similarity. *Psychol. Rev.*, **84**, 327–352.
- Tysklind, M., Lundgren, K., Rappe, C., Eriksson, L., Jonsson, J. and Sjöström, M. (1993) Multivariate quantitative structure–activity relationships for polychlorinated dibenzo-*p*-dioxins and dibenzofurans. *Environ. Toxicol. Chem.*, **12**, 659–672.
- Tysklind, M., Lundgren, K., Rappe, C., Eriksson, L., Jonsson, J., Sjöström, M. and Ahlborg, U.G. (1992) Multivariate characterization and modeling of polychlorinated dibenzo *para* dioxins and dibenzofurans. *Environ. Sci. Technol.*, **26**, 1023–1030.
- Tysklind, M., Tillitt, D., Eriksson, L., Lundgren, K. and Rappe, C. (1994) A toxic equivalency factor scale for polychlorinated dibenzofurans. *Fund. Appl. Toxicol.*, **22**, 277–285.
- Uddameri, V. and Kuchanur, M. (2004) Fuzzy QSARs for predicting log K_{oc} of persistent organic pollutants. *Chemosphere*, **54**, 771–776.
- Ugi, I., Wochner, M., Fontain, E., Bauer, J., Gruber, B. and Karl, R. (1990) Chemical similarity, chemical distance, and computer-assisted formalized reasoning by analogy, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 239–288.
- Unger, S.H., Cheung, P.S., Chiang, G.H. and Cook, J.R. (1986) RP-HPLC determination of 1-octanol. Partition and distribution coefficients: experience and results, in *Partition Coefficient: Determination and Estimation* (eds W.J. Dunn III, J.H. Block and R.S. Pearlman), Pergamon Books, New York, pp. 69–81.
- Unger, S.H., Cook, J.R. and Hollenberg, J.S. (1978) Simple procedure for determining octanol–aqueous partition, distribution, and ionization coefficients by reversed-phase high-pressure liquid chromatography. *J. Pharm. Sci.*, **67**, 1364–1367.
- Unger, S.H. and Hansch, C. (1973) On model building in structure–activity relationships. A reexamination of adrenergic blocking activity of β -halo- β -arylalkylamines. *J. Med. Chem.*, **16**, 745–749.
- Unger, S.H. and Hansch, C. (1976) Quantitative models of steric effects. *Prog. Phys. Org. Chem.*, **12**, 91–118.
- Unity Chemical Information Software, Tripos Associates, Inc., 1699 S Hanley Road, Suite 303, St. Louis, MO.
- Urbano-Cuadrado, M., Carbó, J.J., Maldonado, A.G. and Bo, C. (2007) New quantum mechanics-based three-dimensional molecular descriptors for use in QSSR approaches: application to asymmetric catalysis. *J. Chem. Inf. Model.*, **47**, 2228–2234.
- Urrestarazu Ramos, E., Vaes, W.H.J., Verhaar, H.J.M. and Hermens, J.L.M. (1998) Quantitative structure–activity relationships for the aquatic

- toxicity of polar and nonpolar narcotic pollutants. *J. Chem. Inf. Comput. Sci.*, **38**, 845–852.
- Ursu, O., Costescu, A., Diudea, M.V. and Párv, B. (2004) QSARs of some novel isosteric heterocyclic with antifungal activity. *Carpathian J. Math.*, **20**, 267–274.
- Ursu, O. and Diudea, M.V. (2003) Activity prediction by CLUJ-SIMIL program. *Rev. Roum. Chim.*, **48**, 321–330.
- Ursu, O. and Diudea, M.V. (2004) Topological descriptors in weighted molecular graphs, applications in QSPR modeling. *Studia Univ. Babes-Bolyai*, **49**, 69–74.
- Ursu, O. and Diudea, M.V. (2005) 3D molecular similarity: method, algorithms and case study on dopamine receptor antagonists. *Studia Univ. Babes-Bolyai*, **50**, 175–184.
- Ursu, O., Diudea, M.V. and Nakayama, S. (2006) 3D molecular similarity: method and algorithms. *J. Comp. Chem. (Japan)*, **5**, 39–46.
- Ursu, O., Diudea, M.V. and Nakayma, A. (2004) Quantitative structure–activity relationship study of COX-2 inhibitors. *Carpathian J. Math.*, **20**, 281–288.
- Ursu, O., Don, M., Katona, G., Jäntschi, L. and Diudea, M.V. (2004) QSAR study on dipeptide ace inhibitors. *Carpathian J. Math.*, **20**, 275–280.
- Vaes, W.H.J., Urrestarazu Ramos, E., Verhaar, H.J. M., Cramer, C.J. and Hermens, J.L.M. (1998) Understanding and estimating membrane/water partition coefficients: approaches to derive quantitative structure–property relationships. *Chem. Res. Toxicol.*, **11**, 847–854.
- Valkó, K., Bevan, C. and Reynolds, D. (1997) Chromatographic hydrophobicity index by fast-gradient RP-HPLC: a high-throughput alternative to log *P*/log *D*. *Anal. Chem.*, **69**, 2022–2029.
- Valkó, K., Espinosa, S., Du, C.M., Bosch, E., Rosés, M., Bevan, C. and Abraham, M.H. (2001) Unique selectivity of perfluorinated stationary phases with 2,2,2-trifluoroethanol as organic mobile phase modifier. *J. Chromat.*, **933**, 73–81.
- Valkó, K., Plass, M., Bevan, C., Reynolds, D. and Abraham, M.H. (1998) Relationships between the chromatographic hydrophobicity indices and solute descriptors obtained by using several reversed-phase, diol, nitrile, cyclodextrin and immobilised artificial membrane bonded high-performance liquid chromatography columns. *J. Chromat.*, **797**, 41–55.
- Valkó, K. and Slegel, P. (1992) Chromatographic separation and molecular modeling of triazines with respect to their inhibition of the growth of L1210/R71 cells. *J. Chromat.*, **592**, 59–63.
- Valkova, I., Vračko, M. and Basak, S.C. (2004) Modeling of structure–mutagenicity relationships: counter propagation neural network approach using calculated structural descriptors. *Anal. Chim. Acta*, **509**, 179–186.
- Vallat, P., Fan, W., El Tayar, N., Carrupt, P.-A. and Testa, B. (1992) Solvatochromic analysis of the retention mechanism of two novel stationary phases used for measuring lipophilicity by RP-HPLC. *J. Liquid Chromat.*, **15**, 2133–2151.
- Vallat, P., Gaillard, P., Carrupt, P.-A., Tsai, R.-S. and Testa, B. (1995) Structure–lipophilicity and structure–polarity relationships of amino acids and peptides. *Helv. Chim. Acta*, **78**, 471–485.
- van Aalten, D.M.F., Bywater, R., Findlay, J.B.C., Hendlich, M., Hooft, R.W.W. and Vriend, G. (1996) PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comput. Aid. Mol. Des.*, **10**, 255–262.
- van Bekkum, H., Verkade, P.E. and Wepster, B.M. (1959) Simple re-evaluation of the Hammett ρ – σ relation. *Recueil des Travaux Chimiques des Pays-Bas et de la Belgique*, **78**, 815–850.
- van de Waterbeemd, H. (1986) *Hydrophobicity of Organic Compounds*, Vol. 1, Computdrug, Budapest, Hungary.
- van de Waterbeemd, H. (1992) The history of drug research: from Hansch to the present. *Quant. Struct. -Act. Relat.*, **11**, 200–204.
- van de Waterbeemd, H. (1993) Recent progress in QSAR technology. *Drug Design & Discovery*, **9**, 277–285.
- van de Waterbeemd, H. (ed.) (1994) *Chemometric Methods in Molecular Design*, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 359.
- van de Waterbeemd, H. (ed.) (1995) *Advanced Computer-Assisted Techniques in Drug Discovery*, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 359.
- van de Waterbeemd, H. (ed.) (1996) *Structure–Property Correlations in Drug Research*, Academic Press/R.G. Landes Co., Austin, TX.
- van de Waterbeemd, H., Camenisch, G., Folkers, G. and Raevsky, O.A. (1996) Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quant. Struct. -Act. Relat.*, **15**, 480–490.
- van de Waterbeemd, H., Carrupt, P.-A., Testa, B. and Kier, L.B. (1993) Multivariate data modeling of new steric, topological and CoMFA-derived substituent parameters, in *Trends in QSAR and Molecular Modelling 92* (ed. C.G. Wermuth), ESCOM, Leiden, The Netherlands, pp. 69–75.

- van de Waterbeemd, H., Carter, R.E., Grassby, G., Kubinyi, H., Martin, Y.C., Tute, M.S., Willett, P., Haasnoot, C.A.G., Kier, L.B., Muller, K., Rose, S.V., Weber, J., Wibley, K.S., Wold, S., Boyd, D.B., Clark, D.E., Dehaen, C., Heindel, N.D., Kratochvil, P., Kutscher, B., Lewis, R.A., Mabilia, M., Metanowski, W.V., Polymeropoulos, E.E. and Tollenaere, J.P. (1997) Glossary of terms used in computational drug design. *Prot. Struct. Funct. Gen.*, **69**, 1137–1152.
- van de Waterbeemd, H., Clementi, S., Costantino, G., Carrupt, P.-A. and Testa, B. (1993) CoMFA-derived substituent descriptors for structure–property correlations, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 697–707.
- van de Waterbeemd, H., Costantino, G., Clementi, S., Cruciani, G. and Valigi, R. (1995) Experimental design in synthesis planning and structure–property correlations. Disjoint principal properties of organic substituents, in *Chemometric Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), VCH Publishers, New York, pp. 103–112.
- van de Waterbeemd, H., El Tayar, N., Carrupt, P.-A. and Testa, B. (1989) Pattern recognition study of QSAR substituent descriptors. *J. Comput. Aid. Mol. Des.*, **3**, 111–132.
- van de Waterbeemd, H. and Mannhold, R. (2008) Lipophilicity descriptors for structure–property correlation studies: overview of experimental and theoretical methods and a benchmark of log *P* calculations, in *Lipophilicity in Drug Action and Toxicology* (eds V. Pliška, B. Testa and H. van de Waterbeemd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 401–418.
- van de Waterbeemd, H., Smith, D.A., Beaumont, K. and Walker, D.K. (2001a) Property-based design: optimization of drug absorption and pharmacokinetics. *J. Med. Chem.*, **44**, 1313–1333.
- van de Waterbeemd, H., Smith, D.A. and Jones, B.C. (2001b) Lipophilicity in PK design: methyl, ethyl, futile. *J. Comput. Aid. Mol. Des.*, **15**, 273–286.
- van de Waterbeemd, H. and Testa, B. (1983) The development of a hydration factor “*o*” and its relation to correction terms in current hydrophobic fragmental systems. *Int. J. Pharm.*, **14**, 29–41.
- van de Waterbeemd, H. and Testa, B. (1987) The parametrization of lipophilicity and other structural properties in drug design, in *Advances in Drug Research*, Vol. 16 (ed. B. Testa), Academic Press, London, UK, pp. 85–225.
- van de Waterbeemd, H., Testa, B. and Folkers, G. (eds) (1997) *Computer-Assisted Lead Finding and Optimization*, Wiley-VCH Verlag GmbH, Weinheim, Germany, p. 554.
- Van der Voet, H. (1994) Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst.*, **25**, 313–323.
- van Haelst, A.G., Paulus, R.H.W.L. and Govers, H.A. J. (1997) Calculation of molecular volumes of tetrachlorobenzyltoluenes. *SAR & QSAR Environ. Res.*, **6**, 205–214.
- van Rhee, A.M. (2003) Use of recursion forest in the sequential screening process: consensus selection by multiple recursion trees. *J. Chem. Inf. Model.*, **43**, 941–948.
- van Rhee, A.M., Stocker, J., Printzenhoff, D., Creech, C., Wagoner, P.K. and Spear, K.L. (2001) Retrospective analysis of an experimental high-throughput screening data set by recursive partitioning. *J. Comb. Chem.*, **3**, 267–277.
- Van Vlaardingen, P.L.A., Steinhoff, W.J., de Voogt, P. and Admiraal, W.A. (1996) Property–toxicity relationships of azaarenes to the green alga *Scenedesmus acuminatus*. *Environ. Toxicol. Chem.*, **15**, 2035–2042.
- Vansteenk, B.J., Vanwijngaarden, I., Tulp, M.T. and Soudijn, W. (1994) Structure–affinity relationship studies on 5HT(1A) receptor ligands. 2. Heterobicyclic phenylpiperazines with N4-aralkyl substituents. *J. Med. Chem.*, **37**, 2761–2773.
- Vanyúr, R., Héberger, K. and Jakus, J. (2003) Prediction of anti-HIV-1 activity of a series of tetrapyrrole molecules. *J. Chem. Inf. Comput. Sci.*, **43**, 1829–1836.
- Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Springer, New York.
- Varkony, T.H., Shiloach, Y. and Smith, D.H. (1979) Computer-assisted examination of chemical compounds for structural similarities. *J. Chem. Inf. Comput. Sci.*, **19**, 104–111.
- Varmuza, K. (2003) Multivariate data analysis in chemistry, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1098–1133.
- Varmuza, K., Demuth, W., Karlovits, M. and Scsibraný, H. (2005) Binary substructure descriptors for organic compounds. *Croat. Chem. Acta*, **78**, 141–149.
- Varmuza, K., Karlovits, M. and Demuth, W. (2003) Spectral similarity versus structural similarity: infrared spectroscopy. *Anal. Chim. Acta*, **490**, 313–324.
- Varmuza, K., Penchev, P.N. and Scsibraný, H. (1998) Maximum common substructures of organic compounds exhibiting similar infrared spectra. *J. Chem. Inf. Comput. Sci.*, **38**, 420–427.

- Varmuza, K. and Scsibran, H. (2000) Substructure isomorphism matrix. *J. Chem. Inf. Comput. Sci.*, **40**, 308–313.
- Varnek, A., Fourches, D., Hoonakker, F. and Solov'ev, V.P. (2005) Substructural fragments: a universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aid. Mol. Des.*, **19**, 693–703.
- Varnek, A., Kireeva, N., Tetko, I.V., Baskin, I.I. and Solov'ev, V.P. (2007) Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? *J. Chem. Inf. Model.*, **47**, 1111–1122.
- Varnek, A., Wipff, G., Solov'ev, V.P. and Solotnov, A.F. (2002) Assessment of the macrocyclic effect for the complexation of crown ethers with alkali cations using the substructural molecular fragments method. *J. Chem. Inf. Comput. Sci.*, **42**, 812–829.
- Veber, D.F., Johnson, S.R., Cheng, H.-Y., Smith, B.R., Ward, K.W. and Kopple, K.D. (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, **45**, 2615–2623.
- Vedani, A. and Dobler, M. (2002) Multidimensional QSAR: moving from three- to five-dimensional concepts. *Quant. Struct. -Act. Relat.*, **21**, 382–390.
- Vedani, A., Dobler, M. and Lill, M.A. (2005) *In silico* prediction of harmful effects triggered by drugs and chemicals. *Toxicol. Appl. Pharm.*, **207**, S398–S407.
- Vedani, A., Dobler, M. and Zbinden, P. (1998) Quasi-atomistic receptor surface models: a bridge between 3-D QSAR and receptor modeling. *J. Am. Chem. Soc.*, **120**, 4471–4477.
- Vedani, A., McMasters, D.R. and Dobler, M. (2000) Multi-conformational ligand representation in 4D-QSAR: reducing the bias associated with ligand alignment. *Quant. Struct. -Act. Relat.*, **19**, 149–161.
- Vedrina, M., Marković, S., Medić-Šarić, M. and Trinajstić, N. (1998) TAM: a program for the calculation of topological indices in QSPR and QSAR studies. *Computers Chem.*, **21**, 355–361.
- Veith, G.D. and Mekenyan, O. (1993) A QSAR approach for estimating the aquatic toxicity of soft electrophiles (QSAR for soft electrophiles). *Quant. Struct. -Act. Relat.*, **12**, 349–356.
- Veith, G.D., Mekenyan, O.G., Ankley, G.T. and Call, D.J. (1995) A QSAR analysis of substituent effects on the photoinduced acute toxicity of PAHs. *Chemosphere*, **30**, 2129–2142.
- Veljković, V. (1980) *A Theoretical Approach to Preselection of Carcinogens and Chemical Carcinogenesis*, Gordon and Breach Science Publishers, New York.
- Veljković, V., Mouscadet, J.-F., Veljković, N., Glisic, S. and Debyser, Z. (2007) Simple criterion for selection of flavonoid compounds with anti-HIV activity. *Bioorg. Med. Chem. Lett.*, **17**, 1226–1232.
- Vendrame, R., Braga, R.S., Takahata, Y. and Galvão, D. S. (1999) Structure–activity relationship studies of carcinogenic activity of polycyclic aromatic hydrocarbons using molecular descriptors with principal component analysis and neural network methods. *J. Chem. Inf. Comput. Sci.*, **39**, 1094–1104.
- Vendrame, R., Braga, R.S., Takahata, Y. and Galvão, D. S. (2001) Structure–carcinogenic activity relationship studies of polycyclic aromatic hydrocarbons (PAHs) with pattern-recognition methods. *J. Mol. Struct. (Theochem)*, **539**, 252–265.
- Vendrame, R., Coluci, V.R., Braga, R.S. and Galvão, D.S. (2002) Structure–activity relationship (SAR) studies of the Tripos benchmark steroids. *J. Mol. Struct. (Theochem)*, **619**, 195–205.
- Vendrame, R., Ferreira, M.M.C., Collins, C.H. and Takahata, Y. (2002) Structure–activity relationships (SAR) of contraceptive progestogens studied with four different methods using calculated physico-chemical parameters. *J. Mol. Graph. Model.*, **20**, 345–358.
- Vendrame, R. and Takahata, Y. (1999) Structure–activity relationship (SAR) of substituted 17 α -acetoxyprogesterones studied with principal component analysis and neural networks using calculated physico-chemical. *J. Mol. Struct. (Theochem)*, **489**, 55–66.
- Venturelli, P., Menziani, M.C., Cocchi, M., Fanelli, F. and De Benedetti, P.G. (1992) Molecular modeling and quantitative structure–activity relationship analysis using theoretical descriptors of 1,4-benzodioxan (WB-4101) related compounds alpha-1-adrenergic antagonists. *J. Mol. Struct. (Theochem)*, **276**, 327–340.
- Verhaar, H.J.M., Eriksson, L., Sjöström, M., Schüürmann, G., Seinen, W. and Hermens, J.L.M. (1994) Modeling the toxicity of organophosphates: a comparison of the multiple linear regression and PLS regression methods. *Quant. Struct. -Act. Relat.*, **13**, 133–143.
- Verhaar, H.J.M., Urrestarazu Ramos, E. and Hermens, J.L.M. (1996) Classifying environmental pollutants. 2. Separation of class 1 (baseline toxicity) and class 2 ('polar narcosis') type compounds based on chemical descriptors. *J. Chemom.*, **10**, 149–162.
- Verhaar, H.J.M., Vanleeuwen, C.J. and Hermens, J. L.M. (1992) Classifying environmental pollutants. 1. Structure–activity relationships for prediction of aquatic toxicity. *Chemosphere*, **25**, 471–491.

- Verheij, H.J. (2006) Leadlikeness and structural diversity of synthetic screening libraries. *Mol. Div.*, **10**, 377–388.
- Verloop, A. (1972) The use of linear free energy parameters and other experimental constants in structure–activity studies, in *Drug Design*, Vol. 3 (ed. E.J. Ariens), Academic Press, New York, pp. 133–187.
- Verloop, A. (1985) QSAR and strategy in drug design of bioactive compounds, in *QSAR and Strategies in the Design of Bioactive Compounds* (ed. J.K. Seydel), VCH Publishers, Berlin, Germany, pp. 98–104.
- Verloop, A. (1987) *The STERIMOL Approach to Drug Design*, Marcel Dekker, New York.
- Verloop, A., Hoogenstraaten, W. and Tipker, J. (1976) Development and application of new steric substituent parameters in drug design, in *Drug Design*, Vol. 7 (ed. E.J. Ariens), Academic Press, New York, pp. 165–207.
- Verma, R.P. and Hansch, C. (2007) Understanding human rhinovirus infections in terms of QSAR. *Virology*, **359**, 152–161.
- Verma, R.P., Kapur, S., Berberena, O., Shusterman, A., Hansch, C. and Selassie, C.D. (2003) Synthesis, cytotoxicity, and QSAR analysis of X-thiophenols in rapidly dividing cells. *Chem. Res. Toxicol.*, **16**, 276–284.
- Vidal, D., Thormann, M. and Pons, M. (2005) LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.*, **45**, 386–393.
- Vidal, D., Thormann, M. and Pons, M. (2006) A novel search engine for virtual screening of very large databases. *J. Chem. Inf. Model.*, **46**, 836–843.
- Vieth, M., Siegel, M.G., Higgs, R.E., Watson, I.A., Robertson, D.H., Savin, K.A., Durst, G.L. and Hippskind, P.A. (2004) Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.*, **47**, 224–232.
- Vighi, M., Gramatica, P., Consolaro, F. and Todeschini, R. (2001) QSAR and chemometric approaches for setting water quality objectives for dangerous chemicals. *Ecotox. Environ. Safety*, **49**, 206–220.
- Vilar, S., Estrada, E., Uriarte, E., Santana, L. and Gutierrez, Y. (2005) *In silico* studies toward the discovery of new anti-HIV nucleoside compounds through the use of TOPS-MODE and 2D/3D connectivity indices. 2. Purine derivatives. *J. Chem. Inf. Model.*, **45**, 502–514.
- Villanueva-García, M., Gutiérrez-Parra, R.N., Martínez-Richa, A. and Robles, J. (2005) Quantitative structure–property relationships to estimate nematic transition temperatures in thermotropic liquid crystals. *J. Mol. Struct. (Theochem)*, **727**, 63–69.
- Villemin, D., Cherqaoui, D. and Cense, J.-M. (1993) Neural networks studies. Quantitative structure–activity relationship of mutagenic aromatic nitro compounds. *J. Chim. Phys. Phys-Chim. Biol.*, **90**, 1505–1519.
- Villemin, D., Cherqaoui, D. and Mesbah, A. (1994) Predicting carcinogenicity of polycyclic aromatic hydrocarbons from backpropagation neural network. *J. Chem. Inf. Comput. Sci.*, **34**, 1288–1293.
- Vinogradov, S.N. and Linnell, R.H. (1971) *Hydrogen Bonding*, Van Nostrand Reinhold Company, New York.
- Violon, D. (1999) Multiple regression analysis of octanol/water partition coefficients of non-ionic monomeric radiographic contrast media with the combination of three molecular descriptors. *Brit. J. Radiol.*, **72**, 44–47.
- Visco, D.P., Pophale, R.S., Rintoul, M.D. and Faulon, J.-L. (2002) Developing a methodology for an inverse quantitative structure–activity relationship using the signature molecular descriptor. *J. Mol. Graph. Model.*, **20**, 429–438.
- Vistoli, G., Pedretti, A., Villa, L. and Testa, B. (2005) Range and sensitivity as descriptors of molecular property spaces in dynamic QSAR analyses. *J. Med. Chem.*, **48**, 4947–4952.
- Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. and Robins, R.K. (1989) Atomic physico-chemical parameters for three-dimensional structure directed quantitative structure–activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.*, **29**, 163–172.
- Viswanadhan, V.N., Ghose, A.K., Singh, U.C. and Wendoloski, J.J. (1999) Prediction of solvation free energies of small organic molecules: additive-constitutive models based on molecular fingerprints and atomic constants. *J. Chem. Inf. Comput. Sci.*, **39**, 405–412.
- Viswanadhan, V.N., Ghose, A.K. and Wendoloski, J.J. (2000) Estimating aqueous solvation and lipophilicity of small organic molecules: a comparative overview of atom/group contribution methods. *Persp. Drug Disc. Des.*, **19**, 85–98.
- Viswanadhan, V.N., Mueller, G.A., Basak, S.C. and Weinstein, J.N. (2001) Comparison of a neural net-based QSAR algorithm (PCANN) with hologram and multiple linear regression-based QSAR approaches: application to 1,4-dihydropyridine-

- based calcium channel antagonists. *J. Chem. Inf. Comput. Sci.*, **41**, 505–511.
- Viswanadhan, V.N., Reddy, M.R., Bacquet, R.J. and Erion, M.D. (1993) Assessment of methods used for predicting lipophilicity: application to nucleosides and nucleoside bases. *J. Comput. Chem.*, **14**, 1019–1026.
- Voelkel, A. (1994) Structural descriptors in organic chemistry – New topological parameter based on electrotopological state of graph vertices. *Computers Chem.*, **18**, 1–4.
- Vöge, M., Guttman, A.J. and Jensen, I. (2002) On the number of benzenoid hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **42**, 456–466.
- Vogel, A.I. (1948) Physical properties and chemical constitution. Part XXIII. Miscellaneous compounds. Investigation of the so-called coordinate or dative link in esters of oxy-acids and in nitro-paraffins by molecular refractivity determinations. Atomic, structural, and group parachors and refractivities. *J. Chem. Soc.*, 1833–1855.
- Vogel, A.I., Cresswell, W.T., Jeffery, G.H. and Leicester, J. (1951) Calculation of the refractive indices of liquid organic compounds: bond molecular refraction. *Chem. & Ind.*, **5**, 376.
- Vogt, I. and Bajorath, J. (2007a) Analysis of a high-throughput screening data set using potency-scaled molecular similarity algorithms. *J. Chem. Inf. Model.*, **47**, 367–375.
- Vogt, M. and Bajorath, J. (2007b) Introduction of an information-theoretic method to predict recovery rates of active compounds for Bayesian *in silico* screening: theory and screening trials. *J. Chem. Inf. Model.*, **47**, 337–341.
- Vogt, M., Godden, J.W. and Bajorath, J. (2007) Bayesian interpretation of a distance function for navigating high-dimensional descriptor spaces. *J. Chem. Inf. Model.*, **47**, 39–46.
- Voiculescu, N., Balaban, A.T., Niculescu-Duvaz, I. and Simon, Z. (1990) *Modeling of Cancer Genesis and Prevention*, CRC Press, Boca Raton, FL.
- Voigt, J.H., Bienfait, B., Wang, S. and Nicklaus, M.C. (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.*, **41**, 702–712.
- Voigt, K. (2003) Databases on environmental information, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 722–742.
- Voigt, K., Brüggemann, R. and Pudenz, S. (2004) Chemical databases evaluated by order theoretical tools. *Anal. Bioanal. Chem.*, **380**, 467–474.
- Volkenstein, M.V. (1963) *Configurational Statistics of Polymeric Chains*, Wiley-Interscience, New York.
- von der Lieth, C.-W., Stumpf-Nothof, K. and Prior, U. (1996) A bond flexibility index derived from the constitution of molecules. *J. Chem. Inf. Comput. Sci.*, **36**, 711–716.
- von der Ohe, P.C., Kühne, R., Ebert, R.-U., Altenburger, R., Liess, M. and Schüürmann, G. (2005) Structural alerts – a new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. *Chem. Res. Toxicol.*, **18**, 536–555.
- von Homeyer, A. (2003) Evolutionary algorithms and their applications in chemistry, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1239–1280.
- von Homeyer, A. and Reitz, M. (2003) Databases in biochemistry and molecular biology, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 756–793.
- von Knop, J., Müller, W.R., Jericevic, Z. and Trinajstić, N. (1981) Computer enumeration and generation of trees and rooted trees. *J. Chem. Inf. Comput. Sci.*, **21**, 91–99.
- von Knop, J., Müller, W.R., Szymanski, K. and Trinajstić, N. (1991) On the determinant of the adjacency-plus-distance matrix as the topological index for characterizing alkanes. *J. Chem. Inf. Comput. Sci.*, **31**, 83–84.
- Votano, J.R., Parham, M., Hall, L.H. and Kier, L.B. (2004a) New predictors for several ADME/Tox properties: aqueous solubility, human oral absorption, and Ames genotoxicity using topological descriptors. *Mol. Div.*, **8**, 379–391.
- Votano, J.R., Parham, M., Hall, L.H., Kier, L.B., Oloff, S., Tropsha, A., Xie, Q. and Tong, W. (2004b) Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis*, **19**, 365–377.
- Vračko, M. (1997) A study of structure–carcinogenic potency relationship with artificial neural networks. The using of descriptors related to geometrical and electronic structures. *J. Chem. Inf. Comput. Sci.*, **37**, 1037–1043.
- Vračko, M. and Basak, S.C. (2004) Similarity study of proteomic maps. *Chemom. Intell. Lab. Syst.*, **70**, 33–38.
- Vrakas, D., Panderi, I., Hadjipavlou-Litina, D. and Tsantili-Kakoulidou, A. (2005) Investigation of the relationships between log *P* and various chromatographic indices for a series of substituted coumarins. Evaluation of their similarity/dissimilarity using multivariate statistics. *QSAR Comb. Sci.*, **24**, 254–260.

- Vukicević, D. (2003) Distinction between modifications of Wiener indices. *MATCH Commun. Math. Comput. Chem.*, **47**, 87–105.
- Vukicević, D. (2007) Comparing variable Zagreb indices. *MATCH Commun. Math. Comput. Chem.*, **57**, 633–641.
- Vukicević, D. and Graovac, A. (2004a) On modified Wiener indices of thorn graphs. *MATCH Commun. Math. Comput. Chem.*, **50**, 93–108.
- Vukicević, D. and Graovac, A. (2004b) On molecular graphs with valencies 1, 2 and 4 with prescribed numbers of bonds. *Croat. Chem. Acta*, **77**, 313–319.
- Vukicević, D. and Graovac, A. (2004c) Valence connectivity versus Randić, Zagreb and modified Zagreb index: a linear algorithm to check discriminative properties of indices in acyclic molecular graphs. *Croat. Chem. Acta*, **77**, 501–508.
- Vukicević, D. and Graovac, A. (2005) Compact valence sequences for molecules with single, double and triple covalent bonds. *Croat. Chem. Acta*, **78**, 203–209.
- Vukicević, D. and Graovac, A. (2007) Compact valence sequences for molecules with single, double and triple covalent bonds. II. Graphs with non-trivial cycles. *Croat. Chem. Acta*, **80**, 159–164.
- Vukicević, D. and Gutman, I. (2003) Note on a class of modified Wiener indices. *MATCH Commun. Math. Comput. Chem.*, **47**, 107–117.
- Vukicević, D., Miličević, A., Nikolić, S., Sedlar, J. and Trinajstić, N. (2005) Paths and walks in acyclic structures: plerographs versus kenographs. *ARKIVOC*, (x), 33–44.
- Vukicević, D. and Trinajstić, N. (2003) Modified Zagreb M_2 index – Comparison with the Randić connectivity index for benzenoid systems. *Croat. Chem. Acta*, **76**, 183–187.
- Vukicević, D. and Trinajstić, N. (2004) Wiener indices of benzenoid graphs. *Bull. Chem. Tech. Mac*, **23**, 113–129.
- Vukicević, D. and Trinajstić, N. (2005) Comparison of the Hosoya Z-indices for simple and general graphs of the same size. *Croat. Chem. Acta*, **78**, 235–239.
- Vukicević, D. and Žerovnik, J. (2003) New indices based on the modified Wiener indices. *MATCH Commun. Math. Comput. Chem.*, **47**, 119–132.
- Vukicević, D. and Žerovnik, J. (2005a) Altered Wiener indices. *Acta Chim. Sloven.*, **52**, 272–281.
- Vukicević, D. and Žerovnik, J. (2005b) Variable Wiener indices. *MATCH Commun. Math. Comput. Chem.*, **53**, 385–402.
- Wade, R.C. (1993) Molecular interaction fields, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 486–502.
- Wade, R.C. (2001) Derivation of QSARs using 3D structural models of protein–ligand complexes by COMBINE analysis, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 23–28.
- Wade, R.C. (2006) Calculation and application of molecular interaction fields, in *Molecular Interaction Fields*, Vol. 27 (ed. G. Cruciani), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 27–42.
- Wade, R.C., Clark, K.J. and Goodford, P.J. (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecule of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.*, **36**, 140–147.
- Wade, R.C. and Goodford, P.J. (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecule of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.*, **36**, 148–156.
- Wagener, M., Sadowski, J. and Gasteiger, J. (1995) Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.*, **117**, 7769–7775.
- Wagener, M. and van Geerestein, V.J. (2000) Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.*, **40**, 280–292.
- Wagner, G.C., Colvin, J.T., Allen, J.P. and Stapleton, H.J. (1985) Fractal models of protein structures, dynamics, and magnetic relaxation. *J. Am. Chem. Soc.*, **107**, 5589–5594.
- Waisser, K. (2001) Local parameters in QSAR, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 214–218.
- Wakeham, W.A., Cholakov, G.S. and Stateva, R.P. (2002) Liquid density and critical properties of hydrocarbons estimated from molecular structure. *J. Chem. Eng. Data*, **47**, 559–570.
- Walczak, B. (ed.) (2000) *Wavelet in Chemistry*, Elsevier, Amsterdam, The Netherlands, p. 572.
- Walczak, B. and Massart, D.L. (1999) Rough sets theory. *Chemom. Intell. Lab. Syst.*, **47**, 1–16.
- Walikar, H.B., Shigehalli, V.S. and Ramane, H.S. (2004) Bounds on the Wiener number of a graph. *MATCH Commun. Math. Comput. Chem.*, **50**, 117–132.
- Walker, J.D., Jaworska, J.S., Comber, M.H.I., Schultz, T.W. and Dearden, J.C. (2003) Guidelines for developing and using quantitative structure–

- activity relationships. *Environ. Toxicol. Chem.*, **22**, 1653–1665.
- Walker, P.D., Maggiora, G.M., Johnson, M.A., Petke, J.D. and Mezey, P.G. (1995) Shape group analysis of molecular similarity: shape similarity of six-membered aromatic ring systems. *J. Chem. Inf. Comput. Sci.*, **35**, 568–578.
- Waller, C.L. (1994) A three-dimensional technique for the calculation of octanol–water partition coefficients. *Quant. Struct. -Act. Relat.*, **13**, 172–176.
- Waller, C.L. and Bradley, M.P. (1999) Development and validation of a novel variable selection technique with application to multidimensional quantitative structure–activity relationship studies. *J. Chem. Inf. Comput. Sci.*, **39**, 345–355.
- Waller, C.L., Evans, M.V. and McKinney, J.D. (1996) Modeling the cytochrome P450 mediated metabolism of chlorinated volatile organic compounds. *Drug Metab. Disposition*, **24**, 203–210.
- Waller, C.L. and Kellogg, G.E. (1996) Adding chemical information to CoMFA models with alternative 3D QSAR fields. *Network Science – Computational Chemistry*, <http://www.netsci.org/Science/Compchem/feature10.html>.
- Waller, C.L. and Marshall, G.R. (1993) Three-dimensional quantitative structure–activity relationship of angiotensin converting enzyme and thermolysin inhibitors. 2. A comparison of CoMFA models incorporating molecular orbital fields and desolvation free energies based on active analog and complementary receptor field alignment rules. *J. Med. Chem.*, **36**, 2390–2403.
- Waller, C.L. and McKinney, J.D. (1992) Comparative molecular field analysis of polyhalogenated dibenzo-*p*-dioxins, dibenzofurans, and biphenyls. *J. Med. Chem.*, **35**, 3660–3666.
- Waller, C.L. and McKinney, J.D. (1995) Three-dimensional quantitative structure–activity relationships of dioxins and dioxin-like compounds: model validation and Ah receptor characterization. *Chem. Res. Toxicol.*, **8**, 847–858.
- Waller, C.L., Minor, D.L. and McKinney, J.D. (1995) Using three-dimensional quantitative structure–activity relationships to examine estrogen receptor binding affinities of polychlorinated hydroxy-biphenyls. *Environ. Health Persp.*, **103**, 702–707.
- Waller, C.L., Oprea, T.I., Chae, K., Park, H.-K., Korach, K.S., Laws, S.C., Wiese, T.E., Kelce, W.R. and Gray, L.E., Jr (1996) Ligand-based identification of environmental estrogens. *Chem. Res. Toxicol.*, **9**, 1240–1248.
- Waller, C.L., Oprea, T.I., Giolitti, A. and Marshall, G.R. (1993) Three-dimensional QSAR of human immunodeficiency virus (I) protease inhibitors. 1. A CoMFA study employing experimentally determined alignment rules. *J. Med. Chem.*, **36**, 4152–4160.
- Waller, C.L., Wyrick, S.D., Kemp, W.E., Park, H.M. and Smith, F.T. (1994) Conformational analysis, molecular modeling, and quantitative structure–activity relationship studies of agents for the inhibition of astrocytic chloride transport. *Pharm. Res.*, **11**, 47–53.
- Walsh, D.B. and Claxton, L.D. (1987) Computer-assisted structure–activity relationships of nitrogenous cyclic compounds tested in *Salmonella* assays for mutagenicity. *Mut. Res.*, **182**, 55–64.
- Walters, C.J., Caviness, K. and Hefferlin, R.A. (2004) Global molecular identification from graphs. IV. Molecules with four closed p-shell atoms and beyond. *Croat. Chem. Acta*, **77**, 65–71.
- Walters, D.E. (1998) Genetically evolved receptor models (GERM) as a 3D QSAR tool, in *3D QSAR in Drug Design*, Vol. 3 (eds H. Kubinyi, G. Folkers and Y.C. Martin), Kluwer/ESCOM, Dordrecht, The Netherlands, pp. 159–166.
- Walters, D.E. and Hinds, R.M. (1994) Genetically evolved receptor models: a computational approach to construction of receptor models. *J. Med. Chem.*, **37**, 2527–2537.
- Walters, D.E. and Hopfinger, A.J. (1986) Case studies of the application of molecular shape analysis to elucidate drug action. *J. Mol. Struct. (Theochem)*, **134**, 317–323.
- Walters, W.P., Ajay and Muresan, S. (1999) Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.*, **3**, 384–386.
- Walters, W.P. and Murcko, M.A. (2000) Library filtering systems and prediction of drug-like properties, in *Virtual Screening for Bioactive Molecules*, Vol. 10 (eds H.J. Bohm and G. Schneider), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 15–30.
- Walters, W.P. and Murcko, M.A. (2002) Prediction of ‘drug-likeness’. *Adv. Drug Deliv. Rev.*, **54**, 255–271.
- Walters, W.P., Stahl, M.T. and Murcko, M.A. (1998) Virtual screening – an overview. *Drug Discov. Today*, **3**, 160–178.
- Walther, D. (1974) *J. Prakt. Chem.*, **316**, 604.
- Wan, J., Zhang, L., Yang, G. and Zhan, C.-G. (2004) Quantitative structure–activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: a study of quantum chemical descriptors from density functional theory. *J. Chem. Inf. Comput. Sci.*, **44**, 2009–2105.
- Wang, C.-X., Shi, Y.-Y. and Huang, F.-H. (1990) Fractal study of tertiary structure of proteins. *Phys. Rev. A*, **41**, 7043.

- Wang, G. and Bai, N. (1998) Structure–activity relationships for rat and mouse LD50 of miscellaneous alcohols. *Chemosphere*, **36**, 1475–1483.
- Wang, J. and Wang, W. (2006) New 2-D graphical representation of DNA sequences. *Biophys. Rev. Lett.*, **1**, 133–140.
- Wang, J. and Ramnarayan, K. (1999) Toward designing drug-like libraries: a novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.*, **1**, 524–533.
- Wang, J., Krudy, G., Hou, T.-J., Zhang, W., Holland, G. and Xu, X. (2007) Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.*, **47**, 1395–1404.
- Wang, J., Wang, W., Kollman, P.A. and Case, D.A. (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, **25**, 247–260.
- Wang, R., Fu, Y. and Lai, L. (1997) A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.*, **37**, 615–621.
- Wang, R., Gao, Y. and Lai, L. (2000) Calculating partition coefficient by atom-additive method. *Persp. Drug Disc. Des.*, **19**, 47–66.
- Wang, S. and Milne, G.W.A. (1993) Applications of computers to toxicological research. *Chem. Res. Toxicol.*, **6**, 748–753.
- Wang, S., Milne, G.W.A. and Klopman, G. (1994) Graph theory and group contributions in the estimation of boiling points. *J. Chem. Inf. Comput. Sci.*, **34**, 1242–1250.
- Wang, S., Xue, C., Chen, X., Liu, M. and Hu, Z. (2004) Study on the quantitative relationship between the structures and electrophoretic mobilities of flavonoids in micellar electrokinetic capillary chromatography. *J. Chromat.*, **1033**, 153–159.
- Wang, T. and Wade, R.C. (2001) COMBINE 3D-QSAR analysis of influenza neuramidinase inhibitors, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona, Spain, pp. 78–82.
- Wang, T. and Zhou, J. (1998) 3DFS: a new 3D flexible searching system for use in drug design. *J. Chem. Inf. Comput. Sci.*, **38**, 71–77.
- Wang, X.Z. and Chen, B.H. (1998) Clustering of infrared spectra of lubricating base oils using adaptive resonance theory. *J. Chem. Inf. Comput. Sci.*, **38**, 457–462.
- Wang, X., Dong, Y., Wang, L.-S. and Han, S. (2001) Acute toxicity of substituted phenols to *Rana japonica* tadpoles and mechanism-based quantitative structure–activity relationship (QSAR) study. *Chemosphere*, **44**, 447–455.
- Wang, X., Sun, C., Wang, Yu. and Wang, L.-S. (2002) Quantitative structure–activity relationships for the inhibition toxicity to root elongation of *Cucumis sativus* of selected phenols and interspecies correlation with *Tetrahymena pyriformis*. *Chemosphere*, **46**, 153–161.
- Wang, X., Tang, S., Liu, S., Cui, S. and Wang, L.-S. (2003) Molecular hologram derived quantitative structure–property relationships to predict physico-chemical properties of polychlorinated biphenyls. *Chemosphere*, **51**, 617–632.
- Wang, X., Yin, C.-S. and Wang, L.-S. (2002) Structure–activity relationships and response–surface analysis of nitroaromatics toxicity to the yeast (*Saccharomyces cerevisiae*). *Chemosphere*, **46**, 1045–1051.
- Wang, X., Yu, J., Wang, Yu. and Wang, L.-S. (2002) Mechanism-based quantitative structure–activity relationships for the inhibition of substituted phenols on germination rate of *Cucumis sativus*. *Chemosphere*, **46**, 241–250.
- Wang, Y.-G. and Werstiuik, N.H. (2003) A practical and efficient method to calculate AIM localization and delocalization indices at post-HF levels of theory. *J. Comput. Chem.*, **24**, 379–385.
- Wang, Y.-H., Li, Y., Yang, S.-L. and Yang, L. (2005) An *in silico* approach for screening flavonoids as P-glycoprotein inhibitors based on a Bayesian-regularized neural network. *J. Comput. Aid. Mol. Des.*, **19**, 137–147.
- Wang, Y. and Bajorath, J. (2008) Balancing the influence of molecular complexity on fingerprint similarity searching. *J. Chem. Inf. Model.*, **48**, 75–84.
- Wang, Y., Eckert, H. and Bajorath, J. (2007) Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem*, **2**, 1037–1042.
- Wang, Y., Godden, J.W. and Bajorath, J. (2007) A novel descriptor histogram filtering method for database mining and the identification of active molecules. *Letters in Drug Design and Discovery*, **4**, 286–292.
- Wang, Z.-Y., Zhai, Z. and Wang, L.-S. (2005) Quantitative structure–activity relationship of toxicity of alkyl(1-phenylsulfonyl) cycloalkane-carboxylates using MLSE model and *ab initio*. *QSAR Comb. Sci.*, **24**, 211–217.
- Wania, F. and Dugani, C.B. (2003) Assessing the long-range transport potential of polybrominated diphenyl ethers: a comparison of four multimedia models. *Environ. Toxicol. Chem.*, **22**, 1252–1261.
- Warne, M.S., Boyd, M.A., Meharg, E.M., Osborn, D., Killham, D., Lindon, J.C. and Nicholson, J.K. (1999) Quantitative structure–toxicity

- relationships for halobenzenes in two species of bioluminescent bacteria, *Pseudomonas fluorescens* and *Vibrio fischeri*, using an atom-centered semi-empirical molecular-orbital based model. *SAR & QSAR Environ. Res.*, **10**, 17–38.
- Warne, M.S., Connell, D.W., Hawker, D.W. and Schüürmann, G. (1989a) Prediction of the toxicity of mixtures of shale oil components. *Ecotox. Environ. Safety*, **18**, 121–128.
- Warne, M.S., Connell, D.W., Hawker, D.W. and Schüürmann, G. (1989b) Quantitative structure–activity relationships for the toxicity of selected shale oil components to mixed marine bacteria. *Ecotox. Environ. Safety*, **17**, 133–148.
- Warne, M.S., Osborn, D., Lindon, J.C. and Nicholson, J.K. (1999) Quantitative structure–toxicity relationships for halogenated substituted-benzenes to *Vibrio fischeri*, using atom-based semi-empirical molecular-orbital descriptors. *Chemosphere*, **38**, 3357–3382.
- Warthen, J.D., Schmidt, W.F., Cunningham, R.T., Demilo, A.B. and Fritz, G.L. (1993) Quantitative structure–activity relationships (QSAR) of trimedlure isomers. *J. Chem. Ecol.*, **19**, 1323–1335.
- Wayner, D.D.M. and Arnold, D.R. (1984) Substituent effects on benzyl radical hyperfine coupling constants. Part 2. The effect of sulphur substituents. *Can. J. Chem.*, **62**, 1164–1168.
- Weber, A., Teckentrup, A. and Briem, H. (2002) Flexsim-R: a virtual affinity fingerprint descriptor to calculate similarities of functional groups. *J. Comput. Aid. Mol. Des.*, **16**, 903–916.
- Weber, K.C., Honório, K.M., Bruni, A.T. and da Silva, A.B.F. (2006) The use of classification methods for modeling the antioxidant activity of flavonoid compounds. *J. Mol. Model.*, **12**, 915–920.
- Weckwerth, J.D., Vitha, M.F. and Carr, P.W. (2001) The development and determination of chemically distinct solute parameters for use in linear solvation energy relationships. *Fluid Phase Equil.*, **183–184**, 143–157.
- Wegner, J.K., Fröhlich, H., Mielenz, H.M. and Zell, A. (2006) Data and graph mining in chemical space for ADME and activity data sets. *QSAR Comb. Sci.*, **25**, 205–220.
- Wegner, J.K. and Zell, A. (2003) Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.*, **43**, 1077–1084.
- Wehrens, R., Pretsch, E. and Buydens, L. (1998) Quality criteria of genetic algorithms for structure optimization. *J. Chem. Inf. Comput. Sci.*, **38**, 151–157.
- Wehrens, R., Pretsch, E. and Buydens, L. (1999) The quality of optimization by genetic algorithms. *Anal. Chim. Acta*, **388**, 265–271.
- Wehrens, R., Putter, H. and Buydens, L. (2000) The bootstrap: a tutorial. *Chemom. Intell. Lab. Syst.*, **54**, 35–52.
- Wei, D., Zhang, A., Wu, C., Han, S. and Wang, L.-S. (2001) Progressive study and robustness test of QSAR model based on quantum chemical parameters for predicting BCF of selected polychlorinated organic compounds (PCOCs). *Chemosphere*, **44**, 1421–1428.
- Weiner, M.L. and Weiner, P.H. (1973) A study of structure–activity relationships of a series of diphenylaminopropanols by factor analysis. *J. Med. Chem.*, **16**, 655–661.
- Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
- Weininger, D. (1990) SMILES. 3. DEPICT: graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.*, **30**, 237–243.
- Weininger, D. (2003) SMILES – a language for molecules and reactions, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 195–205.
- Weininger, D., Weininger, A. and Weininger, J.L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, **29**, 97–101.
- Weininger, D. and Weininger, J.L. (1990) Chemical structures and computers, in *Quantitative Drug Design*, Vol. 4 (ed. C.A. Ramsden), Pergamon Press, Oxford, UK, pp. 59–82.
- Weininger, S.J. (1984) The molecular structure conundrum: can classical chemistry be reduced to quantum chemistry? *J. Chem. Educ.*, **61**, 939–944.
- Weinstein, J.N., Kohn, K.W., Grever, M.R., Viswanadhan, V.N., Rubinstein, L.V., Monks, A.P., Scudiero, D.A., Welch, L., Koutsoukos, A.D., Chiousa, A.J. and Paull, K.D. (1992) Neural computing in cancer drug development: predicting mechanism of action. *Science*, **258**, 447–451.
- Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J., Kohn, K.W., Fojo, T., Bates, S. E., Rubinstein, L.V., Anderson, N.L., Buolamwini, J.K., van Osdol, W.W., Monks, A.P., Scudiero, D.A., Sausville, E.A., Zaharevitz, D.W., Bunow, B., Viswanadhan, V.N., Johnson, G.S., Wittes, R.E. and Paull, K.D. (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science*, **275**, 343–349.

- Weis, D.C., Faulon, J.-L., LeBorne, R.C. and Visco, D. P., Jr (2005) The signature molecular descriptor. 5. The design of hydrofluoroether foam blowing agents using inverse-QSAR. *Ind. Eng. Chem. Res.*, **44**, 8883–8891.
- Wells, M.J.M., Clark, C.R. and Patterson, R.M. (1981) Correlation of reversed-phase capacity factors for barbiturates with biological activities, partition coefficients, and molecular connectivity indices. *J. Chromatogr. Sci.*, **19**, 573.
- Wells, M.J.M., Clark, C.R. and Patterson, R.M. (1982) Investigation of N-alkylbenzamides by reversed-phase liquid chromatography. III. Correlation of chromatographic parameters with molecular connectivity indices for C₁ to C₅ N-alkylbenzamides. *J. Chromat.*, **235**, 61–74.
- Wells, P.R. (1968a) Group electronegativities. *Prog. Phys. Org. Chem.*, **6**, 111–145.
- Wells, P.R. (1968b) *Linear Free Energy Relationships*, Academic Press, New York.
- Wellsow, J., Machulla, H.-J. and Kovar, K.-A. (2002) 3D QSAR of serotonin transporter ligands: CoMFA and CoMSIA studies. *Quant. Struct. -Act. Relat.*, **21**, 577–589.
- Wenlock, M.C., Austin, R.P., Barton, P., Davis, A.M. and Leeson, P.D. (2003) A comparison of physico-chemical property profiles of development and marketed oral drugs. *J. Med. Chem.*, **46**, 1250–1256.
- Wentang, C., Ying, Z. and Feibai, Yu. (1993) New computer representation for chemical structures: two-level compact connectivity tables. *J. Chem. Inf. Comput. Sci.*, **33**, 604–608.
- Wermuth, C.G. (ed.) (1993) *Trends in QSAR and Molecular Modelling* 92, ESCOM, Leiden, The Netherlands, p. 595.
- Wermuth, C.G. (ed.) (1996) *The Practice of Medicinal Chemistry*, Academic Press, Cambridge, UK, p. 968.
- Werther, W., Demuth, W., Krueger, F.R., Kissel, J., Schmid, E.R. and Varmuza, K. (2002) Evaluation of mass spectra from organic compounds assumed to be present in cometary grains. Exploratory data analysis. *J. Chemom.*, **16**, 99–110.
- Wessel, M.D. and Jurs, P.C. (1994) Prediction of reduced ion mobility constants from structural information using multiple linear regression analysis and computational neural networks. *Anal. Chem.*, **66**, 2480–2487.
- Wessel, M.D. and Jurs, P.C. (1995a) Prediction of normal boiling points for a diverse set of industrially important organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **35**, 841–850.
- Wessel, M.D. and Jurs, P.C. (1995b) Prediction of normal boiling points of hydrocarbons from molecular structure. *J. Chem. Inf. Comput. Sci.*, **35**, 68–76.
- Wessel, M.D., Jurs, P.C., Tolan, J.W. and Muskal, S.M. (1998) Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, **38**, 726–735.
- Wessel, M.D., Sutter, J.M. and Jurs, P.C. (1996) Prediction of reduced ion mobility constants of organic compounds from molecular structure. *Anal. Chem.*, **68**, 4237–4243.
- Westheimer, F.H. and Kirkwood, J.J. (1938) The electrostatic influence of substituents on the dissociation constants of organic acids. II. *J. Chim. Phys.*, **6**, 513–517.
- Wheland, G.W. (1955) *Resonance in Organic Chemistry*, John Wiley & Sons, Inc., New York.
- White, J.H. (1969) Self-linking and the Gauss integral in higher dimension. *Am. J. Math.*, **91**, 693.
- Whitley, D.C. (1998) van der Waals surface graphs and the shape of small rings. *J. Chem. Inf. Comput. Sci.*, **38**, 906–914.
- Whitley, D.C., Ford, M.G. and Livingstone, D.J. (2000) Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.*, **40**, 1160–1168.
- Wiberg, K.B. (1968) Application of the Pople–Santry–Segal CNDO method to the cyclopropylcarbiny and cyclobutyl cation and to bicyclobutane. *Tetrahedron*, **24**, 1083–1096.
- Wiener, H. (1947a) Correlation of heat of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.*, **69**, 2636–2638.
- Wiener, H. (1947b) Influence of interatomic forces on paraffin properties. *J. Chim. Phys.*, **15**, 766.
- Wiener, H. (1947c) Structural determination of paraffin boiling points. *J. Am. Chem. Soc.*, **69**, 17–20.
- Wiener, H. (1948a) Relationship of physical properties of isomeric alkanes to molecular structure surface tension, specific dispersion and critical solution temperature in aniline. *J. Phys. Colloid Chem.*, **52**, 1082–1089.
- Wiener, H. (1948b) Vapour pressure–temperature relations among the branched paraffin hydrocarbons. *J. Phys. Chem.*, **52**, 425–430.
- Wierl, K. (1931) Elektronenbeugung und Molekulbau. *Ann. Phys. (Leipzig)*, **8**, 521–564.
- Wiese, M. (1993) The hypothetical active-site lattice, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 431–442.

- Wiggins, G.D. (2003) Overview of databases/data sources, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 496–506.
- Wikel, J.H. and Dow, E.R. (1993) The use of neural networks for variable selection in QSAR. *Bioorg. Med. Chem. Lett.*, **3**, 645–651.
- Wilcox, C.F., Jr (1968) Solubility of molecules containing (4n)-rings. *Tetrahedron Lett.*, **7**, 795–800.
- Wilcox, C.F., Jr (1969) Stability of molecules containing nonalternant rings. *J. Am. Chem. Soc.*, **91**, 2732–2736.
- Wild, D.J. and Blankley, C.J. (2000) Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.*, **40**, 155–162.
- Wilding, W.V. and Rowley, R.L. (1986) A four parameter corresponding states method for the prediction of thermodynamic properties of polar and nonpolar fluids. *Int. J. Thermophys.*, **7**, 525–539.
- Wildman, S.A. and Crippen, G.M. (1999) Prediction of physico-chemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.*, **39**, 868–873.
- Wildman, S.A. and Crippen, G.M. (2001) Evaluation of ligand overlap by atomic parameters. *J. Chem. Inf. Comput. Sci.*, **41**, 446–450.
- Wildman, S.A. and Crippen, G.M. (2002) Three-dimensional molecular descriptors and a novel QSAR method. *J. Mol. Graph. Model.*, **21**, 161–170.
- Wildman, S.A. and Crippen, G.M. (2003) Validation of DAPPER for 3D QSAR: conformational search and chirality metric. *J. Chem. Inf. Comput. Sci.*, **43**, 629–636.
- Wilkerson, W.W. (1995) A quantitative structure–activity relationship analysis of a series of 2'-(2,4 difluorophenoxy)-4'-substituted methanesulfonilides. *Eur. J. Med. Chem.*, **30**, 191–197.
- Wilkerson, W.W., Copeland, R.A., Covington, M. and Trzaskos, J.M. (1995) Antiinflammatory 4,5-diarylpyrroles. 2. Activity as a function of cyclooxygenase-2 inhibition. *J. Med. Chem.*, **38**, 3895–3901.
- Wilkerson, W.W., Galbraith, W., Gansbrangs, K., Grubb, M., Hewes, W.E., Jaffee, B., Kenney, J.P., Kerr, J. and Wong, N. (1994) Antiinflammatory 4,5-diarylpyrroles synthesis and QSAR. *J. Med. Chem.*, **37**, 988–998.
- Wilkins, C.L. and Randić, M. (1980) A graph theoretical approach to structure–property and structure–activity correlations. *Theor. Chim. Acta*, **58**, 69–71.
- Wilkins, C.L., Randić, M., Schuster, S.M., Markin, R.S., Steiner, S. and Dorgan, L. (1981) A graph-theoretic approach to quantitative structure–activity/ reactivity studies. *Anal. Chim. Acta*, **133**, 637–645.
- Willauer, H.D., Huddleston, J.G. and Rogers, R.D. (2002) Solvent properties of aqueous biphasic systems composed of polyethylene glycol and salt characterized by the free energy of transfer of a methylene group between the phases and by a linear solvation energy relationship. *Ind. Eng. Chem. Res.*, **41**, 2591–2601.
- Willett, P. (1987) *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth, UK, p. 254.
- Willett, P. (1988) Ranking and clustering of chemical structure databases, in *Physical Property Prediction in Organic Chemistry* (eds C. Jochum, M.G. Hicks and J. Sunkel), Springer-Verlag, Berlin, Germany, pp. 191–207.
- Willett, P. (1990) Algorithms for the calculation of similarity in chemical structure databases, in *Concepts and Applications of Molecular Similarity* (eds M.A. Johnson and G.M. Maggiora), John Wiley & Sons, Inc., New York, pp. 43–63.
- Willett, P. (1991) *Three-Dimensional Chemical Structure Handling*, Research Studies Press–Wiley, Taunton, UK.
- Willett, P. (1997) Computational tools for the analysis of molecular diversity. *Persp. Drug Disc. Des.*, **7/8**, 1–11.
- Willett, P. (2003a) Similarity searching in chemical structure databases, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 904–916.
- Willett, P. (2003b) Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.*, **31**, 603–606.
- Willett, P., Barnard, J.M. and Downs, G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, **38**, 983–996.
- Willett, P. and Winterman, V.A. (1986) A comparison of some measures for the determination of inter-molecular structural similarity. *Quant. Struct. -Act. Relat.*, **5**, 18–25.
- Willett, P., Winterman, V.A. and Bawden, D. (1986) Implementation of nearest neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.*, **26**, 36–41.
- Williams, C. (2006) Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol. Div.*, **10**, 311–332.
- Williford, C.J. and Stevens, E.P. (2006) Strain energies as a steric descriptor in QSAR calculations. *QSAR Comb. Sci.*, **23**, 495–505.

- Wilson, L.Y. and Famini, G.R. (1991) Using theoretical descriptors in quantitative structure–activity relationships: some toxicological indices. *J. Med. Chem.*, **34**, 1668–1674.
- Wilson, N.S., Dolan, J.W., Snyder, L.R., Carr, P.W. and Sander, L.C. (2002) Column selectivity in reversed-phase liquid chromatography. III. The physico-chemical basis of selectivity. *J. Chromat.*, **961**, 217–236.
- Wilson, N.S., Nelson, M.D., Dolan, J.W., Snyder, L.R. and Carr, P.W. (2002a) Column selectivity in reversed-phase liquid chromatography. II. Effect of a change in conditions. *J. Chromat.*, **961**, 195–215.
- Wilson, N.S., Nelson, M.D., Dolan, J.W., Snyder, L.R., Wolcott, R.G. and Carr, P.W. (2002b) Column selectivity in reversed-phase liquid chromatography. I. A general quantitative relationship. *J. Chromat.*, **961**, 171–193.
- Wilson, R.J. (1972) *Introduction to Graph Theory*, Oliver & Boyd, Edinburgh, UK.
- Wilton, D. and Willett, P. (2003) Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci.*, **43**, 469–474.
- Winberg, N. and Mislow, K. (1995) A unification of chirality measures. *J. Math. Chem.*, **17**, 35–53.
- Winget, P., Cramer, C.J. and Truhlar, D.G. (2000) Prediction of soil sorption coefficients using a universal solvation model. *Environ. Sci. Technol.*, **34**, 4733–4740.
- Winiwarter, S., Ax, F., Lennernäs, H., Hallberg, A., Pettersson, C. and Karlén, A. (2003) Hydrogen bonding descriptors in the prediction of human *in vivo* intestinal permeability. *J. Mol. Graph. Model.*, **21**, 273–287.
- Winiwarter, S., Bonham, N.M., Ax, F., Hallberg, A., Lennernäs, H. and Karlén, A. (1998) Correlation of human jejunal permeability (*in vivo*) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. *J. Med. Chem.*, **41**, 4939–4949.
- Winkler, D.A. and Burden, F.R. (1998) Holographic QSAR of benzodiazepines. *Quant. Struct. -Act. Relat.*, **17**, 224–231.
- Winkler, D.A., Burden, F.R. and Watkins, A.J.R. (1998) Atomistic topological indices applied to benzodiazepines using various regression methods. *Quant. Struct. -Act. Relat.*, **17**, 14–19.
- Wintner, E.A. and Moallemi, C.C. (2000) Quantized surface complementarity diversity (QSCD): a model based on small molecule-target complementarity. *J. Med. Chem.*, **43**, 1993–2006.
- Wipke, W.T. and Dyott, T.M. (1974a) Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry. *J. Am. Chem. Soc.*, **96**, 4825–4834.
- Wipke, W.T. and Dyott, T.M. (1974b) Stereochemically unique naming algorithm. *J. Am. Chem. Soc.*, **96**, 4834–4842.
- Wipke, W.T., Krishnan, S. and Ouchi, G.I. (1978) Hash functions for rapid storage and retrieval of chemical structures. *J. Chem. Inf. Comput. Sci.*, **18**, 32–37.
- Wirth, K. (1986) Coding of relational descriptions of molecular structures. *J. Chem. Inf. Comput. Sci.*, **26**, 242–249.
- Wise, S.A., Bonnett, W.J., Guenther, F.R. and May, W. E. (1981) A relationship between reversed-phase C₁₈ liquid chromatographic retention and the shape of polycyclic aromatic hydrocarbons. *J. Chromatogr. Sci.*, **19**, 457–465.
- Wisniewski, J.L. (2003) Chemical nomenclature and structure representation: algorithmic generation and conversion, in *Handbook of Chemoinformatics*, Vol. 1 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 51–79.
- Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, **20**, 397–405.
- Wold, S. (1991) Validation of QSAR's. *Quant. Struct. -Act. Relat.*, **10**, 191–193.
- Wold, S. (1995) PLS for multivariate linear modeling, in *Chemometric Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 195–218.
- Wold, S. and Dunn, W.J. III (1983) Multivariate quantitative structure–activity relationships (QSAR): conditions for their applicability. *J. Chem. Inf. Comput. Sci.*, **23**, 6–13.
- Wold, S. and Eriksson, L. (1995) Statistical validation of QSAR results. Validation tools, in *Chemometrics Methods in Molecular Design*, Vol. 2 (ed. H. van de Waterbeemd), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 309–318.
- Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M., Skagerberg, B. and Wikström, C. (1987) Principal property values for six non-coded amino acids and their application to a structure–activity relationship for oxytocin peptide analogues. *Can. J. Chem.*, **65**, 1814–1820.
- Wold, S., Johansson, E. and Cocchi, M. (1993) PLS – partial least squares projection of latent structures, in *3D QSAR in Drug Design. Theory, Methods and Applications* (ed. H. Kubinyi), ESCOM, Leiden, The Netherlands, pp. 523–550.
- Wold, S., Jonsson, J., Sjöström, M., Sandberg, M. and Rännar, S. (1993) DNA and peptide sequences and chemical processes multivariately modelled by

- principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta*, **277**, 239–253.
- Wold, S., Kettaneh-Wold, N. and Tjessem, K. (1996) Hierarchical multiblock PLS and PC models for easier interpretation and as an alternative to variable selection. *J. Chemom.*, **10**, 463–482.
- Wold, S., Ruhe, A., Wold, H. and Dunn, W.J. III (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.*, **5**, 735–743.
- Wold, S. and Sjöström, M. (1978) Linear free energy relationships as tools for investigating chemical similarity. Theory and practice, in *Correlation Analysis in Chemistry* (eds N.B. Chapman and J. Shorter), Plenum Press, New York, pp. 1–54.
- Wold, S. and Sjöström, M. (1998) Chemometrics, present and future success. *Chemom. Intell. Lab. Syst.*, **44**, 3–14.
- Wold, S., Sjöström, M. and Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, **58**, 109–130.
- Wold, S., Trygg, J., Berglund, A. and Antti, H. (2001) Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.*, **58**, 131–150.
- Wolfenden, R., Andersson, L., Cullis, P.M. and Southgate, C.C.B. (1981) Affinities of amino acid side chains for solvent water. *Biochemistry*, **20**, 849–855.
- Wolohan, P. and Reichert, D.E. (2003) CoMFA and docking study of novel estrogen receptor subtype selective ligands. *J. Comput. Aid. Mol. Des.*, **17**, 313–328.
- Wolohan, P., Yoo, J., Welch, M.J. and Reichert, D.E. (2005) QSAR studies of copper azamacrocycles and thiosemicarbazones: MM3 parameter development and prediction of biological properties. *J. Med. Chem.*, **48**, 5561–5569.
- Wolohan, P.R.N. and Clark, R.D. (2003) Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA. *J. Comput. Aid. Mol. Des.*, **17**, 65–76.
- Woolfrey, J.R., Avery, M.A. and Doweyko, A.M. (1998) Comparison of 3D quantitative structure–activity relationship methods: analysis of the *in vitro* antimalarial activity of 154 artemisinin analogues by hypothetical active-site lattice and comparative molecular field analysis. *J. Comput. Aid. Mol. Des.*, **12**, 165–181.
- Woolley, R.G. (1978a) Further remarks on molecular structure in quantum theory. *Chem. Phys. Lett.*, **55**, 443–446.
- Woolley, R.G. (1978b) Must a molecule have a shape? *J. Am. Chem. Soc.*, **100**, 1073–1078.
- Wootton, R., Cranfield, R., Sheppey, G.C. and Goodford, P.J. (1975) Physico-chemical–activity relationships in practice. 2. Rational selection of benzenoid substituents. *J. Med. Chem.*, **18**, 607–613.
- Worrall, F. and Thomsen, M. (2004) Quantum vs. topological descriptors in the development of molecular models of groundwater pollution by pesticides. *Chemosphere*, **54**, 585–596.
- Worth, A.P., Bassan, A., Fabjan, E., Gallegos Saliner, A., Netzeva, T.I., Patlewicz, G., Pavan, M. and Tsakovska, I. (2008) The use of computational methods in the grouping and assessment of chemicals – preliminary investigations. European Technical Report Eur N. 22941 EN.
- Worth, A.P. and Cronin, M.T.D. (1999) Embedded cluster modelling – A novel method for analysing embedded data sets. *Quant. Struct. -Act. Relat.*, **18**, 229–235.
- Worth, A.P. and Cronin, M.T.D. (2003) The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *J. Mol. Struct. (Theochem)*, **622**, 97–111.
- Wu, J. and Aluko, R.E. (2007) Quantitative structure–activity relationship study of bitter di- and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity. *J. Pestic. Sci.*, **13**, 63–69.
- Wu, W., Walczak, B., Massart, D.L., Heuerding, S., Erni, F., Last, I.R. and Prebble, K.A. (1996) Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemom. Intell. Lab. Syst.*, **33**, 35–46.
- Wypych, G. (ed.) (2001) *Handbook of Solvents*, ChemTec Publishing, Toronto (Canada), p. 1675.
- Xiang, Y.H., Liu, M., Zhang, X., Zhang, R. and Hu, Z. (2002) Quantitative prediction of liquid chromatography retention of N-benzylideneanilines based on quantum chemical parameters and radial basis function neural network. *J. Chem. Inf. Comput. Sci.*, **42**, 592–597.
- Xiao, W. (2004) Relations between resistance and Laplacian matrices. *MATCH Commun. Math. Comput. Chem.*, **51**, 119–127.
- Xiao, W. and Gutman, I. (2003) On resistance matrices. *MATCH Commun. Math. Comput. Chem.*, **49**, 67–81.
- Xiao, Y.-D., Qiao, Y., Zhang, J., Lin, S. and Zhang, W. (1997) A method for substructure search by atom-centered multilayer code. *J. Chem. Inf. Comput. Sci.*, **37**, 701–704.

- Xiao, Z., Varma, S., Xiao, Y.-D. and Tropsha, A. (2004) Modeling of p38 mitogen-activated protein kinase inhibitors using the CatalystTM HypoGen and *k*-nearest neighbor QSAR methods. *J. Mol. Graph. Model.*, **23**, 129–138.
- Xiao, Z., Xiao, Y.-D., Feng, J., Golbraikh, A., Tropsha, A. and Lee, K.-H. (2002) Antitumor agents. 213. Modeling of epipodophyllotoxin derivatives using variable selection *k* nearest neighbor QSAR method. *J. Med. Chem.*, **45**, 2294–2309.
- Xie, H.-P., Jiang, J.-H., Cui, H., Shen, G.-L. and Yu, R.-Q. (2002) A new redundant variable pruning approach – minor latent variable perturbation – PLS used for QSAR studies on anti-HIV drugs. *Computers Chem.*, **26**, 591–600.
- Xie, Q., Sun, H., Xie, G. and Zhou, J. (1995) An iterative method for calculation of group electronegativities. *J. Chem. Inf. Comput. Sci.*, **35**, 106–109.
- Xing, L. and Glen, R.C. (2002) Novel methods for the prediction of log *P*, p*K*_a, and log *D*. *J. Chem. Inf. Comput. Sci.*, **42**, 796–805.
- Xing, L., Glen, R.C. and Clark, R.D. (2003) Predicting p*K*_a by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.*, **43**, 870–879.
- Xu, J., Guo, B., Chen, B. and Zhang, Q. (2005) A QSPR treatment for the thermal stabilities of second-order NLO chromophore molecules. *J. Mol. Model.*, **12**, 65–75.
- Xu, J., Liu, L., Xu, W., Zhao, S. and Zuo, D. (2007) A general QSPR model for the prediction of θ (lower critical solution temperature) in polymer solutions with topological indices. *J. Mol. Graph. Model.*, **26**, 352–359.
- Xu, J. (1996) GMA: a generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. *J. Chem. Inf. Comput. Sci.*, **36**, 25–34.
- Xu, J. (2003) Two-dimensional structure and substructure searching, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH Weinheim, Germany, pp. 868–884.
- Xu, J. and Hagler, A. (2002) Chemoinformatics and drug discovery. *Molecules*, **7**, 566–600.
- Xu, J. and Stevenson, J. (2000) Drug-like index: a new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.*, **40**, 1177–1187.
- Xu, L. (1992) Molecular topological index a_N and its extension. *J. Serb. Chem. Soc.*, **57**, 485–495.
- Xu, L., Ball, J., Dixon, S.L. and Jurs, P.C. (1994) Quantitative structure–activity relationships for toxicity of phenols using regressions analysis and computational networks. *Environ. Toxicol. Chem.*, **13**, 841–851.
- Xu, L., Wang, H.-Y. and Su, Q. (1992a) A newly proposed molecular topological index for the discrimination of *cis/trans* isomers and for the studies of QSAR/QSPR. *Computers Chem.*, **16**, 187–194.
- Xu, L., Wang, H.-Y. and Su, Q. (1992b) Correlation analysis in structure and chromatographic data of organophosphorus compounds by GAI. *Computers Chem.*, **16**, 195–199.
- Xu, L., Yang, J.-A. and Wu, Y.-P. (2002) Effective descriptions of molecular structures and the quantitative structure–activity relationship studies. *J. Chem. Inf. Comput. Sci.*, **42**, 602–606.
- Xu, L., Yao, Y.-Y. and Wang, H.-M. (1995) New topological index and prediction of phase transfer energy for protonated amines and tetraalkylamines ions. *J. Chem. Inf. Comput. Sci.*, **35**, 45–49.
- Xu, L., Zhang, Q.-Y., Wang, J. and Dong, L. (2006) Extended topological indices and prediction of activities of chiral compounds. *Chemom. Intell. Lab. Syst.*, **82**, 37–43.
- Xu, M., Zhang, A., Han, S. and Wang, L.-S. (2002) Studies of 3D-quantitative structure–activity relationships on a set of nitroaromatic compounds: CoMFA, advanced CoMFA and CoMSIA. *Chemosphere*, **48**, 707–715.
- Xu, Q.-S., Massart, D.L., Liang, Y.-Z. and Fang, K.-T. (2003) Two-step multivariate adaptive regression splines for modeling a quantitative relationship between gas chromatography retention indices and molecular descriptors. *J. Chromat.*, **998**, 155–167.
- Xu, S. and Nirmalakhandan, N.N. (1998) Use of QSAR models in predicting joint effects in multi-component mixtures of organic chemicals. *Water Res.*, **32**, 2391–2399.
- Xu, Y.-J. and Gao, H. (2003) Dimension related distance and its application in QSAR/QSPR model error estimation. *QSAR Comb. Sci.*, **22**, 422–429.
- Xu, Y.-J. and Johnson, M. (2002) Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.*, **42**, 912–926.
- Xu, Y. and Brereton, R.G. (2005) A comparative study of cluster validation indices applied to genotyping data. *Chemom. Intell. Lab. Syst.*, **78**, 30–40.
- Xue, C., Zhang, R., Liu, H., Yao, X.-J., Liu, M., Hu, Z. and Fan, B.T. (2004) QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. *J. Chem. Inf. Comput. Sci.*, **44**, 1693–1700.

- Xue, L. and Bajorath, J. (2000) Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.*, **40**, 801–809.
- Xue, L. and Bajorath, J. (2002) Accurate partitioning of compounds belonging to diverse activity classes. *J. Chem. Inf. Comput. Sci.*, **42**, 757–764.
- Xue, L., Godden, J.W. and Bajorath, J. (1999a) Database searching for compounds with similar biological activity using short binary string representations of molecules. *J. Chem. Inf. Comput. Sci.*, **39**, 881–886.
- Xue, L., Godden, J.W. and Bajorath, J. (2000) Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.*, **40**, 1227–1234.
- Xue, L., Godden, J.W. and Bajorath, J. (2003a) Mini-fingerprints for virtual screening: design principles and generation of novel prototypes based on information theory. *SAR & QSAR Environ. Res.*, **14**, 27–40.
- Xue, L., Godden, J.W., Gao, H. and Bajorath, J. (1999b) Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.*, **39**, 699–704.
- Xue, L., Godden, J.W., Stahura, F.L. and Bajorath, J. (2003b) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.*, **43**, 1151–1157.
- Xue, L., Godden, J.W., Stahura, F.L. and Bajorath, J. (2003c) Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.*, **43**, 1218–1225.
- Xue, L., Stahura, F.L., Godden, J.W. and Bajorath, J. (2001a) Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.*, **41**, 746–753.
- Xue, L., Stahura, F.L., Godden, J.W. and Bajorath, J. (2001b) Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci.*, **41**, 394–401.
- Xue, Y., Li, Z.R., Yap, C.W., Sun, L.Z., Chen, X. and Chen, Y.Z. (2004) Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.*, **44**, 1630–1638.
- Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A. and Giralt, F. (2001) A fuzzy ARTMAP based on quantitative structure–property relationships (QSPRs) for predicting aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.*, **41**, 1177–1207.
- Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A. and Giralt, F. (2002) Fuzzy ARTMAP and back-propagation neural networks based quantitative structure–property relationships (QSPRs) for octanol–water partition coefficient of organic compounds. *J. Chem. Inf. Comput. Sci.*, **42**, 162–183.
- Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A. and Giralt, F. (2003) A Fuzzy ARTMAP-based quantitative structure–property relationship (QSPR) for the Henry's law constant of organic compounds. *J. Chem. Inf. Comput. Sci.*, **43**, 85–112.
- Yalkowsky, S.H. (1999) *Solubility and Solubilization in Aqueous Media*, Oxford University Press, New York, p. 480.
- Yalkowsky, S.H., Dannenfelser, R.-M., Myrdal, P. and Simamora, P. (1994) Unified physical property estimation relationships (UPPER). *Chemosphere*, **28**, 1657–1673.
- Yalkowsky, S.H., Johnson, J.L.H., Snaghvi, T. and Machatha, S.G. (2006) A 'rule of unity' for human intestinal absorption. *Pharm. Res.*, **23**, 2475–2481.
- Yalkowsky, S.H., Myrdal, P., Dannenfelser, R.-M. and Simamora, P. (1994) UPPER II: calculation of physical properties of the chlorobenzenes. *Chemosphere*, **28**, 1675–1688.
- Yalkowsky, S.H. and Valvani, S.C. (1979) Solubilities and partitioning. 2. Relationships between aqueous solubilities, partition coefficients, and molecular surface areas of rigid aromatic hydrocarbons. *J. Chem. Eng. Data*, **24**, 127–129.
- Yamagami, C., Kawase, K. and Fujita, T. (1999) Hydrophobicity parameters determined by reversed-phase liquid chromatography. XIII. A new hydrogen-accepting scale of monosubstituted (di)azines for the relationship between retention factor and octanol–water partition coefficient. *Quant. Struct. -Act. Relat.*, **18**, 26–34.
- Yamamoto, Y. and Otsu, T. (1967) Effects of substituents in radical reactions: extension of the Hammett equation. *Chem. & Ind.*, 787–789.
- Yamashita, F., Fujiwara, S. and Hashida, M. (2002) The "latent membrane permeability" concept: QSPR analysis of inter/intralaboratorically variable Caco-2 permeability. *J. Chem. Inf. Comput. Sci.*, **42**, 408–413.
- Yamashita, F., Itoh, T., Hara, H. and Hashida, M. (2006) Visualization of large-scale aqueous

- solubility data using a novel hierarchical data visualization technique. *J. Chem. Inf. Model.*, **46**, 1054–1059.
- Yan, A. and Gasteiger, J. (2003) Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.*, **43**, 429–434.
- Yan, S.T., Wang, J.S., Niknejad, A., Lu, C.X., Jin, N. and Ho, Y.K. (2003) DNA sequence representation without degeneracy. *Nucleic Acids Res.*, **31**, 3078–3080.
- Yan, W. and Yeh, Y.-N. (2006) Connections between Wiener index and matchings. *J. Math. Chem.*, **39**, 389–399.
- Yang, C. and Zhong, C. (2003) Modified connectivity indices and their application to QSPR study. *J. Chem. Inf. Comput. Sci.*, **43**, 1998–2004.
- Yang, F., Wang, Z.-D. and Huang, Y.-P. (2003a) Modification of the Wiener index. 2. *J. Chem. Inf. Comput. Sci.*, **43**, 1337–1341.
- Yang, F., Wang, Z.-D., Huang, Y.-P. and Ding, X.-R. (2003b) Modification of Wiener index and its application. *J. Chem. Inf. Comput. Sci.*, **43**, 753–756.
- Yang, G.-Z., Lien, E.J. and Guo, Z.-R. (1986) Physical factors contributing to hydrophobic constant π . *Quant. Struct. -Act. Relat.*, **5**, 12–18.
- Yang, J.A. and Kiang, Y.-S. (1983) *Acta Chim. Sin.*, **41**, 884.
- Yang, P. (1992) *Distribution and Physical Property in Molecule*, Union Press of Shanxi University, Tai Yuan, China.
- Yang, S., Lu, W., Chen, N. and Hu, Q.-N. (2005) Support vector regression based QSPR for the prediction of some physico-chemical properties of alkyl benzenes. *J. Mol. Struct. (Theochem)*, **719**, 119–127.
- Yang, S., Bumgarner, J.G., Kruk, L.F.R. and Khaledi, M.G. (1996) Quantitative structure–activity relationships studies with micellar electrokinetic chromatography. Influence of surfactant type and mixed micelles on estimation of hydrophobicity and bioavailability. *J. Chromat.*, **721**, 323–335.
- Yang, W. and Mortier, W.J. (1986) The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines. *J. Am. Chem. Soc.*, **108**, 5708–5711.
- Yang, Y., Lin, J. and Wang, C. (2002) Small regular graphs having the same path layer matrix. *J. Graph Theory*, **39**, 219–221.
- Yang, Y.-Q., Xu, L. and Hu, C.-Y. (1994) Extended adjacency matrix indices and their applications. *J. Chem. Inf. Comput. Sci.*, **34**, 1140–1145.
- Yao, X.-J., Fan, B.T., Doucet, J.P., Panaye, A., Liu, M., Zhang, R., Zhang, X. and Hu, Z. (2003) Quantitative structure–property relationship models for the prediction of liquid heat capacity. *QSAR Comb. Sci.*, **22**, 29–48.
- Yao, X.-J., Wang, Y., Zhang, X., Zhang, R., Liu, M., Hu, Z. and Fan, B.T. (2002) Radial basis function neural network-based QSPR for the prediction of critical temperature. *Chemom. Intell. Lab. Syst.*, **62**, 217–225.
- Yao, X.-J., Zhang, X., Zhang, R., Liu, M., Hu, Z. and Fan, B.T. (2001) Prediction of enthalpy of alkanes by the use of radial basis function neural networks. *Computers Chem.*, **25**, 475–482.
- Yao, X.-J., Zhang, X., Zhang, R., Liu, M., Hu, Z. and Fan, B.T. (2002) Radial basis function neural network based QSPR for the prediction of critical pressure of substituted benzenes. *Computers Chem.*, **26**, 159–169.
- Yao, Y.-Y., Xu, L., Yang, Y.-Q. and Yuan, X.-S. (1993a) Study on structure–activity relationships of organic compounds: three new topological indices and their applications. *J. Chem. Inf. Comput. Sci.*, **33**, 590–594.
- Yao, Y.Y., Xu, L. and Yuan, X.-S. (1993b) A new topological index for research on structure–property relationships of alkane. *Acta Chim. Sin.*, **51**, 463–469.
- Yap, C.W. and Chen, Y.Z. (2005a) Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.*, **45**, 982–992.
- Yap, C.W. and Chen, Y.Z. (2005b) Quantitative structure–pharmacokinetic relationships for drug distribution properties by using general regression neural network. *J. Pharm. Sci.*, **94**, 153–168.
- Yap, C.W., Li, Z.R. and Chen, Y.Z. (2006) Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods. *J. Mol. Graph. Model.*, **24**, 383–395.
- Yasri, A. and Hartsough, D. (2001) Toward an optimal procedure for variable selection and QSAR model building. *J. Chem. Inf. Comput. Sci.*, **41**, 1218–1227.
- Yen, T.E., Agatonovic-Kustrin, S., Evans, A.M., Nation, R.L. and Ryand, J. (2005) Prediction of drug absorption based on immobilized artificial membrane (IAM) chromatography separation and calculated molecular descriptors. *J. Pharm. Biomed. Anal.*, **38**, 472–478.
- Yin, S.-W., Shuai, Z. and Wang, Y. (2003) A quantitative structure–property relationship study of the glass transition temperature of OLED materials. *J. Chem. Inf. Comput. Sci.*, **43**, 970–977.
- Yiyu, C., Minjun, C. and Welsh, W.J. (2003) Fractal fingerprinting of chromatographic profiles based on wavelet analysis and its application to

- characterize the quality grade of medicinal herbs. *J. Chem. Inf. Comput. Sci.*, **43**, 1959–1965.
- Yokono, S., Shieh, D.D., Goto, H. and Arakawa, K. (1982) Hydrogen bonding and anesthetic potency. *J. Med. Chem.*, **25**, 873–876.
- Yoneda, Y. (1979) An estimation of the thermodynamic properties of organic compounds in the ideal gas state. I. Acyclic compounds and cyclic compounds with a ring of cyclopentane, cyclohexane, benzene or naphthalene. *Bull. Chem. Soc. Jap.*, **52**, 1297–1314.
- Yoshida, F. and Topliss, J.G. (2000) QSAR model for drug human oral bioavailability. *J. Med. Chem.*, **43**, 2575–2585.
- Young, S.S. (2003) Design of diverse and focused combinatorial libraries using an alternating algorithm. *J. Chem. Inf. Comput. Sci.*, **43**, 1916–1921.
- Young, S.S., Gombar, V.K., Emptage, M.R., Cariello, N.F. and Lambert, C. (2002) Mixture deconvolution and analysis of Ames mutagenicity data. *Chemom. Intell. Lab. Syst.*, **60**, 5–11.
- Young, S.S. and Hawkins, D.M. (1995) Analysis of a 2⁹ full factorial chemical library. *J. Med. Chem.*, **38**, 2784–2788.
- Young, S.S. and Hawkins, D.M. (1998) Using recursive partitioning to analyze a large SAR data set. *SAR & QSAR Environ. Res.*, **8**, 183–193.
- Young, S.S., Profeta, S., Unwalla, R.J. and Kosh, J.W. (1997) Exploratory analysis of chemical structure, bacterial mutagenicity and rodent tumorigenicity. *Chemom. Intell. Lab. Syst.*, **37**, 115–124.
- Yu, A., Lu, M. and Tian, F. (2005) New upper bounds for the energy of graphs. *MATCH Commun. Math. Comput. Chem.*, **53**, 441–448.
- Yu, S.J., Keenan, S.M., Tong, W. and Welsh, W.J. (2002) Influence of the structural diversity of data sets on the statistical quality of three-dimensional quantitative structure–activity relationship (3D-QSAR) models: predicting the estrogenic activity of xenoestrogens. *Chem. Res. Toxicol.*, **15**, 1229–1234.
- Yu, X., Wang, X., Wang, H., Li, X. and Gao, J. (2006) Prediction of solubility parameters for polymers by a QSPR model. *QSAR Comb. Sci.*, **25**, 156–161.
- Yu, X., Yi, B., Xie, Z., Wang, X. and Liu, F. (2007) Prediction of the conformational property for polymers using quantum chemical descriptors. *Chemom. Intell. Lab. Syst.*, **87**, 247–251.
- Yuan, C.X., Liao, B. and Wang, T. (2003) New 3-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.*, **379**, 412–417.
- Yuan, H. and Parrill, A.L. (2002) QSAR studies of HIV-1 integrase inhibition. *Bioorg. Med. Chem.*, **10**, 4169–4183.
- Yuan, H. and Cao, C. (2003) Topological indices based on vertex, edge, ring, and distance: application to various physico-chemical properties of diverse hydrocarbons. *J. Chem. Inf. Comput. Sci.*, **43**, 501–512.
- Yukawa, Y. and Tsuno, Y. (1959) Resonance effect in Hammett relation. III. The modified Hammett relation for electrophilic reactions. *Bull. Chem. Soc. Jap.*, **32**, 971–981.
- Yukawa, Y., Tsuno, Y. and Sawada, M. (1966) Resonance effect in Hammett relation. IV. Linear free energy based on the normal substituent constants. *Bull. Chem. Soc. Jap.*, **39**, 2274–2286.
- Yukawa, Y., Tsuno, Y. and Sawada, M. (1972a) The substituent effect. I. Normal substituent constants from the hydrolysis of substituted-benzyl benzoates. *Bull. Chem. Soc. Jap.*, **45**, 1198–1205.
- Yukawa, Y., Tsuno, Y. and Sawada, M. (1972b) The substituent effect. III. The basicities of polynuclear aryl methyl ketones. *Bull. Chem. Soc. Jap.*, **45**, 1210–1216.
- Zabrodsky, H. and Avnir, D. (1995) Continuous symmetry measures. 4. Chirality. *J. Am. Chem. Soc.*, **117**, 462–473.
- Zakarya, D., Belkhadir, M. and Fkih-Tetouani, S. (1993) Quantitative structure–biodegradability relationships (QSBRs) using modified autocorrelation method (MAM). *SAR & QSAR Environ. Res.*, **1**, 21–27.
- Zakarya, D. and Chastrette, M. (1998) Contribution of structure–odor relationships to the elucidation of the origin of musk fragrance activity, in *Comparative QSAR* (ed. J. Devillers), Taylor & Francis, Washington, DC, pp. 169–195.
- Zakarya, D., Larfaoui, E.M., Boulaamail, A., Tollabi, M. and Lakhlifi, T. (1998) QSARs for a series of inhibitory anilides. *Chemosphere*, **36**, 2809–2818.
- Zakarya, D., Nohair, M. and Nyassi, H. (2000) On the DZ^{kP} molecular descriptors. *Lab. Rob. Autom.*, **12**, 37–40.
- Zakarya, D., Tiyal, F. and Chastrette, M. (1993) Use of the multifunctional autocorrelation method to estimate molar volumes of alkanes and oxygenated compounds comparison between components of autocorrelation vectors and topological indexes. *J. Phys. Org. Chem.*, **6**, 574–582.
- Zaliani, A. and Gancia, E. (1999) MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.*, **39**, 525–533.
- Zamora, I., Oprea, T.I., Cruciani, G., Pastor, M. and Ungell, A.-L. (2003) Surface descriptors for protein–ligand affinity prediction. *J. Med. Chem.*, **46**, 25–33.

- Zamora, I., Oprea, T.I. and Ungell, A.-L. (2001) Prediction of oral drug permeability, in *Rational Approaches to Drug Design* (eds H.-D. Höltje and W. Sippl), Prous Science, Barcelona (Spain), pp. 271–280.
- Zar, J.H. (1984) *Biostatistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ, p. 718.
- Zass, E. (2003) Databases of chemical reactions, in *Handbook of Chemoinformatics*, Vol. 2 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 667–699.
- Zauhar, R.J., Moyna, G., Tian, L.F., Li, Z.-J. and Welsh, W.J. (2003) Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.*, **46**, 5674–5690.
- Zefirov, N., Palyulin, V.A., Skvortsova, M.I. and Baskin, I.I. (1995) Inverse problems in QSAR, in *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications* (eds F. Sanz, J. Giraldo and F. Manaut), Prous Science, Barcelona, Spain, pp. 40–41.
- Zefirov, N.S., Kirpichenok, M.A., Izmailov, F.F. and Trofimov, M.I. (1987) Scheme for the calculation of the electronegativities of atoms in a molecule in the framework of Sanderson's principle. *Dokl. Akad. Nauk. SSSR*, **296**, 883–887.
- Zefirov, N.S. and Palyulin, V.A. (2001) QSAR for boiling points of "small" sulfides. Are the "high-quality structure–property–activity regressions" the real high quality QSAR models? *J. Chem. Inf. Comput. Sci.*, **41**, 1022–1027.
- Zefirov, N.S. and Palyulin, V.A. (2002) Fragmental approach in QSPR. *J. Chem. Inf. Comput. Sci.*, **42**, 1112–1122.
- Zefirov, N.S., Palyulin, V.A. and Radchenko, E.V. (1991) Problem of generation of structures with specified properties. Solution of the inverse problem for Balaban centric index. *Dokl. Akad. Nauk. SSSR*, **316**, 921–924.
- Zefirov, N.S., Palyulin, V.A. and Radchenko, E.V. (1997) Molecular field topology analysis in studies of quantitative structure–activity relationships for organic compounds. *Dokl. Akad. Nauk. SSSR*, **352**, 23–26.
- Zefirov, N.S. and Tratch, S.S. (1997) Some notes on Randić–Razinger's approach to characterization of molecular shape. *J. Chem. Inf. Comput. Sci.*, **37**, 900–912.
- Zenkevich, I.G. (1998) Application of molecular dynamics to chromatographic–spectral identification of isomeric products of organic reactions. *Russ. J. Org. Chem.*, **34**, 1403–1409.
- Zenkevich, I.G. (1999) New applications of the retention index concept in gas and high performance liquid chromatography. *Fresen. J. Anal. Chem.*, **365**, 305–309.
- Zernov, V.V., Balakin, K.V., Ivashchenko, A.A., Savchuk, N.P. and Pletnev, I.V. (2003) Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.*, **43**, 2048–2056.
- Zhai, H.L., Chen, X. and Hu, Z. (2006) A new approach for the identification of important variables. *Chemom. Intell. Lab. Syst.*, **80**, 130–135.
- Zhai, Z., Wang, Z.-Y. and Chen, S.-D. (2006) Quantitative structure–retention relationship for gas chromatography of polychlorinated naphthalenes by *ab initio* quantum mechanical calculations and a Cl substitution position method. *QSAR Comb. Sci.*, **25**, 7–14.
- Zhai, Z., Wang, Z.-Y. and Wang, L.-S. (2005) Quantitative structure–property relationship study of GC retention indices for PCDFs by DFT and relative position of chlorine substitution. *J. Mol. Struct. (Theochem)*, **724**, 115–124.
- Zhang, H., Qu, X. and Ando, H. (2005) A simple method for reaction rate prediction of ester hydrolysis. *J. Mol. Struct. (Theochem)*, **725**, 31–37.
- Zhang, J., Kleinöder, T. and Gasteiger, J. (2006) Prediction of pK_a values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *J. Chem. Inf. Model.*, **46**, 2256–2266.
- Zhang, J., Aizawa, M., Amari, S., Iwasawa, Y., Nakano, T. and Nakata, K. (2004) Development of KiBank, a database supporting structure-based drug design. *Comp. Biol. Chem.*, **28**, 401–407.
- Zhang, K. and Shasha, D. (1989) Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal of Computing*, **18**, 1245–1262.
- Zhang, Q.-Y. and Aires-de-Sousa, J. (2005) Structure-based classification of chemical reactions without assignment of reaction centers. *J. Chem. Inf. Model.*, **45**, 1775–1783.
- Zhang, Q.-Y. and Aires-de-Sousa, J. (2007) Random forest prediction of mutagenicity from empirical physico-chemical descriptors. *J. Chem. Inf. Model.*, **47**, 1–8.
- Zhang, R., Liu, S., Liu, M. and Hu, Z. (1997) Neural network-molecular descriptors approach to the prediction of properties of alkenes. *Computers Chem.*, **21**, 335–341.
- Zhang, S., Golbraikh, A., Oloff, S., Kohn, H. and Tropsha, A. (2007) A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of

- chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.*, **46**, 1984–1995.
- Zhang, S., Golbraikh, A. and Tropsha, A. (2006) Development of quantitative structure–binding affinity relationship models based on novel geometrical chemical descriptors of the protein–ligand interfaces. *J. Med. Chem.*, **49**, 2713–2724.
- Zhang, T.-L., Ding, Y.-S. and Chou, K.-C. (2008) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. *J. Theor. Biol.*, **250**, 186–193.
- Zhang, X., Luo, J. and Yang, L. (2007) New invariant of DNA sequence based on 3DD-curves and its application on phylogeny. *J. Comput. Chem.*, **28**, 2342–2346.
- Zhang, Y. (1982a) Electronegativities of elements in valence states and their applications. 1. Electronegativities of elements in valence states. *Inorg. Chem.*, **21**, 3886–3889.
- Zhang, Y. (1982b) Electronegativities of elements in valence states and their applications. 2. A scale of strengths of Lewis acids. *Inorg. Chem.*, **21**, 3889–3893.
- Zhang, Y. (2007) On 2D graphical representation of RNA secondary structure. *MATCH Commun. Math. Comput. Chem.*, **57**, 697–710.
- Zhang, Y. and Chen, W. (2006) Invariants of DNA sequences based on 2DD-curves. *J. Theor. Biol.*, **242**, 382–388.
- Zhang, Y., Liao, B. and Ding, K. (2005) On 2D graphical representation of DNA sequence of nondegeneracy. *Chem. Phys. Lett.*, **411**, 28–32.
- Zhang, Y., Liao, B. and Ding, K. (2006) On 3DD-curves of DNA sequences. *Mol. Simulat.*, **32**, 29–34.
- Zhao, C.Y., Zhang, H., Zhang, X., Liu, M., Hu, Z. and Fan, B.T. (2006) Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology*, **217**, 105–119.
- Zhao, H. and Liu, R. (2006) On the Merrifield–Simmons index of graphs. *MATCH Commun. Math. Comput. Chem.*, **56**, 617–624.
- Zhao, H., Chen, J., Quan, X., Yang, F. and Peijnenburg, W.J.G.M. (2001) Quantitative structure–property relationship study on reductive dehalogenation of selected halogenated aliphatic hydrocarbons in sediment slurries. *Chemosphere*, **44**, 1557–1563.
- Zhao, L. and Yalkowsky, S.H. (1999) A combined group contribution and molecular geometry approach for predicting melting points of aliphatic compounds. *Ind. Eng. Chem. Res.*, **38**, 3581–3584.
- Zhao, W.-N., Yu, Q.-S., Zou, J.-W., Ma, M. and Zheng, K.-W. (2005) Three-dimensional quantitative structure–activity relationship study for analogues of TQXs using CoMFA and CoMSIA. *J. Mol. Struct. (Theochem)*, **723**, 69–78.
- Zhao, Y.H., Abraham, M.H. and Zissimos, A.M. (2003a) Determination of McGowan volumes for ions and correlation with van der Waals volumes. *J. Chem. Inf. Comput. Sci.*, **43**, 1848–1854.
- Zhao, Y.H., Abraham, M.H. and Zissimos, A.M. (2003b) Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. *J. Org. Chem.*, **68**, 7368–7373.
- Zhao, Y.-H., Le, J., Abraham, M.H., Hersey, A., Eddershaw, P.J., Luscombe, C.N., Boutina, D., Beck, G., Sherborne, B., Cooper, I. and Platts, J.A. (2001) Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure–activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.*, **90**, 749–784.
- Zhao, Y.-H., Yuan, X., Ji, G.-D. and Sheng, L.-X. (1997) Quantitative structure–activity relationships of nitroaromatic compounds to four aquatic organisms. *Chemosphere*, **34**, 1837–1844.
- Zheng, F., Bayram, E., Sumithran, S.P., Ayers, J.T., Zhan, C.-G., Schmitt, J.D., Dwoskin, L.P. and Crooks, P.A. (2006) QSAR modeling of mono- and bis-quaternary ammonium salts that act as agonists at neuronal nicotinic acetylcholine receptors mediating dopamine release. *Bioorg. Med. Chem.*, **14**, 3017–3037.
- Zheng, F., Zheng, G., Deaciuc, A.G., Zhan, C.-G., Dwoskin, L.P. and Crooks, P.A. (2007) Computational neural network analysis of the affinity of lobeline and tetrabenazine analogs for the vesicular monoamine transporter-2. *Bioorg. Med. Chem.*, **15**, 2975–2992.
- Zheng, S., Luo, X., Chen, G., Zhu, W., Shen, J., Chen, K. and Jiang, H. (2005) A new rapid and effective chemistry space filter in recognizing a druglike database. *J. Chem. Inf. Model.*, **45**, 856–862.
- Zheng, W., Cho, S.J. and Tropsha, A. (1998) Rational combinatorial library design. 1. Focus-2D: a new approach to the design of targeted combinatorial chemical libraries. *J. Chem. Inf. Comput. Sci.*, **38**, 251–258.
- Zheng, W. and Tropsha, A. (2000) Novel variable selection quantitative structure–property relationship approach based on the *k*-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.*, **40**, 185–194.
- Zhokova, N.I., Palyulin, V.A., Baskin, I.I., Zefirov, A. N. and Zefirov, N.S. (2007) Fragment descriptors in the QSPR method: their use for calculating the enthalpies of vaporization of organic substances. *Russ. J. Phys. Chem.*, **81**, 9–12.

- Zhong, C., He, J., Xia, Z. and Li, Y. (2004) Modeling of activity of efavirenz with the mutant of HIV reverse transcriptase using variable connectivity indices. *QSAR Comb. Sci.*, **23**, 650–654.
- Zhong, C., Yang, C. and Li, Q. (2002) Correlation of Henry's constants of nonpolar and polar solutes in molten polymers using connectivity indices. *Ind. Eng. Chem. Res.*, **41**, 2826–2833.
- Zhou, B. (2004a) Energy of a graph. *MATCH Commun. Math. Comput. Chem.*, **51**, 111–118.
- Zhou, B. (2004b) Zagreb indices. *MATCH Commun. Math. Comput. Chem.*, **52**, 113–118.
- Zhou, B. (2007) Remarks on Zagreb indices. *MATCH Commun. Math. Comput. Chem.*, **57**, 591–596.
- Zhou, B. and Gutman, I. (2004a) Estimating the modified Hosoya index. *MATCH Commun. Math. Comput. Chem.*, **52**, 183–192.
- Zhou, B. and Gutman, I. (2004b) Relations between Wiener, hyper-Wiener and Zagreb indices. *Chem. Phys. Lett.*, **394**, 93–95.
- Zhou, B. and Gutman, I. (2005) Further properties of Zagreb indices. *MATCH Commun. Math. Comput. Chem.*, **54**, 233–239.
- Zhou, B. and Gutman, I. (2007) On Laplacian energy of graphs. *MATCH Commun. Math. Comput. Chem.*, **57**, 211–220.
- Zhou, B., Gutman, I., De La Peña, J.A., Rada, J. and Mendoza, L. (2007) On spectral moments and energy graphs. *MATCH Commun. Math. Comput. Chem.*, **57**, 183–191.
- Zhou, B. and Stevanović, D. (2006) A note on Zagreb indices. *MATCH Commun. Math. Comput. Chem.*, **56**, 571–578.
- Zhou, C., Nie, C., Li, S. and Li, Z. (2007) A novel semi-empirical topological descriptor Nt and the application to study on QSPR/QSAR. *J. Comput. Chem.*, **28**, 2413–2423.
- Zhou, J., Xie, Q., Sun, D.M., Xie, G., Cao, L.X. and Xu, Z. (1993) Structure–activity relationships on pesticides: a development in methodology and its software system. *J. Chem. Inf. Comput. Sci.*, **33**, 310–319.
- Zhou, P., Zhou, Y., Wu, S., Li, B., Tian, F. and Li, Z. (2006) A new descriptor of amino acids based on the three-dimensional vector of atomic interaction fields. *Chinese Sci. Bull.*, **51**, 524–529.
- Zhou, Y.-X., Xu, L., Wu, Y.-P. and Liu, B.-L. (1999) A QSAR study of the antiallergic activities of substituted benzamides and their structures. *Chemom. Intell. Lab. Syst.*, **45**, 95–100.
- Zhou, Z., Dai, Q. and Gu, T. (2003) A QSAR model of PAHs carcinogenesis based on thermodynamic stabilities of biactive sites. *J. Chem. Inf. Comput. Sci.*, **43**, 615–621.
- Zhu, H.Y. and Klein, D.J. (1996) Graph-geometric invariants for molecular structures. *J. Chem. Inf. Comput. Sci.*, **36**, 1067–1075.
- Zhu, H.Y., Klein, D.J. and Lukovits, I. (1996) Extensions of the Wiener number. *J. Chem. Inf. Comput. Sci.*, **36**, 420–428.
- Zhu, H., Sedykh, A., Chakravarti, S.K. and Klopman, G. (2005) A new group contribution approach to the calculation of log P . *Curr. Comput. -Aided Drug Des.*, **1**, 3–9.
- Zhu, L., Hou, T.-J., Chen, L. and Xu, X. (2001) 3D QSAR analyses of novel tyrosine kinase inhibitors based on pharmacophore alignment. *J. Chem. Inf. Comput. Sci.*, **41**, 1032–1040.
- Zissimos, A.M., Abraham, M.H., Barker, M.C., Box, K.J. and Tam, K.Y. (2002a) Calculation of Abraham descriptors from solvent–water partition coefficients in four different systems: evaluation of different methods of calculation. *J. Chem. Soc. Perkin Trans. 2*, 470–477.
- Zissimos, A.M., Abraham, M.H., Du, C.M., Valko, K., Bevan, C., Reynolds, D., Wood, J. and Tam, K.Y. (2002b) Calculation of Abraham descriptors from experimental data from seven HPLC systems: evaluation of five different methods of calculation. *J. Chem. Soc. Perkin Trans. 2*, 2001–2010.
- Zissimos, A.M., Abraham, M.H., Klamt, A., Eckert, F. and Wood, J. (2002c) A comparison between the two general sets of linear free energy descriptors of Abraham and Klamt. *J. Chem. Inf. Comput. Sci.*, **42**, 1320–1331.
- Zou, J.-W., Zhao, W.-N., Shang, Z.-C., Huang, M.-L., Guo, M. and Yu, Q.-S. (2002) A quantitative structure–property relationship analysis of log P for disubstituted benzenes. *J. Phys. Chem. A*, **106**, 11550–11557.
- Zuegge, J., Fechner, U., Roche, O., Parrott, N.J., Engkvist, O. and Schneider, G. (2002) A fast virtual screening filter for cytochrome P450 3A4 inhibition liability of compound libraries. *Quant. Struct. -Act. Relat.*, **21**, 249–256.
- Zupan, J. (2002) 2D mapping of large quantities of multi-variate data. *Croat. Chem. Acta*, **75**, 503–515.
- Zupan, J. (2003) Neural networks, in *Handbook of Chemoinformatics*, Vol. 3 (ed. J. Gasteiger), Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 1167–1215.
- Zupan, J. and Gasteiger, J. (1999) *Neural Networks for Chemistry and Drug Design*, Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 380.
- Zupan, J. and Novič, M. (1997) General type of a uniform and reversible representation of chemical structures. *Anal. Chim. Acta*, **348**, 409–418.

- Zupan, J., Novič, M. and Gasteiger, J. (1995) Neural networks with counter-propagation learning strategy used for modelling. *Chemom. Intell. Lab. Syst.*, **27**, 175–187.
- Zupan, J., Novič, M. and Ruisánchez, I. (1997) Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemom. Intell. Lab. Syst.*, **38**, 1–23.
- Zupan, J., Vračko, M. and Novič, M. (2000) New uniform and reversible representation of 3-D chemical structures. *Acta Chim. Sloven.*, **47**, 19–37.
- Zweerszeilmaker, W.M., Horbach, G.J. and Witkamp, R.F. (1997) Differential inhibitory effects of phenytoin, diclofenac, phenylbutazone and a series of sulfonamides on hepatic cytochrome P4502c activity *in vitro*, and correlation with some molecular descriptors in the dwarf goat (*Caprus hircus aegagrus*). *Xenobiotica*, **27**, 769–780.
- Zyrianov, Y. (2005) Distribution-based descriptors of the molecular shape. *J. Chem. Inf. Model.*, **45**, 657–672.
- Žerovnik, J. (1996) Computing the Szeged index. *Croat. Chem. Acta*, **69**, 837–843.
- Žerovnik, J. (1999) Szeged index of symmetric graphs. *J. Chem. Inf. Comput. Sci.*, **39**, 77–80.
- Žigert, P., Klavžar, S. and Gutman, I. (2000) Calculating the hyper-Wiener index of benzenoid hydrocarbons. *Acta Chim. Hung.-Mod. Chem.*, **137**, 83–94.
- Živković, T. (1990) On the evaluation of the characteristic polynomial of a chemical graph. *J. Comput. Chem.*, **11**, 217–222.
- Živković, T., Trinajstić, N. and Randić, M. (1975) On conjugated molecules with identical topological spectra. *Mol. Phys.*, **30**, 517–532.
- Župerl, Š., Pristovšek, P., Menart, V., Gaberc-Porekar, V. and Novič, M. (2007) Chemometric approach in quantification of structural identity/similarity of proteins in biopharmaceuticals. *J. Chem. Inf. Model.*, **47**, 737–743.

Appendix A.

Greek alphabets

Lower	Upper	Pronounce	Lower	Upper	Pronounce
α	A	Alpha	ν	N	Ni
β	B	Beta	ξ	Ξ	Xi
γ	Γ	Gamma	$\omicron$	O	Omikron
δ	Δ	Delta	π	Π	Pi
ϵ	E	Epsilon	ρ	ρ	Rho
ζ	Z	Zeta	σ	Σ	Sigma
η	H	Eta	τ	T	Tau
θ, ϑ	Θ	Theta	υ	Y	Upsilon
ι	I	Iota	ϕ	Φ	Phi
κ	K	Kappa	χ	X	Ki
λ	Λ	Lambda	φ	Ψ	Psi
μ	M	Mi	ω	Ω	Omega

Appendix B.

Acronyms

The most known acronyms used to define research fields, methods, statistical indices, and molecular descriptors are listed below, in alphabetic order. Acronyms beginning with numbers are at the end of the list.

AA	Augmented Atoms	CAIMAN	Classification And Influence Matrix ANalysis
AAA	Active Analog Approach		
AAC	Augmented Atom Codes	CAMD	Computer-Aided Molecular Design
AD	Applicability Domain		
ADME	Absorption, Distribution, Metabolism, Excretion properties	CAMM	Computer-Aided Molecular Modeling
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity properties	CART	Classification And Regression Trees
		CASE	Computer Automated Structure Evaluation
AIC	Akaike Information Content	CAST	CANonical representation of STereochemistry
AID	Atomic ID number		
AIM	Atom In Molecules	CATS	Chemically Advanced Template Search
ALOGP	Ghose–Crippen LOGP		
AMSP	Autocorrelation of Molecular Surface Properties	CEP	Conformational Ensemble Profile
ANN	Artificial Neural Networks	CFM	Compressed Feature Matrix
AP	Atom Pairs	CHEMICALC	Combined Handling of Estimation Methods Intended for Completely Automated Log P Calculation
ATS	Autocorrelation of a Topological Structure		
AWC	Atomic Walk Count		
BP-ANN	Back-Propagation Artificial Neural Networks	CIC	Complementary Information Content
BIC	Bonding Information Content	CID	Connectivity ID number
BID	Balaban ID number	CLOGP	Calculated LOGP
BLOGP	Bodor LOGP	CoMFA	Comparative Molecular Field Analysis
CADD	Computer-Aided Drug Design	CoMMA	Comparative Molecular Moment Analysis

CoMSA	Comparative Molecular Similarity Analysis	EEVA	Electronic EigenValue descriptors
CoMSIA	Comparative Molecular Similarity Indices Analysis	EFD	Electrophilic Frontier electron Density
CoRSA	Comparative Receptor Surface Analysis	EFVCI	External Factor Variable Connectivity Indices
CoSA	Comparative Spectra Analysis	ESD	Electrophilic SuperDelocalizability
COSV	Common Overlap Steric Volume	ER	Error Rate
CP-ANN	Counter-Propagation Artificial Neural Networks	ETA	Extended Topochemical Atom indices
CPK	Corey–Pauling–Koltun volume	EVA	EigenValue descriptors
CPSA	Charged Partial Surface Areas	FCFP	Functional Connectivity FingerPrints
CR	Continuum Regression	FEVA	First EigenValue Algorithm
CS	Column Sum	FLAP	Fingerprints for Ligands And Proteins
CSA	Cluster Significance Analysis	FPE	Final Prediction Error
CV	Cross-Validation	FRAU	Field-characterization for Reaction Analysis and Understanding
CWLIMG	Correlation Weights of the Local Invariants of Molecular Graphs	FW	Free–Wilson analysis
DA	Discriminant Analysis	GA	Genetic Algorithms
DAI	Distance-based Atom-type topological Index	GAI	General a_N -Index
DARC	Description, Acquisition, Retrieval Computer system	GAO	Graph of Atomic Orbitals
DD	Drug Design	GA-VSS	Genetic Algorithms–Variable Subset Selection
DFT	Density Functional Theory	GCM	Group Contribution Method
DFPS	Daylight-FingerPrint druglike Score	GCOD	Grid Cell Occupancy Descriptor
DG	Distance Geometry	GCSA	Generalized Cluster Significance Analysis
DiP	Distance Profiles descriptors	GERM	Genetically Evolved Receptor Models
DOS	Density Of States	GETAWAY	GEometric, Topological and Atomic Weighted Assembly descriptors
EA	Electronic Affinity	GFA	Genetic Function Approximation
EAID	Extended Adjacency ID number	GIPF	General Interaction Properties Function
EC	Extended Connectivity	GOLPE	Generating Optimal Linear PLS Estimations
ECA	Extended Connectivity Algorithm	GRIND	GRIld INdependent Descriptors
ECI	Electronic Charge Index		
ECFP	Extended Connectivity FingerPrints		

G-WHIM	Grid-Weighted Holistic Invariant Molecular descriptors	LDOS	Local Density Of States
HASL	Hypothetical Active Site Lattice	LFER	Linear Free Energy Relationships
HBA	Hydrogen Bond Acceptor	LHFG	Labeled Hydrogen-Filled Graphs
HBD	Hydrogen Bond Donor	LHSG	Labeled Hydrogen-Suppressed Graphs
HDG	Hydrogen Depleted Graph	LOEI	LOcal Edge Invariant
HFG	Hydrogen Filled Graph	LOGP	LOGarithm of the octanol–water Partition coefficient ($\log P$)
HQSAR	Hologram QSAR	LOMO	Lowest Occupied Molecular Orbital
H-QSAR	Hierarchical-QSAR	LOVI	LOcal Vertex Invariant
HFED	Hydration Free Energy Density	LS	Least Squares
HINT	Hydrophatic INteractions	LSER	Linear Solvation Energy Relationship
HOC	Hierarchically Ordered extended Connectivity	LUMO	Lowest Unoccupied Molecular Orbital
HOMO	Highest Occupied Molecular Orbital	MaP	Mapping Property distributions of molecular surfaces
HQSAR	Hologram Quantitative Structure-Activity Relationships	MARS	Multivariate Adaptive Regression Splines
HSA	Hydrated Surface Area	MCD	MonteCarlo version of MTD
HTS	High-Throughput Screening	MCI s	Molecular Connectivity Indices
HXID	Hu–Xu ID number	MCS	Maximum Common Substructure
IC	Information Content	MDDM	Main Distance-Dependent Matrix
ILGS	Iterated Line Graph Sequence	MDS	Molecular Dynamic Simulation
IP	Ionization Potential	MDS	MultiDimensional Scaling
IPE	Interaction Pharmacophore Element	MEDNE	Markovian Electron Delocalization NEgentropy
ISA	Isotropic Surface Area	MEP	Molecular Electrostatic Potential
IVEC	Iterative Vertex and Edge Centricity algorithm	MFP	MiniFingerPrints
IVS-PLS	Interactive Variable Selection–Partial Least Squares	MFTA	Molecular Field Topology Analysis
JEDA	Joint Entropy-based Diversity Analysis	MID	Molecular ID number
K-ANN	Kohonen Artificial Neural Networks	MIM	Molecular Influence Matrix
KLOGP	Klopman LOG P		
KNN	Kth Nearest Neighbor method		
LDA	Linear Discriminant Analysis		

MI-QSAR	Membrane Interaction–Quantitative Structure-Activity Relationships	OCWLI	Optimization of Correlation Weights of the Local Invariants
MLOGP	Moriguchi LOG P	OLS	Ordinary Least Squares regression
MLP	Molecular Lipophilicity Potential	PAR	Property-Activity Relationships
MLR	Multiple Linear Regression	PASS	Prediction of Activity Spectra of Substances
MNA	Multilevel Neighborhoods of Atoms descriptors	PCA	Principal Component Analysis
MO	Molecular Orbital	PCR	Principal Component Regression
MOA	Mode Of Action	PDT	Pharmacophore Definition Triplets
MPR	Matrix-Property-Response descriptors	PEI	Polarizability Effect Index
MQSA	Molecular Quantum Similarity Analysis	PELCO	Pérturbation d'un Environnement Limité Concentrique Ordonné
MQSI	Molecular Quantum Similarity Indices	PEOE	Partial Equalization of Orbital Electronegativities
MQSM	Molecular Quantum Similarity Measures	PEST	Property Encoded Surface Translator
MR	Misclassification Risk	PID	Prime ID number
MSA	Molecular Shape Analysis	PLS	Partial Least Squares regression
MSD	Minimal Steric Difference	PLS-DA	Partial Least Squares–Discriminant Analysis
MSE	Mean Square Error	PPF	Property and Pharmacophore Features fingerprints
MSG	Molecular SuperGraph	PPFS	Property and Pharmacophore Features Score
MSS	Model Sum of Squares	PPP	Potential Pharmacophore Point
MTD	Minimal Topological Difference	PRESS	Prediction Error Sum of Squares
MTI	Molecular Topological Index	PSA	Polar Surface Area
MUSEUM	MUtation and SElection Uncover Models	QDA	Quadratic Discriminant Analysis
MW	Molecular Weight	QMSA	Quantitative Molecular Similarity Analysis
MWC	Molecular Walk Count	QSAM	Quantitative Sequence Activity Model
NER	Nonrror Rate	QSAR	Quantitative Structure-Activity Relationship
NFD	Nucleophilic Frontier electron Density		
NN	Neural Networks		
NNC	Nearest Neighbor Code		
NOAEL	No-Observed-Adverse-Effect Level		
NOEL	No-Observed-Effect Level		
NSD	Nucleophilic SuperDelocalizability		
OASIS	Optimized Approach based on Structural Indices Set		

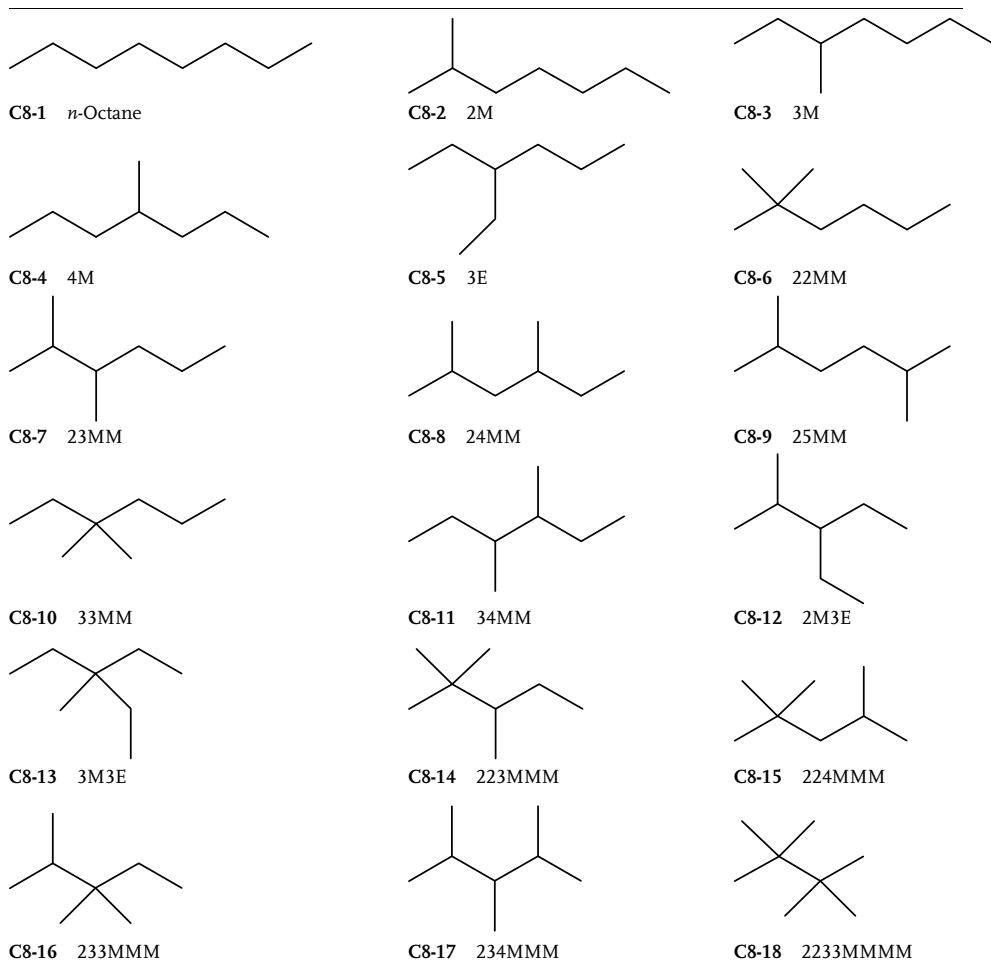
QSERR	Quantitative Structure-Enantioselective Retention Relationship	SAR	Structure-Activity Relationships
QShAR	Quantitative SHape-Activity Relationship	SASA	Solvent-Accessible Surface Area
QSiAR	Quantitative Similarity-Activity Relationship	SAVOL	Solvent-Accessible VOLUME
QSMR	Quantitative Structure-Mobility Relationship	SBL	Smallest Binary Label
QSPR	Quantitative Structure-Property Relationship	SDEC	Standard Deviation Error in Calculation
QSRC	Quantitative Structure/Response Correlation	SDEP	Standard Deviation Error in Prediction
QSRR	Quantitative Structure-Reactivity Relationship	SE	Shannon's Entropy
QSRR	Quantitative Structure-Retention Relationship	SEC	Standard Error in Calculation
QSTR	Quantitative Structure-Toxicity Relationship	SEP	Standard Error in Prediction
RBf-ANN	Radial Basis Function–Artificial Neural Network	SIBIS	Steric Interactions in BIological Systems
RBSM	Receptor Binding Site Model	SIC	Structural Information Content
RDA	Regularized Discriminant Analysis	SID	Self-returning ID number
RDF	Radial Distribution Function	SIMCA	Soft-Independent Modeling of Class Analysis
REC	Relative Error in Calculation	SMF	Substructural Molecular Fragment descriptors
REP	Relative Error in Prediction	SOM	Self-Organizing Maps
RID	Ring ID number	SOM-CP	Self-Organizing Maps–Counter-Propagation method
RMS	Root Mean Square error	SOMFA	Self-Organizing Molecular Field Analysis
RMSD	Root Mean Square Deviation	SOMO	Singly Occupied Molecular Orbital
RMSDP	Root Mean Square Deviation in Prediction	SPP	Submolecular Polarity Parameter
RMSE	Root Mean Square Error	SPR	Structure-Property Relationships
RMSEC	Root Mean Square Error in Calculation	SRC	Structure/Response Correlations
RMSEP	Root Mean Square Error in Prediction	SRR	Structure-Reactivity Relationship
ROC	Receiver Operator Characteristic curve	SRW	Self-Returning Walk
RR	Ridge Regression	SVM	Support Vector Machines
RSD	Residual Standard Deviation	SWC	StepWise Classification
RSM	Receptor Surface Model	SWIM	Spectral Weighted Invariant Molecular descriptors
RSS	Residual Sum of Squares	SWM	Spectral Weighted Molecular signals
SA	Surface Area		

SWR	StepWise Regression	TSS	Total Sum of Squares
TAEs	Transferable Atom Equivalents	UVE-PLS	Uniformative Variable Elimination by PLS
TAUs	Topochemically Arrived Unique indices	VEM	Valence Electron Mobile environment
TI	Topological Index	VFA	Voronoi Field Analysis
TIC	neighborhood Total Information Content	VR	Variable Reduction
TLP	Topological Lipophilicity Potential	VS	Variable Selection
TLSER	Theoretical Linear Solvation Energy Relationship	VS	Vertex Sum
TMSA	Total Molecular Surface Area	VSS	Variable Subset Selection
TOMOCOMD	Topological MOlecular COMputer Design	WHIM	Weighted Holistic Invariant Molecular descriptors
ToPD	Total Pharmacophore Diversity	WID	Weighted ID number
TOPP	Triplets Of Pharmacophoric Points	WLN	Wiswesser Line-formula Notation
TOSS-MODE	Topological SubStructure MOlecular DEsign	1D	one-dimensional
TPSA	Topological Partial Surface Area	2D	bi-dimensional
		3D	three-dimensional
		3D-MoRSE	3D-Molecule Representa tion of Structures based on Electron diffraction descriptors
		nD	n-dimensional

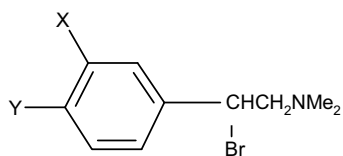
Appendix C.

Molecular structures

Set 1 – 18 octane isomers (C8)



M = methyl; E = ethyl.

Set 2 – 22 *N,N*-dimethyl- α -bromo-phenethylamines

No.	X	Y	log(1/C)	No.	X	Y	log(1/C)
1	H	H	7.46	12	Cl	F	8.19
2	H	F	8.16	13	Br	F	8.57
3	H	Cl	8.68	14	Me	F	8.82
4	H	Br	8.89	15	Cl	Cl	8.89
5	H	I	9.25	16	Br	Cl	8.92
6	H	Me	9.30	17	Me	Cl	8.96
7	F	H	7.52	18	Cl	Br	9.00
8	Cl	H	8.16	19	Br	Br	9.35
9	Br	H	8.30	20	Me	Br	9.22
10	I	H	8.40	21	Me	Me	9.30
11	Me	H	8.46	22	Br	Me	9.52

X and Y are the substituent groups and 1/C is the biological activity.