View Article Online / Journal Homepage / Table of Contents for this issue

PAPER

NJC
www.rsc.org/njc

A natural graph-theory model for partition and kinetic coefficients

Lionello Pogliani†

Dipartimento di Chimica, Università della Calabria, 87036 Rende (CS), Italy. E-mail: lionp@unical.it

Received (in Montpellier, France) 29th April 2005, Accepted 31st May 2005
First published as an Advance Article on the web 8th July 2005

A model of six partition coefficients and two metabolic kinetic parameters of a class of halogenated compounds has been performed with the aid of molecular connectivity concepts, which are based on complete graphs and general graphs. The complete graphs are used to encode the core electrons into the main parameter of molecular connectivity theory, the valence delta number, δ^v. The present model, which is solely based on graph concepts, confirms the central importance of a complete graph conjecture, which is based on an odd number of vertices for the core electrons. Two slightly different algorithms for δ^v, both centered on this conjecture, compete in deriving an optimal model for the six sets of partition coefficients. An algorithm, the K_p-(*p*-odd) algorithm, is valid for a model based on a linear combination of four connectivity and pseudoconnectivity basis indices. This linear combination is able to model, in a satisfactory way, all the six sets of partition coefficients. The other algorithm, the K_p-(*pp*-odd) algorithm, is, instead, able to derive a molecular connectivity term, which is able to adequately model four sets of partition coefficients, and in a less satisfactory way another set of partition coefficients. The two metabolic kinetic constants are, instead, optimally modeled by different basis indices based on the K_p-(*pp*-odd) algorithm. Underlying are also (i) the importance, in nearly all cases, of the $^1\chi^v$ basis index as best single descriptor, and (ii) the overall improvement the model undergoes when the only *cis*-compound of the class of chemicals is deleted form the class. Preliminary results show the possibility of improving the model with a δ^v algorithm, which encodes also the bonded hydrogens.

Downloaded by University of California - Los Angeles on 01 January 2013
Published on 08 July 2005 on http://pubs.rsc.org | doi:10.1039/B506091P

Introduction

Recently, a complete K_p graph representation for the core electrons of any atom of any row with principal quantum number $n \geq 2$ has been successfully used to model many properties of different classes of compounds, including inorganic compounds.[1–5] In these studies the molecular connectivity indices have been derived using a natural valence delta number, δ^v, *i.e.*, a δ^v based solely on graph concepts, and specifically on the concepts of complete graphs and of general graphs (also known as pseudographs[6]). The given algorithm for δ^v, which for second row atoms equals the previous quantum algorithms,[7,8] is four-valued, when both odd and sequential complete graphs are considered, as we shall see in the methodology section. Now, throughout the model of many properties[1–4] an intriguing preference for odd complete graphs showed up, a preference that could only be ascribed to the particular choice of the properties, as a theoretical study on the ability of δ^v to encode the core electrons[7] showed no preference either for odd complete graphs or for sequential complete graphs. Up to now only one property of a set of inorganic compounds[3] has been detected whose model could only be satisfactorily achieved with descriptors based on sequential complete graphs for the core electrons.

The model of a property of a class of compounds which assumes different values in different media could throw more light both on the preference and the consistency of either odd or sequential complete graphs, when used to encode the core electrons. Long ago Cargas *et al.*[9] published six sets of experimental values for the solvent/air and tissue/air partition coefficient, as log *P*, for twenty-five halogenated compounds, as well as the experimental metabolic rate constants (*MK*) for nineteen halogenated compounds. It should here be remembered that the partition constants of a given compound between different media can vary by many orders of magnitude. For example, octane has a hexane/air partition coefficient of 9300 whereas its ethanediol/air partition coefficient is only 13.[8] We will here try to test which complete graph conjecture gives rise to the best model for the six sets of log *P* values, and the two sets of *MK* values, having in mind the guideline that *similar properties of similar compounds should be described by similar descriptors*.

Methodology

Before entering into the details of the method let us review some graph concepts, and especially the concept of complete graphs, which will be used throughout the present paper.[7,11–15] A simple graph can be defined as a set of vertices, and a set of edges that connect these vertices. A simple chemical graph is a graph representation of a molecule. Normally, but not always, in chemical graph theory use is made of hydrogen-suppressed or -depleted chemical graphs, either simple graphs or general graphs. The degree of a vertex of a graph is the number of edges that occur with it. The degree of a vertex in a chemical graph is in some cases called its valence. The hydrogen-depleted graph of 1,2-dichloroethylene (either *cis* or *trans*) is shown in Fig. 1A. Here the degree of the two vertices encoding the carbon atoms is two, while the degree of the chlorine vertices is just one. Molecular connectivity theory[7,8] defines a delta number, δ, for the degree of a vertex in a chemical graph. It should here be stressed that the normal Euclidean metrics in graph theory has no meaning, *i.e.*, edge lengths and angles have no meaning. A general graph (or pseudograph) is a graph that allows for multiple connections and self-connections

† On sabbatical leave, present address: L. Pogliani, Unidad de Investigación en Diseño de Farmacos y Conectividad Molecular, Facultad de Farmacia, Dept. de Química Física, Av. V. A. Estellés s/n, 46100 Burjassot (València), Spain.

DOI: 10.1039/b506091p

This journal is © The Royal Society of Chemistry and the Centre National de la Recherche Scientifique 2005

View Article Online

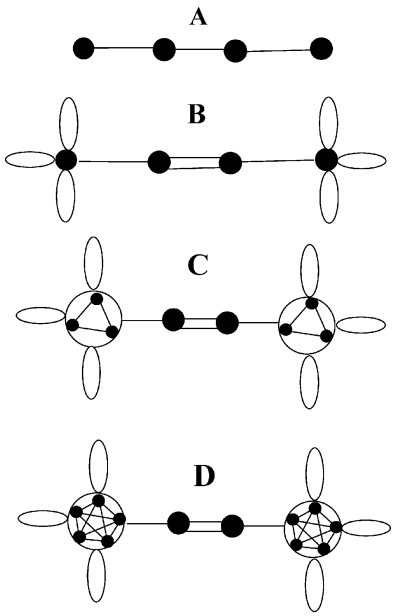

Fig. 1 A: The chemical graph of 1,2 dichloroethylene; B: the chemical pseudograph of 1,2 dichloroethylene; C, and D: the pseudograph plus complete graph of 1,2 dichloroethylene, and 1,2 dibromoethylene, respectively.

(or loops), and a pseudograph is the representation of a molecule with a general graph. The degree of a vertex (or valence) in a general graph is the number of edges that occur with it. Even if this definition is the same for a graph and a general graph , the end result is not always the same. In Fig. 1B is shown the hydrogen-suppressed (HS) chemical general graph of 1,2-dichloroethylene; here the degree of the vertices is three for the carbon atoms and seven for the chlorine atoms, as each loop on the chlorine atoms contributes twice to its degree. Molecular connectivity theory defines the degree of a vertex in a chemical general graph as its valence delta number, δ^v, which, for our purposes, will be redefined as δ^v(ps), where *ps* means pseudograph. Now, the general graph of 1,2-dichloroethylene could also be used to encode 1,2-difluoroethylene, 1,2-dibromoethylene, and 1,2-diiodoethylene. It is thus evident that even if the concept of general graph allows a rather faithful encoding of a molecule, allowing for multiple bonds and non-bonding electrons to be 'graph' encoded, a problem arises at the level of the core electrons. In fact, with higher-row atoms, *i.e.*, for atoms with $n > 2$ (n is the principal quantum number), the contribution of the core electrons has to be taken into due account.

The graph representation for the core electrons is centered on the concept of complete graphs. A complete graph, K_p, of order p is a graph where every pair of its vertices is adjacent. A complete graph is always r-regular ($r = p - 1$), *i.e.*, it has all vertices with the same degree, r. The contrary is not true, *i.e.*, a regular graph is not always complete. In Fig. 1C the chemical general graph plus complete graph of 1,2-dichloroethylene is shown. Here, the two carbons are represented with a K_1 complete graph, *i.e.*, a vertex, while the two chlorine atoms are represented with a K_3 complete graph. Finally, in Fig. 1D the chemical pseudograph (or general graph) plus complete graph of 1,2-dibromoethylene is shown. The two carbons are, again, represented with a K_1 complete graph, while the two bromine atoms are represented with a K_5 odd complete graph. Thus, different core electrons have been encoded with different 'graph' concepts, *i.e.*, in practice p mimics, in a broad sense, the principal quantum number n. The dimension of the halogen vertices has no metrical meaning, it is just a zoom of this vertex, whose dimensions are as meaningless as the dimensions of the K_1 vertex. The circle encircling the complete graph has also no meaning; it is just a frame to underline the property. The depleted hydrogens of a graph could be considered sort of null graphs, K_0, which possess no points and no edges.[16] The δ^v algorithm, which includes the complete graph contribution, can now be defined in the following way:[1–4]

$$\delta^v = q \cdot \delta^v(\mathrm{ps})/(p \cdot r + 1) \qquad (1)$$

The parameter $p \cdot r$ equals the sum of all vertex degrees in complete graphs, and this sum equals twice the number of connections. The most successful type of complete graph, to date, was the odd complete graphs with $p = 1, 3, 5, 7, \ldots$ while the sequential complete graphs with $p = 1, 2, 3, 4, \ldots$, seem to play a minor role. Even complete graphs with $p = 2, 4, 6, \ldots$, which give rise to a series of conceptual difficulties, as K_1 vertices should be replaced by K_2 vertices, will be discarded. Two values have been chosen for the adjustable q parameter, $q = 1$ and $q = p$. For $q = p$, the resulting δ^v values are rather similar to the $\delta^v = (2/n)^2\delta^v(\mathrm{ps})$ values of the E_S concept, while for $q = 1$, the resulting δ^v are rather similar to the $\delta^v = \delta^v(\mathrm{ps})/(Z - Z^v - 1)^4$ values of the molecular connectivity concept. Clearly, for second-row atoms, $\delta^v = \delta^v(\mathrm{ps})$, as it should be. Algorithm (1) can thus be four-valued. The case with $q = 1$ and p odd will be denoted with K_p-(p-odd), the case with $q = p$ and p odd with K_p-(pp-odd), the case with $q = 1$ and p sequential with K_p-(p-seq), and the case $q = p$ and p sequential with K_p-(pp-seq).

The following subsets of basis indices will be used to model the molecular connectivity indices, the molecular pseudoconnectivity indices and the dual indices:[5,6,15–20]

$$\begin{aligned}\{\chi\} &= \{\{D, {}^0\chi, {}^1\chi, \chi_t, D^v, {}^0\chi^v, {}^1\chi^v, \chi_t^v\},\\ \{\psi\} &= \{{}^S\psi_I, {}^0\psi_I, {}^1\psi_I, {}^T\psi_I, {}^S\psi_E, {}^0\psi_E, {}^1\psi_E, {}^T\psi_E\}\\ \{\beta_d\} &= \{{}^0\chi_d, {}^1\chi_d, {}^1\chi_s, {}^0\chi_d^v, {}^1\chi_d^v, {}^1\chi_s^v,\\ &\quad {}^0\psi_{Id}, {}^1\psi_{Id}, {}^1\psi_{Is}, {}^0\psi_{Ed}, {}^1\psi_{Ed}, {}^1\psi_{Es}\}\end{aligned} \qquad (2)$$

Basis χ and ψ indices are formally similar, in fact,

$$D = \Sigma_i\delta_i,\ {}^0\chi = \Sigma_i(\delta_i)^{-0.5},\ {}^1\chi = \Sigma(\delta_i\delta_j)^{-0.5},\ \chi_t = (\Pi\ \delta_i)^{-0.5}$$

$${}^S\psi_I = \Sigma_i I_i,\ {}^0\psi_I = \Sigma_i(I_i)^{-0.5},\ {}^1\psi_I = \Sigma(I_iI_j)^{-0.5},\ {}^T\psi_I = (\Pi\ I_i)^{-0.5}$$

Sums in the first and second equations, as well as products (Π), are taken over all vertices of the hydrogen-suppressed chemical graph. Sums in the third equations (in both rows) are over all edges of the chemical graph (σ bonds in a molecule). Replacing δ with δ^v the subset of valence χ^v indices is obtained, while replacing I_i with S_i the ψ_E subset is obtained. Superscripts S and T mean sum and total, while the other sub- and superscripts follow the established denomination for χ indices.[5] Throughout the present study also three subgraph indices, for F, Cl, and Br, will be used:

$$D_F = \Sigma_i\delta_F,\ D_{Cl} = \Sigma_i\delta_{Cl},\ D_{Br} = \Sigma_i\delta_{Br}$$

The dual basis β_d indices are defined in the following way,[20]

$${}^0\chi_d = (-0.5)^N\Pi_i(\delta_i),\ {}^1\chi_d = (-0.5)^{(N+\mu-1)}\Pi(\delta_i + \delta_j),$$
$${}^1\chi_s = \Pi(\delta_i + \delta_j)^{-0.5}$$

$${}^0\psi_{Id} = (-0.5)^N\Pi_i(I_i),\ \psi_{Id} = (-0.5)^{(N+\mu-1)}\Pi\ (I_i + I_j),$$
$${}^1\psi_{Is} = \Pi\ (I_i + I_j)^{-0.5}$$

Replacing δ by δ^v and I_i by S_i the corresponding χ^v valence dual and ψ_E dual indices are obtained. N is the number of vertices, and exponent μ is the cyclomatic number. It indicates the number of cycles of a chemical graph and it is equal to the minimal number of edges necessary to be removed in order to convert the (poly)cyclic graph to an acyclic subgraph.

Basis ψ indices are indirectly related to the δ^v number, defined in eqn. (1), through the I-state (ψ_I subset) and S-state (ψ_E subset) indices:[8]

$$I = (\delta^v + 1)/\delta,\ S = I + \Sigma\Delta I, \text{ with: } \Delta I = (I_i - I_j)/r_{ij}^2 \qquad (3)$$

Downloaded by University of California - Los Angeles on 01 January 2013
Published on 08 July 2005 on http://pubs.rsc.org | doi:10.1039/B506091P

View Article Online

Table 1 Liquid and rat tissue: air partition coefficients (PC) at 37 °C, given as log *P*, for twenty-five halo compounds. In the last two columns are the metabolic rate constants for sixteen halo compounds

Molecule[a]	Saline	Olive oil	Blood	Liver	Muscle	Fat	V_{max}/mg h^{-1}	K_M/mg l^{-1}
MeCl	−0.056	0.933	0.393	0.540	−0.013	1.130	—	—
$MeCl_2$	0.775	2.117	1.288	1.152	0.899	2.079	4.0	0.4
$CHCl_3$	0.529	2.604	1.318	1.324	1.143	2.307	7.0	0.25
CCl_4	−0.456	2.573	0.655	1.152	0.657	2.555	0.4	0.25
CH_2═$CHCl$	−0.367	1.387	0.225	0.204	0.342	1.301	2.5	0.1
CCl_2═CH_2	−0.456	1.808	0.699	0.645	0.312	1.836	7.5	0.1
CHCl═CHCl(*c*)	0.512	2.444	1.334	1.185	0.785	2.356	3.0	0.5
CHCl═CHCl(*t*)	0.149	2.250	0.981	0.952	0.547	2.170	3.0	0.1
CCl_2═$CHCl$	−0.081	2.743	1.340	1.435	1.004	2.744	11.0	0.25
CCl_2═CCl_2	−0.102	3.329	1.276	1.847	1.301	3.214	—	—
CH_3–CH_2Cl	0.037	1.590	0.611	0.558	0.508	1.587	4.0	0.1
$CHCl_2$–CH_3	0.389	2.270	1.049	1.033	0.709	2.215	7.5	0.2
CH_2Cl–CH_2Cl	1.057	2.563	1.483	1.553	1.369	2.537	3.15	0.25
CCl_3–CH_3	−0.125	2.470	0.760	0.934	0.498	2.420	—	—
$CHCl_2$–CH_2Cl	1.124	3.249	1.763	1.863	1.360	3.158	7.7	0.78
$CHCl_2$–$CHCl_2$	1.369	3.803	2.152	2.292	2.004	3.576	12.0	0.8
CCl_3–CH_2Cl	0.548	3.429	1.620	1.945	1.597	3.332	6.5	0.92
CH_2F_2	0.117	0.678	0.204	0.439	0.158	0.155		
CH_2FCl	0.489	1.348	0.706	0.537	0.391	1.188		
CH_2BrCl	0.937	2.558	1.618	1.465	1.045	2.512		
CH_2Br_2	1.158	2.981	1.870	1.833	1.607	2.899		
CF_3-CHClBr	−0.301	2.297	0.721	0.882	0.649	2.260		
CH_2═$CHBr$	−0.357	1.748	0.607	0.522	0.354	1.692		
CH_2Br–CH_2Cl	0.950	2.755	1.722	1.631	1.405	2.982		
CF_3–CH_2Cl	−0.377	1.380	0.104	0.265	0.090	1.326		
$CHCl_2$–Cl_3							9.2	0.9
CCl_3–CCl_3							2.0	0.8

[a] *c* stands for *cis*, and *t* for *trans*.

Downloaded by University of California - Los Angeles on 01 January 2013
Published on 08 July 2005 on http://pubs.rsc.org | doi:10.1039/B506091P

The *S* values for F and Cl atoms here will play an interesting role. When several Cl or F atom are present in a molecule the corresponding S_F and S_{Cl} values will be the average over all Cl or F atoms, respectively. r_{ij} counts the atoms in the minimum path length separating two atoms *i* and *j*, which is equal to the usual graph distance, $d_{ij} + 1$. Factor $\Sigma\Delta I$ incorporates the information about the influence of the remainder of the molecular environment, and, as it can be negative, *S* can also be negative. To avoid imaginary ψ_E values, every *S* value of our classes of compounds (as some atoms have $S < 0$), has been rescaled to the *S* value in CF_4 for the studied halogenated compounds, *i.e.*, $S(C) = -5.5$. Normally the rescaling procedure has minor effects on the model.[18]

The linear combination of basis indices (LCBI) normally used to model a property *P* can be written as a dot product: $P = \boldsymbol{\beta} \cdot \boldsymbol{C}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, U_0)$ is the vector of the molecular connectivity basis indices (dual indices excluded), and $\boldsymbol{C} = (c_1, c_2, \ldots, c_0)$ is the vector of the coefficient of the linear regression used for the model. $U_0 \equiv 1$ is the unitary index, and c_0 its regression parameter. The components of vector $\boldsymbol{\beta}$ are chosen by the aid of a combinatorial procedure, which searches the best basis β_i indices. It is now possible to construct by the aid of a trial-and-error procedure a series of higher-order descriptors, known as terms, $X = f(\chi)$, $Y = f(\psi)$, $Z = f(X, Y)$, and $Z' = f(Z, \beta_d)$. To avoid a huge combinatorial problem the dual indices will only be used to improve the quality of the *X*, *Y*, or *Z* terms. The model equation, when terms are used normally, simplifies into $P = c_1T + c_0U_0$, where $T = X$, *Y*, *Z*, or Z'.[17–20] The quality of a description is tested by *r* (correlation coefficient), *s* (standard deviation of the estimates), and *F* (the Fisher ratio) statistics. Further, to check how much *s* improves with successive descriptors, $s_R = s_0/s_i$, is introduced, where s_0 is the *s* value for the best single-basis-index, and s_i the *s* value of the other descriptors. Now, all statistics increase with increasing quality of the model. To check how far the regression parameters, c_i, are meaningful, their utility will be given as $u_i = c_i/e_i$, where e_i is the error of c_i. The u_i values of the linear regression parameters used to model have been collected into the form of a vector, $\boldsymbol{u} = (u_1, u_2, \ldots, u_0)$.

Results and discussion

In the following we will start to model the partition coefficients by considering case after case, *i.e.*, the six sets of partition coefficients will be separately modeled in two different ways. Then, the overall case will be discussed *i.e.*, all six cases together will also be modeled with a unique descriptor in two different ways. Finally, some recent results obtained with a δ^v algorithm, which is able to encode the hydrogen atoms, will be presented.

The individual models

The logarithms of the partition coefficients, log *P*, of a set of twenty-five halo compounds as well as the metabolic kinetic constants, the maximum velocity, V_{max}, and the Michaelis–Menten constant, K_m, for sixteen halo compounds are collected in Table 1. The original values are taken from ref. 9. Superscript *po*, *ps*, *ppo*, and *pps* will be used for basis indices based on the K_p-(*p*-odd), K_p-(*p*-seq), K_p-(*pp*-odd), and K_p-(*pp*-seq), representation for the core electrons, respectively. The best single basis indices for the six different cases of twenty-five values of log *P*, *i.e.*, saline (Sal), olive oil (Oil), blood (Bl), liver (Liv), muscle (Mus), and fat, are collected in Table 2. Two things should be noted here: (i) the importance of the K_p-(*p*-odd) representation for the core electrons throughout the six cases, and (ii) the importance, with the exception of the saline case, of the $^1\chi^v$ index, which is the best single-index for five cases.

In Table 3 are collected the best multi-basis index descriptors for the individual log *P* cases. Here, for every case, the prediction coefficient q^2 has also been given, where the prediction has been made with the leave-one-out cross-validation

View Article Online

Table 2 The best single-index descriptor for the six individual log *P* cases

	Sal	Oil	Bl	Liv	Mus	Fat
{β}	$^{po}\{^{T}\psi_I\}$	$^{po}\{^{1}\chi^{v}\}$	$^{po}\{^{1}\chi^{v}\}$	$^{po}\{^{1}\chi^{v}\}$	$^{po}\{^{1}\chi^{v}\}$	$^{po}\{^{1}\chi^{v}\}$
F	5.1	99	27	48	44	124
R	0.427	0.901	0.737	0.822	0.809	0.918
s_0	0.53	0.35	0.39	0.34	0.32	0.33

method (note that $s_R = s_0/s$). A value $q^2 > 0.5$ is normally accepted as a satisfactory result.[21] Note (i) the improvement over the previous one-index case and the very satisfactory q^2 values, especially for the fat, oil, and liver cases, (ii) the multi-index descriptors are all based on a K_p-(*p*-odd) representation for the core electrons, and (iii) the $^{po}\{^{0}\chi, {}^{0}\psi_I, D_F, D_{Cl}\}$ multi-index combination is the best descriptor for Sal, Bl, and Liv cases. Furthermore, (iv) this four-index K_p-(*p*-odd) combination shows that the number of fluorine and chlorine atoms, D_F, D_{Cl}, is important in improving the model quality, while (v) of the four indices only the $^{0}\psi_I$ index is dependent on δ^v, *i.e.*, only this index includes a K_p description for the core electrons. In fact, the other three indices are δ-dependent only. The best three-index linear combination for the olive oil log *P*, $^{po}\{^{0}\chi, D^{v}, {}^{0}\psi_I\}$, is also a good descriptor, even if inferior, for fat-log *P* with: $F = 210$, $r = 0.984$, $s_R = 2.2$, $N = 25$. Thus, linear combination $^{po}\{^{0}\chi, D^{v}, {}^{0}\psi_I\}$ is a rather good descriptor of $N = 50$ points. The contrary, instead, is not true: the $^{po}\{^{0}\chi, {}^{1}\chi^{v}, {}^{0}\psi_I,\}$ for the fat tissue is not a good descriptor for olive oil-log *P*.

The correlation vectors for these descriptors for the six cases, which will allow us to obtain the corresponding calculated log *P* values, are:

Sal: $^{po}\{^{0}\chi, {}^{0}\psi_I, D_F, D_{Cl}\}$:
$\mathbf{C} = (-3.7718, 3.70213, 1.36892, 0.35451, 1.93033)$

Oil: $^{po}\{^{0}\chi, D^{v}, {}^{0}\psi_I\}$:
$\mathbf{C} = (-1.8722, 0.04176, 2.58623, 1.05843)$

Bl: $^{po}\{^{0}\chi, {}^{0}\psi_I, D_F, D_{Cl}\}$:
$\mathbf{C} = (-2.60992, 2.92293, 0.70755, 0.14607, 1.37927),$

Liv: $^{po}\{^{0}\chi , {}^{0}\psi_I, D_F, D_{Cl}\}$:
$\mathbf{C} = (-2.4135, 2.74399, 0.70903, 0.22563, 1.10583),$

Mus: $^{po}\{D, {}^{0}\chi^{v}, {}^{1}\chi^{v}, {}^{1}\psi_E\}$:
$\mathbf{C} = (0.35769, -0.94262, 2.08192, -2.80391, -0.38763),$

Fat: $^{po}\{^{0}\chi, {}^{1}\chi^{v}, {}^{0}\psi_I,\}$:
$\mathbf{C} = (-1.08143, 0.42748, 1.64032, 0.53332)$

Note the similarity among the correlation values for the blood and liver cases, which share with the saline case the same descriptor. Actually even the correlation vector for the saline case is not that different from these two.

Molecular connectivity terms of the *X*-type offer a good alternative for an optimal model; nearly all these terms are obtained by the aid of the K_p-(*pp*-odd) representation for the core electrons, with some interesting exceptions, like the saline and blood cases, where the K_p-(*pp*-seq) representation gives rise to the best description. In Table 4 the statistics of the best *X* terms are given, where in parentheses throughout the *F*, *r*, s_R, and *N* columns are shown the statistics for the deletion of the *cis*-1,2-dichloroethylene. Note that every time these statistics increase the corresponding statistics for the deletion of *trans*-1,2-dichloroethylene decrease. For the muscle case excluding either the *cis*- or *trans*-1,2-dichloroethylene there is, practically, no improvement.

In the following are the exact form for the different *X* terms. Note the improvement that the *X*-term for liver and fat undergoes when the dual index $^{0}\chi_d^{v}$ is inserted in it. The statistics of the corresponding *X′* terms are shown in the Table 4 (the *X*-statistics are shown below the corresponding term). With the $^{ppo}X'$(Liv) term and its ***C*** vector the calculated values of Fig. 2 have been obtained. Most of the residuals (▲) are within a $\pm s$ range, and all of them are within a $\pm 2s$ range.

$$^{pps}X(\text{Sal}) = [3.3D_{Cl} + 2.3(D_{Br})^{0.7} + 0.8S_F - 0.1S_{Cl} - 0.3(^{0}\chi)^{2}]^{2.3} (^{1}\chi)^{10.5}/(D^{v})^{8.1}$$

$$^{ppo}X(\text{Oil}) = [0.4 \cdot D + 2.6 \cdot (D_{Cl} + D_{Br}) + 0.2 \cdot (S_F - S_{Cl})]^{0.65}(^{1}\chi)^{1.1}/(D^{v})^{0.8}$$

$$^{pps}X(\text{Bl}) = [3.2D_{Cl} + 2.7(D_{Br})^{0.8} + 0.5S_F - 0.1S_{Cl} - 0.3(^{0}\chi)^{1.8}]^{0.7}(^{1}\chi)^{2.7}/(D^{v})^{2}$$

$$^{ppo}X(\text{Liv}) = [3D_{Cl} + 2.6(D_{Br})^{0.8} + 0.6S_F - 0.1S_{Cl} - 0.3(^{0}\chi)^{1.8}]^{0.7} (^{1}\chi)^{3.2}/(D^{v})^{1.9}$$

$F = 580$, $r = 0.981$, $s_R = 2.8$, $N = 25$, $\boldsymbol{u} = (24, 1.1)$

$$^{ppo}X'(\text{Liv}) = [^{ppo}X(\text{Liv}) - 0.0001\,^{0}\chi_d^{v}]: \mathbf{C} = (7.41210, -0.0648)$$

$$^{ppo}X(\text{Mus}) = [2.9D_{Cl} + 2.6(D_{Br})^{0.8} + 0.7S_F - 0.07S_{Cl} - 0.3(^{0}\chi)^{1.8}](^{1}\chi)^{5.2}/(D^{v})^{3.1}$$

$$^{ppo}X(\text{Fat}) = \{[0.6 \cdot D + 3.7 \cdot (D_{Cl} + D_{Br})^{0.9} + 0.2 \cdot (S_F - S_{Cl})]^{0.65}(^{1}\chi)^{1.1}/(D^{v})^{0.80}\}^{0.7}$$

$F = 1169$, $r = 0.990$, $s_R = 3.0$, $N = 25$, $\boldsymbol{u} = (34, 14)$

$$^{ppo}X'(\text{Fat}) = [^{ppo}X(\text{Fat}) - 0.0002\,^{0}\chi_d^{v}]$$

The metabolic kinetic constants, V_{max} and K_M

In Table 1 are also shown the metabolic Michaelis–Menten kinetic constants of nineteen halogenated hydrocarbons, V_{max} and K_M. As already suggested[9] only sixteen compounds give rise to a satisfactory model, and their values are shown in Table 1. Table 5 shows that the two properties can satisfactorily be modeled with a linear combination of basis indices, but this time derived by the aid of the K_p-(*pp*-odd) representation for the core electrons. The best single-index also belongs to this representation for the core electrons. For V_{max} an electrotopological-state index, $^{ppo}S_{Cl}$ (see eqn. (3)) is the best single-basis index. The *cis*- and *trans*-compounds are rather similar

Table 3 The best combination of basis indices for the six individual log *P* cases

Case	LCBI	*F*	*r*	s_R	q^2	***u***
Sal	$^{po}\{^{0}\chi, {}^{0}\psi_I, D_F, D_{Cl}\}$	26	0.917	2.1	0.765	(9.3, 9.6, 7.0, 3.7, 6.7)
Oil	$^{po}\{^{0}\chi, D^{v}, {}^{0}\psi_I\}$	135	0.975	1.8	0.935	(5.7, 2.1, 8.5, 5.5)
Bl	$^{po}\{^{0}\chi , {}^{0}\psi_I, D_F, D_{Cl}\}$	38	0.941	1.9	0.828	(7.6, 9.1, 4.3, 1.8, 5.7)
Liv	$^{po}\{^{0}\chi , {}^{0}\psi_I, D_F, D_{Cl}\}$	52	0.955	1.8	0.850	(7.9, 9.5, 4.8, 3.2, 5.1)
Mus	$^{po}\{D, {}^{0}\chi^{v}, {}^{1}\chi^{v}, {}^{1}\psi_E\}$	47	0.951	1.8	0.838	(4.9, 4.5, 7.3, 6.2, 2.8)
Fat	$^{po}\{^{0}\chi, {}^{1}\chi^{v} , {}^{0}\psi_I,\}$	264	0.987	2.4	0.965	(10, 3.5, 9.8, 4.5)

Published on 08 July 2005 on http://pubs.rsc.org | doi:10.1039/B506091P
Downloaded by University of California - Los Angeles on 01 January 2013

View Article Online

Table 4 The best molecular connectivity terms for the six individual log *P* cases

Case	*X*-type term	*F*	*r*	s_R	*N*	*u*
Sal	^{pps}X(Sal)	86 (90)	0.889 (0.896)	2.0 (2.0)	25 (24)	(9.3, 3.8)
Oil	^{ppo}X(Oil)	585 (634)	0.981 (0.984)	2.3 (2.3)	25 (24)	(24, 2.2)
Bl	^{pps}X(Bl)	175 (184)	0.940 (0.945)	1.9_5 (2.1)	25 (24)	(13, 3.4)
Liv	$^{ppo}X'$(Liv)	669 (725)	0.983 (0.985)	3.1 (3.4)	25 (24)	(26, 1.3)
Mus	^{ppo}X(Mus)	196 (192)	0.946 (0.947)	1.8 (1.8)	25 (24)	(14, 1.0)
Fat	$^{ppo}X'$(Fat)	1270 (1409)	0.991 (0.992)	3.0 (3.3)	25 (24)	(36, 14)

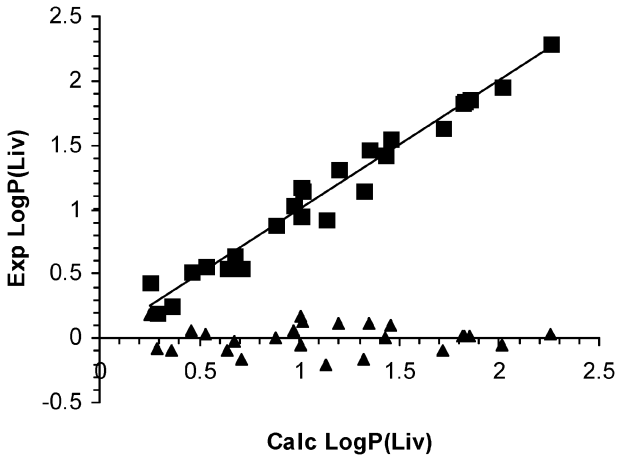

Fig. 2 Plot of the experimental *vs.* calculated log *P* (with $^{ppo}X'_{pp}$(Liv) descriptor) for Liver, for twenty five halo compounds, together with the corresponding residual plot (▲).

V_{max} values, and this explains the lack of improvement in eliminating either the *cis* or the *trans* form.

For K_M, whose LCBI show an unsatisfactory decreasing *F* value, an interesting K_p-(*pp*-seq) higher-order $^{pps}X(K_M)$ term could be detected:

$$^{pps}X(K_M) = [2.5D_{Cl} - 0.08S_{Cl}{}^{1.3} - 0.3(^0\chi)^{1.8}]^2(^1\chi)^{8.1}/(D^v)^{5.8}$$

$$F = 66,\ r = 0.907,\ s_R = 1.3,\ N = 16$$

Here, excluding *cis*-1,2-dichloroethylene the model improves: $F = 98$, $r = 0.940$, $s_R = 1.5$, $N = 15$.

The overall model

The possibility to derive with a descriptor trained on twenty-five values of a single case a model for the remaining cases will now be further perused. In Table 6 are shown the interrelation values, *r*, among the different partition coefficients. Remembering that an accepted collinearity criterion tells that a value of $r > 0.98$ denotes strongly interrelated parameters, and that even strongly interrelated indices can help to improve the description,[22,23] then this matrix indicates that, quite probably, a single descriptor could satisfactorily model most of the cases, and perhaps even all six cases. In fact, a closer look at the bottom matrix of this table it is easy to note that the $^{ppo}X'$(Liv) term is a good descriptor for the olive oil, muscle, and fat cases, and also, even if in a poorer way, of the blood case, *i.e.*, the following eqn. (4) is the optimal fit model equation for $N = 125$ points:

$$\log P_i = c_{1i}{}^{ppo}X'(\text{Liv}) + c_{0i} \qquad (4)$$

with i = Liv, Oil, Mus, Fat, and Bl. The regression vectors for Fig. 3, where only the optimal Liv, Oil, Mus, and Fat case are plotted, are:

$$\boldsymbol{C}(\text{Oil}) = (9.66578, 0.73730),\ \boldsymbol{C}(\text{Mus}) = (6.50886, -0.21827),$$
$$\boldsymbol{C}(\text{Fat}) = (9.81295, 0.64253)$$

For $\boldsymbol{C}$(Liv) see the liver X' term. In this figure most of the residuals (▲) are within a $\pm s_{aver}$ range.

A closer examination of the optimal combination, $^{ppo}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$, for the saline, blood, and liver cases tells us that with the exclusion of an interrelation of $r = 0.94$ between $^0\psi_I$ and $^0\chi$, all the other interrelations among these indices have $r \leq 0.74$, *i.e.*, practically, following the criterion of ref. 22, our indices are orthogonal to each other. Combination $^{po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ gives also rise to an interesting fit-model for the three remaining cases, for a total of $N = 150$ points, with the following statistics (for Sal, Bl, and Liv cases see the corresponding paragraph):

Olive Oil: $F = 99$, $r = 0.976$, $s_R = 1.8$,
$\boldsymbol{C} = (-1.8217, 2.56576, 0.29914, 0.06943, 1.06648)$

Muscle: $F = 41$, $r = 0.944$, $s_R = 1.7$,
$\boldsymbol{C} = (-2.34576, 2.69296, 0.68124, 0.14847, 0.86830)$

Fat: $F = 150$, $r = 0.984$, $s_R = 2.1$,
$\boldsymbol{C} = (-1.62782, 2.47560, 0.12546, -0.0317, 0.80659)$

The calculated/observed plot for the same $n = 100$ points of the previous case, to allow an easy comparison, is shown in Fig. 4. This plot is better than the plot of Fig. 3: the residuals are much more even. A 'broad' external validation test will now be tried, *i.e.*, five compounds chosen at random (CH_2=CHCl, CH_2Cl–CH_2Cl, CCl_3–CH_2Cl, CH_2Br_2, and CF_3–CH_2Cl), will, now, be left out and the model equation derived by the aid of twenty training points will be used to externally model the log *P* of these five compounds, for the cases liver, muscle, oil, and fat, *i.e.*, for a total of twenty externally

Table 5 The best single- and multi-index descriptors for the V_{max} and K_M kinetic constants

	{β}	*F*	*r*	s_0	*N*
V_{max}	$^{ppo}\{S_{Cl}\}$	3	0.418	3.18	16
K_M	$^{ppo}\{^1\chi^v\}$	30	0.824	0.18	16

	LCBI	*F*	*r*	s_R	*N*	*U*
V_{max}	$^{ppo}\{\chi_t, {}^0\psi_E, {}^1\psi_E, D_{Cl}\}$	22	0.944	2.4	16	(6.7, 8.2, 8.4, 3.5, 7.4)
K_M	$^{ppo}\{\chi_t{}^v, {}^1\psi_I, {}^1\psi_E, {}^T\psi_E\}$	16	0.926	1.3	16	(3.1, 4.3, 3.5, 3.1, 1.6)

Downloaded by University of California - Los Angeles on 01 January 2013
Published on 08 July 2005 on http://pubs.rsc.org | doi:10.1039/B506091P

View Article Online

Table 6 The interrelation values among the different log *P* values (top), and the interrelation values between the optimal terms and log *P*

	Sal	Oil	Bl	Liv	Mus	Fat
Saline	1	0.54	0.84	0.72	0.77	0.50
Olive Oil		1	0.87	0.94	0.91	0.98
Blood			1	0.94	0.94	0.86
Liver				1	0.96	0.92
Muscle					1	0.88
Fat						1
$^{pps}X_{pp}$(Sal)	0.89	0.63	0.85	0.80	0.80	0.63
$^{ppo}X_{pp}$(Oil)	0.46	0.98	0.82	0.92	0.87	0.98
$^{pps}X_{pp}$(Bl)	0.75	0.87	0.94	0.95	0.91	0.87
$^{ppo}X'_{pp}$(Liv)	0.66	0.95	0.91	0.98	0.94	0.94
$^{ppo}X_{pp}$(Mus)	0.70	0.91	0.91	0.97	0.95	0.90
$^{ppo}X'_{pp}$(Fat)	0.45	0.97	0.82	0.90	0.86	0.99

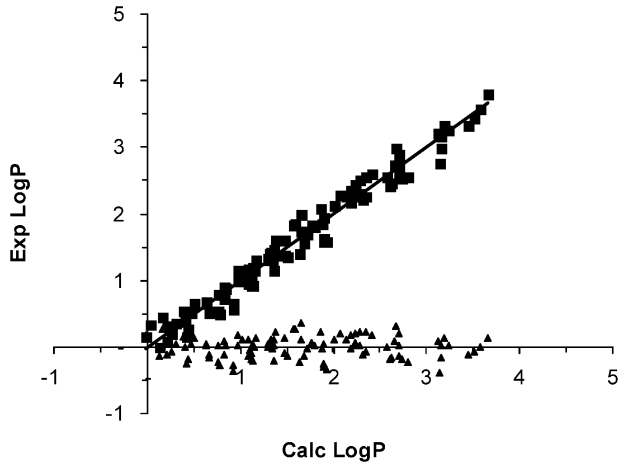

Fig. 3 Plot of the experimental *vs.* calculated log *P* (with $^{ppo}X'_{pp}$(Liv) descriptor) for Liver, Olive oil, Muscle, and Fat, for a total of $n = 100$ points, together with the corresponding residual plot (▲).

validated points. For the twenty points (instead of twenty-five) of saline, blood and liver cases this combination continues to be the optimal combination. The model, for the same hundred points, is shown in Fig. 5, where the twenty external validated points are characterized by open figures. The model and the validation seem quite good.

We have here proposed two optimal fit-models, to underline the flexibility of the molecular connectivity method. One fit-model is based on a rather convoluted connectivity *X* term, which maximizes most of the statistics for four sets of partition coefficients. The other fit-model, which is based on a combination of four basis indices, is able to derive a satisfactory fit-model for all six sets of partition coefficients. Which fit-model to use depends on the 'taste' of the reader. Both models could be improved, the basis index combinatorial model with a rather easy introduction of more '*ad hoc*' parameters, the term model with a more time-consuming trial-and-error strategy.

An alternative hypothesis for δ^v

In a recent paper[24] a δ^v algorithm for the bonded hydrogens has been proposed, which has, as a bonus, the possibility to 'graph'-differentiate compounds like BHF_2 and CH_2F_2, whose either simple or general graphs are identical. The algorithm is the following:

$$\delta^v = \frac{(q + f_\delta^2)\delta^v(ps)}{(pr + 1)} \quad (5)$$

Here,

$$f_\delta = n_H/\delta^v_m(ps) \quad (6)$$

where $\delta^v_m(ps)$ is the maximal $\delta^v(ps)$ value a heteroatom can have in a chemical HS general graph when all bonded hydrogens are substituted with heteroatoms, and n_H is the number of bonded hydrogen atoms. Note (i) that for a completely substituted heteroatom: $n_H = 0$, *i.e.*, $f_\delta = 0$, and eqn. (1) is retrieved, and (ii) that for saturated hydrocarbons δ and δ^v are no more equal even if $\delta^v(ps)$ equals the δ value of a simple graph. In fact, for hydrocarbons simple chemical graphs and general graphs coincide, and we have, as $p = 1$: $\delta^v = (1 + f_\delta^2)\delta$. The square power of algorithm (5) can be seen as a special case of a more

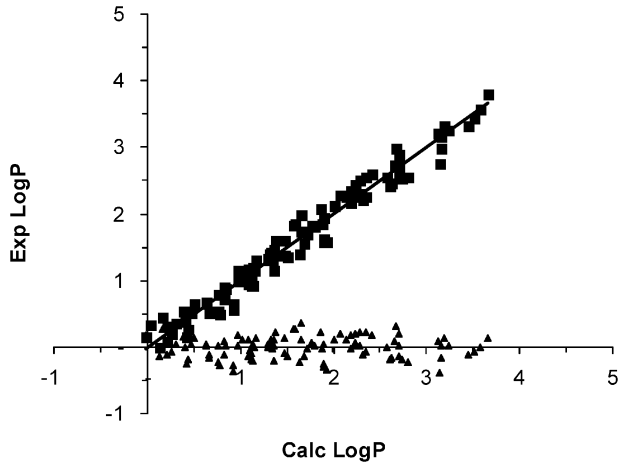

Fig. 4 Plot of the experimental *vs.* calculated log *P* (with $^{po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ descriptor), for liver, olive oil, muscle, and fat, for a total of $n = 100$ points, together with the corresponding residual plot (▲).

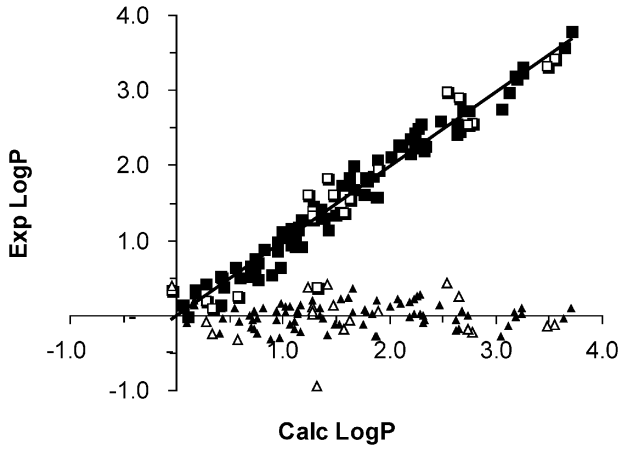

Fig. 5 Plot of the experimental *vs.* calculated log *P* (with $^{po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ descriptor), for liver, olive oil, muscle, and fat, with twenty externally validated points (open symbols), together with the corresponding residual plot (▲).

Downloaded by University of California - Los Angeles on 01 January 2013
Published on 08 July 2005 on http://pubs.rsc.org | doi:10.1039/B506091P

View Article Online

Table 7 The F-r-s_R values for the six individual log P cases when algorithm (5) for δ^v is used

$f_\delta{}^2$	Saline F-r-s_R	Blood F-r-s_R	Liver F-r-s_R	Oil F-r-s_R	Muscle F-r-s_R	Fat F-r-s_R
0	26-0.917-2.1	38-0.941-1.9	52-0.955-1.8	99-0.976-1.8	41-0.944-1.7	150-0.984-2.1
$\neq 0$	19-0.888-1.9	53-0.956-2.1	81-0.970-2.3	219-0.989-2.5	48-0.951-1.8	283-0.991-2.5

general case where f_δ is elevated to the nth power. Preliminary calculations[25] show that with algorithm (5) the combination $^{ppo}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ gives rise to the statistics of Table 7. Here, throughout the line $f_\delta = 0$ are, once again, shown the statistics for the hypothesis analysed in this paper.

Conclusions

The importance of the $^1\chi^v$ basis index as single descriptor throughout the studied cases is rewarding. This additive bond parameter, which encodes information on π, non-bonding, and core electrons, may be seen as the key physical parameter of the model. On the other side the zeroth-order type of basis indices are quite relevant in many LCBIs and in all X terms, *i.e.*, the zeroth-order index, $^0\chi$, a pure graph index, and the zeroth-order I-state index, $^0\psi_I$-K_p-(p-odd). This last index is the only index of the LCBI which is dependent on a K_p representation for the core electrons. The following results should also be underlined: (i) the relevance of the zeroth-order dual basis index $^0\chi_d^v$ in improving the description for the liver and fat case (with $^{ppo}X'$ terms), (ii) the worsening of the model due to the *cis*-ethylene compound, which is in line with an earlier result, which says that molecular connectivity indices work better with *trans*-compounds,[17] and finally (iii) the main aim of this work has been achieved, *i.e.*, to describe similar properties of similar compounds with similar descriptors. Here it is worth noting the fact that the K_p-(p-odd) algorithm is optimal for the linear combination of basis indices (LCBI), while the K_p-(pp-odd) algorithm is, instead, preferred by the molecular connectivity X terms. Even if the fundamental role of the odd complete graph conjecture for the core electrons seems, here, to be confirmed, sequential complete graphs cannot be underestimated as, in three cases, the Sal, Bl, and K_M cases they give rise to the best X term.

The theory behind the concept of partition coefficient is rather convoluted, but some rules of thumb have been proposed to understand qualitatively the trends observed with the partition coefficient.[10] These rules are centered on van der Waals interactions and H-bonding, and more precisely on concepts like apolar, monopolar, and dipolar, which describe the hydrogen-bond polarity, not to be confused with permanent dipole moment. Basis sub-graph indices, D_{Cl}, D_{Br}, and D_F represent nothing other than the number of chlorine, bromine, and fluorine atoms, which play an important role in determining the polar/apolar character of a compound. The X terms, either pp-odd or pp-sequential, are highly dependent on the pseudograph-complete-graph indices, S_F, S_{Cl}, D^v, as well as on pure graph basis indices, D_{Cl}, D_{Br}, $^0\chi$, and $^1\chi$. The electrotopological state indices, S_F, and S_{Cl} describe in a direct way the electronic environment of F and Cl atoms, while D^v gives an overall electronic information about the compound. The pure graph basis indices, $^0\chi$ and $^1\chi$, underline, instead, the importance of the simple graph structure of the molecule. The description of the partition coefficient of the examined compounds resides, then, in a mixing of pure graph, general graph and complete graph characteristics.

The model of the partition coefficients with mathematical methods has awakened the interest of other groups also.[26–31] Thus, models were published about specific sets of partition coefficients, which either use combinations of χ-non-K_p indices plus other non-χ indices. Furthermore, in these cited QSAR studies no attempt has been made to model the different sets of log P with a common descriptor, or to include the information about the hydrogen atoms.

References

1 L. Pogliani, *New J. Chem.*, 2003, **27**, 919.
2 L. Pogliani, *J. Comput. Chem.*, 2003, **24**, 1097.
3 L. Pogliani, *Int. J. Quant. Chem.*, 2005, **102**, 38.
4 L. Pogliani, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 42.
5 L. Pogliani, *Croat. Chem. Acta*, 2004, **77**, 193.
6 F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 2nd edn, 1971.
7 L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York, 1986.
8 L. B. Kier and L. H. Hall, *Molecular Structure Description. The Electrotopological State*, Academic Press, New York, 1999.
9 M. L. Cargas, P. G. Seybold and M. E. Andersen, *Toxicol. Lett.*, 1988, **43**, 235.
10 K.-U. Gass and R. P. Schwarzenbach, *J. Chem. Educ.*, 2003, **80**, 450.
11 M. Randić, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
12 A. T. Balaban, *Chemical Applications of Graph Theory*, Academic Press, New York, 1986.
13 A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, 1985, **29**, 334.
14 M. Randić and N. Trinajstić, *Croat. Chem. Acta*, 1989, **185**, 187.
15 N. Trinastić, *Chemical graph Theory*, CRC Press, Boca Raton, FL, 1992.
16 F. Harary and R. Read, *Proceedings of the Graphs and Combinatorics Conference*, George Washington University, Springer, New York, 1977, cited in J. D. Barrow, *The Book of Nothing*, Vintage Books, New York, 2000, p. 155.
17 L. Pogliani, *Chem. Rev.*, 2000, **100**, 3827.
18 L. Pogliani, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 836.
19 L. Pogliani, *J. Phys. Chem.*, 2000, **104**, 9029.
20 L. Pogliani, *J. Mol. Struct. (THEOCHEM)*, 2002, **581**, 87.
21 R. Carbó-Dorca, D. Robert, Ll. Amat, X. Girones and E. Besalu, *Molecular Quantum Similarity in QSAR and Drug Design*, Springer, Berlin, 20000.
22 Z. Mihalić, S. Nikolić and N. Trinastić, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 28.
23 S. C. Peterangelo and P. G. Seybold, *Int. J. Quant. Chem.*, 2004, **96**, 1.
24 L. Pogliani, *Adv. Quant. Chem.*, in press.
25 L. Pogliani, private results.
26 B. D. Gute, G. D. Grunwald and S. C. Basak, *SAR QSAR Environ. Res.*, 1999, **10**, 1.
27 S. C. Basak, B. D. Gute, D. Mills and D. M. Hawkins, *J. Mol. Struct. (THEOCHEM)*, 2002, **622**, 127.
28 S. C. Basak, D. Mills, D. M. Hawkins and H. A. El-Masri, *SAR QSAR Environ. Res.*, 2002, **13**, 649, and references therein.
29 S. C. Basak, B. D. Gute, D. Mills and D. M. Hawkins, *J. Mol. Struct. (THEOCHEM)*, 2002, **622**, 127.
30 P. G. Seybold, *SAR QSAR Environ. Res.*, 1999, **10**, 101.
31 B. Ren, *J. Comput. Aided Mol. Des.*, 2003, **17**, 607.

Published on 08 July 2005 on http://pubs.rsc.org | doi:10.1039/B506091P
Downloaded by University of California - Los Angeles on 01 January 2013