

Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity

Yovani Marrero-Ponce,^{a,b,*} Ricardo Medina-Marrero,^b Francisco Torrens,^c
Yamile Martinez,^a Vicente Romero-Zaldivar^d and Eduardo A. Castro^e

^aDepartment of Pharmacy, Faculty of Chemical-Pharmacy, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba

^bDepartment of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba

^cInstitut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E-46100 Burjassot (València), Spain

^dFaculty of Informatics, University of Cienfuegos, Cienfuegos, 55500 Cuba

^eINIFTA, División Química Teórica, Suc.4, C.C. 16, La Plata 1900, Buenos Aires, Argentina

Received 13 December 2004; accepted 9 February 2005

Abstract—The TOPological MOlecular COMputer Design (TOMOCOMD-CARDD) approach has been introduced for the classification and design of antimicrobial agents using computer-aided molecular design. For this propose, atom, atom-type, and total quadratic indices have been generalized to codify chemical structure information. In this sense, stochastic quadratic indices have been introduced for the description of the molecular structure. These stochastic fingerprints are based on a simple model for the intramolecular movement of all valence-bond electrons. In this work, a complete data set containing 1006 antimicrobial agents is collected and presented. Two structure-based antibacterial activity classification models have been generated. The models (including nonstochastic and stochastic indices) classify correctly more than 90% of 1525 compounds in training sets. These models permit the correct classification of 92.28% and 89.31% of 505 compounds in an external test sets. The TOMOCOMD-CARDD approach, also, satisfactorily compares with respect to nine of the most useful models for antimicrobial selection reported to date. Finally, a virtual screening of 87 new compounds reported in the anti-infective field with antibacterial activities is developed showing the ability of the TOMOCOMD-CARDD models to identify new leads as antibacterial.

© 2005 Published by Elsevier Ltd.

1. Background

The introduction of antibiotics in the 1940s was thought to have eliminated the scourge of all infectious diseases. However, due to the widespread use and misuse of antibiotics, bacterial resistance to antibiotics has become a serious public health problem. Some of these resistant strains, such as vancomycin-resistant enterococci (VRE) and multidrug resistant *Staphylococcus aureus* (MRSA), are capable of surviving the effects of most, if not all, antibiotics currently in use.^{1–19} This recent increase in resistant bacterial infections has created a crit-

ical need to develop novel antibacterial drugs that elude existing mechanisms of resistance. For this reason, many researchers worldwide have been interested in the search and evaluation of novel lead antibacterial compounds.^{20–35}

Because the experimental tests (based on ‘trial and error’ screening), which must be performed, especially pharmacological and toxicological test, are usually expensive and time consuming, during the past two decades the pharmaceutical industry has reoriented its research strategy to the development of methods enabling rational selection or design of novel agents with the desired properties. In addition, the great quantity of chemicals synthesized in laboratories (many of which have not found biological or pharmaceutical application) and the broad diversity of possible chemical structures to the test imposed on us the necessity for development of an alternative technique to classical ‘trial and error’ screenings.

Keywords: TOMOCOMD-CARDD software; Nonstochastic and stochastic quadratic indices; Classification model; LDA-based QSAR; Antibacterial activity.

* Corresponding author. Tel.: +53 42 281192/281473; fax: +53 42 281130/281455; e-mail addresses: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es

In particular, the search of antibacterial compounds has always been on the desktop of molecular modeling and drug design specialists. In spite of this intensive search, the discovery of selective antibacterial agents has remained a largely elusive goal of antimicrobial research. Subsequently, new approaches are needed in order to make an efficient search for candidates to be assayed as antibacterial drugs.

In this sense, several *in silico* methods have been used to develop QSARs on antimicrobial activity.^{36–41} The effort in this area has been placed mainly into the development of structure-based classification methods, utilizing pattern recognition techniques to predict biologically active molecules. This is a logic behavior, because compounds with antibacterial activity are structurally diverse and are the class of pharmaceuticals that clearly do not share a common mode of action.⁴² In these approaches, several pattern recognition techniques have been already applied to examine structural features that underlie patterns associated with biological effects of antimicrobial agents.^{36–41} Among these methods, we can point out the linear discriminant analysis (LDA), the binary logistic regression analysis (BLR), and the artificial neural networks (ANNs). Many 2D-physicochemical and structural descriptors were calculated in these studies to classify the compounds into active (antibacterial) or inactive ones.^{36–41} However, in all cases the spectrum of structural patterns (diversity of chemical families) considered was small. In one of these works, Tomás-Vert et al.³⁷ developed a set of algorithms that make it possible to calculate as many as 62 descriptors for each structure directly from a novel matrix representation (combined matrix).

In the context of novel methods based on chemical graph and algebra theory developed for modeling physicochemical and biological properties, our research group has recently introduced the novel computer-aided molecular design scheme TOMOCOMD-CARDD.⁴³ The TOPological MOlecular COMputer Design-Computer Aided 'Rational' Drug Design (TOMOCOMD-CARDD) has been applied to the description of physicochemical and biological properties of chemicals and drugs.^{44–56,61,62} The TOMOCOMD-CARDD strategy is very useful for the selection of novel subsystems of compounds having a desired property/activity, which can be further optimized by using some of the many molecular modeling methods at the disposition of the medicinal chemists. In this sense, it was successfully applied to the virtual (computational) screening of novel anthelmintic compounds, which were then synthesized and *in vivo* evaluated on *Fasciola hepatica*.^{44,45}

Studies for the *in silico* screening and design of leads paramphistomocides were also conducted with this theoretical approach.⁴⁶ It was also able to identify 100% of Ras farnesyl transferase (FTase) inhibitors in a simulated virtual screening experiment.⁴⁷ By using the same and similar models we were able to design a set of aryl-aminomethylenemalonates compounds with antimalarial activity.^{47,48} These compounds were then synthesized and tested for antimalarial activity in two different *Plasmodium falciparum* strains. Chemicals predicted as

inactive for the models were also prepared and tested showing lack of activity in the two cellular lines.

The prediction of the pharmacokinetics properties of organic compounds is a problem that can also be addressed using this approach. In this sense, this method has been used to estimate the intestinal-epithelial transport of drug in Caco-2 cell culture of a heterogeneous series of drug-like compounds.^{49–51} The obtained results suggest that the TOMOCOMD-CARDD method is able to predict the permeability values and it proved to be a good tool for studying the oral absorption of drug candidates during the drug development process.

In another experiments with TOMOCOMD-CARDD several physical, physicochemical, and chemical properties of organic compounds were effectively modeled.^{52–55} In addition, TOMOCOMD-CARDD has been extended to consider three-dimensional features of small/medium sized molecules based on the trigonometric 3D-chirality correction factor approach.⁵⁶

Chemical graph theory^{57,58} and graph-theoretic molecular descriptors^{59,60} have been criticized by several authors due to their 'over-simplification' of the molecular structure as well as their lack of physical meaning. In this sense, the latter opportunity has allowed the description of the significance interpretation and the comparison to other molecular descriptors.^{53,55} The approach describes changes in the electron distribution with time throughout the molecular backbone. Specifically, the features of the *k*th total and local quadratic and linear indices were illustrated by examples of various types of molecular structures, including chain-lengthening, branching, heteroatoms-content, and multiple bonds. Additionally, the linear independence of the atom-type quadratic and linear fingerprints to other 229 0D-3D 'DRAGON' molecular descriptors was demonstrated. That is, it was concluded that the local fingerprints are an independent indices containing important structural information to be used in QSPR/QSAR and drug design studies.^{53,55}

Finally, the method is very flexible and makes also possible the study of macromolecules such as protein and nucleic acid.^{61,62} The calculation of several macromolecular fingerprints has been implemented in two different subprograms of the TOMOCOMD software: CANAR (Computed-Aided Nucleic Acid Research) and CAMPS (Computed-Aided Modeling in Protein Science). These macromolecular descriptors have been successfully employed in studies related to proteomics and nucleic acid-drug interactions. In this sense, promising results have been found in the modeling of the interaction between paromomycin and HIV-1Ψ-RNA packaging region in the field of bioinformatics⁶¹ and in the prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor.⁶²

In the current work we continue by testing the possibilities of this approach for modeling and predicting activities of structurally diverse organic compounds. The main objectives of this paper are first to find rationality

in the search of novel antibacterial drug-like compounds using the TOMOCOMD-CARDD approach, and second but not less important, to introduce the stochastic molecular quadratic indices as a novel component of the TOMOCOMD-CARDD scheme.

2. Theoretical framework

2.1. Atom, atom-type, and total nonstochastic quadratic fingerprints

Implemented in the subprogram CARDD of the TOMOCOMD software, the quadratic indices (nonstochastic) can be calculated from both molecular pseudograph's atom adjacency matrix and molecular vector of small-to-medium sized organic compounds. The general principles of the quadratic indices have been explained in some detail elsewhere.^{44,46,47,49–53,55,56,61,62} However, an overview of this approach will be given. For a given molecule composed of n atoms, the 'molecular vector' (X) is constructed and the k th total quadratic indices, $q_k(x)$ are calculated as quadratic forms as shown in Eq. 1,

$$q_k(x) = \sum_{i=1}^n \sum_{j=1}^n {}^k a_{ij} x_i x_j \quad (1)$$

where, n is the number of atoms of the molecule and x_1, \dots, x_n are the coordinates or components of the 'molecular vector' (X) in a system of canonical basis vectors of \mathfrak{R}^n . The components of the molecular vector are numeric values, which can be considered as weights (atom-labels) for the vertices of the pseudograph. Certain atomic properties (electronegativity, atomic radii, etc) can be used with this propose. In this work Pauling electronegativities are selected as atom weights.⁶³

The coefficients ${}^k a_{ij}$ are the elements of the k th power of the symmetric square matrix $\mathbf{M}(G)$ of the molecular pseudograph (G) and are defined as follows:

$$\begin{aligned} a_{ij} &= P_{ij} \text{ if } i \neq j \text{ and } \exists e_k \in E(G) \\ &= L_{ii} \text{ if } i = j \\ &= 0 \text{ otherwise} \end{aligned} \quad (2)$$

where, $E(G)$ represents the set of edges of G . P_{ij} is the number of edges (bonds) between vertices (atoms) v_i and v_j , and L_{ii} is the number of loops in v_i (see Table 1).

Eq. 1 for $q_k(x)$ can be written as the single matrix equation:

$$q_k(x) = \mathbf{X}^t \mathbf{M}^k \mathbf{X} \quad (3)$$

where \mathbf{X} is a column vector (a $n \times 1$ matrix), \mathbf{X}^t the transpose of \mathbf{X} (a $1 \times n$ matrix) and \mathbf{M}^k the k th power of the matrix \mathbf{M} of the molecular pseudograph G (mathematical quadratic form's matrix).

In addition to total quadratic indices, computed for the whole-molecule, local-fragment (atom and atom-type) formalisms can be developed. These descriptors are termed local quadratic indices, $q_{kL}(x)$.^{44,46,47,49–53,55,56,61,62} The definition of these descriptors is as follows:

$$q_{kL}(x) = \sum_{i=1}^m \sum_{j=1}^m {}^k a_{ijL} x_i x_j \quad (4)$$

where, m is the number of atoms of the fragment of interest and ${}^k a_{ijL}$ is the element of the row ' i ' and column ' j ' of the matrix \mathbf{M}_L^k . This matrix is extracted from the \mathbf{M}^k matrix and contains the information referred to the vertices (atoms) of the specific molecular fragments and also of the molecular environment. The matrix $\mathbf{M}_L^k = [{}^k a_{ijL}]$ with elements ${}^k a_{ijL}$ is defined as follows:

$$\begin{aligned} {}^k a_{ijL} &= {}^k a_{ij} \text{ if both } v_i \text{ and } v_j \text{ are atoms} \\ &\quad \text{contained within the molecular fragment} \\ &= 1/2 {}^k a_{ij} \text{ if } v_i \text{ or } v_j \text{ is an atom contained} \\ &\quad \text{within the molecular fragment but not both} \\ &= 0 \text{ otherwise} \end{aligned} \quad (5)$$

These local analogues can also be expressed in matrix form by the expression

$$q_{kL}(x) = \mathbf{X}^t \mathbf{M}_L^k \mathbf{X} \quad (6)$$

Note that the above scheme follows the spirit of a Mulliken population analysis.⁶⁴ Also note that for every partitioning of a molecule into Z molecular fragment there will be Z local molecular fragment matrices. In this case, if a molecule is partitioned into Z molecular fragments, the matrix \mathbf{M}^k can be partitioned into Z local matrices \mathbf{M}_L^k , $L = 1, \dots, Z$, and the k th power of matrix \mathbf{M} is exactly the sum of the k th power of the local Z matrices. In this way, the total quadratic indices are the sum of the quadratic indices of the Z molecular fragments:

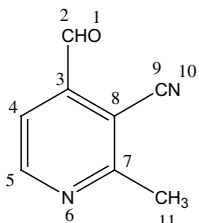
$$q_k(x) = \sum_{L=1}^Z q_{kL}(x) \quad (7)$$

Atom and atom-type quadratic indices are specific cases of local quadratic indices. In this sense, the k th atom-type quadratic indices are calculated by summing the k th atom quadratic indices of all atoms of the same atom type in the molecule.

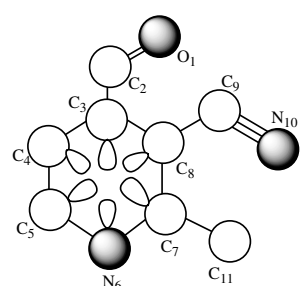
In the atom-type quadratic indices formalism, each atom in the molecule is classified into an atom-type (fragment), such as heteroatoms, heteroatoms H-bonding to heteroatoms (O, N, and S), halogens, aliphatic carbon chain, aromatic atoms (aromatic rings), and so on. For all data sets, including those with a common molecular scaffold as well as those with very diverse structure, the k th atom-type quadratic indices provide important information.

As mentioned above, the present approach codifies 3D information by the introduction of a local *trigonometric 3D-chirality correction factor* in molecular vector, X .⁵⁶ In these sense, a chirality molecular vector is obtained ($*X$), where the components of X (for instance, Pauling electronegativity (x_A)⁶³ of the atom A) are substituted by the following term $[x_A + \sin((\omega_A + 4\Delta)\pi/2)]$. The trigonometric 3D-chirality correction factor use a dummy variable, ω_A and an integer parameter, Δ .⁵⁶

Table 1. Calculation of $M^k(G)$ and $S^k(G)$ for 2-formyl-6-methyl-benzonitrile When k varies from 0 to 2 and i is a specific atom in the molecule



Molecular Structure



Molecular Pseudograph (G)

a_{ij}	O ₁	C ₂	C ₃	C ₄	C ₅	N ₆	C ₇	C ₈	C ₉	N ₁₀	C ₁₁	$k \delta_i$	O ₁	C ₂	C ₃	C ₄	C ₅	N ₆	C ₇	C ₈	C ₉	N ₁₀	C ₁₁
$M^0(G)$													$S^0(G)$										
O ₁	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
C ₂	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
C ₃	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
C ₄	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
C ₅	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
N ₆	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
C ₇	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
C ₈	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
C ₉	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0
N ₁₀	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0
C ₁₁	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
$M^1(G)$													$S^1(G)$										
O ₁	0	2	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0
C ₂	2	0	1	0	0	0	0	0	0	0	0	3	0.66	0	0.33	0	0	0	0	0	0	0	0
C ₃	0	1	1	1	0	0	0	1	0	0	0	4	0	0.25	0.25	0.25	0	0	0	0.25	0	0	0
C ₄	0	0	1	1	1	0	0	0	0	0	0	3	0	0	0.33	0.33	0.33	0	0	0	0	0	0
C ₅	0	0	0	1	1	1	0	0	0	0	0	3	0	0	0	0.33	0.33	0.33	0	0	0	0	0
N ₆	0	0	0	0	1	1	1	0	0	0	0	3	0	0	0	0	0.33	0.33	0.33	0	0	0	0
C ₇	0	0	0	0	0	1	1	1	0	0	1	4	0	0	0	0	0	0.25	0.25	0.25	0	0	0.25
C ₈	0	0	1	0	0	0	1	1	1	0	0	4	0	0	0.25	0	0	0	0.25	0.25	0.25	0	0
C ₉	0	0	0	0	0	0	0	1	0	3	0	4	0	0	0	0	0	0	0	0.25	0	0.75	0
N ₁₀	0	0	0	0	0	0	0	0	3	0	0	3	0	0	0	0	0	0	0	0	1	0	0
C ₁₁	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
$M^2(G)$													$S^2(G)$										
O ₁	4	0	2	0	0	0	0	0	0	0	0	6	0.66	0	0.33	0	0	0	0	0	0	0	0
C ₂	0	5	1	1	0	0	0	1	0	0	0	8	0	0.625	0.125	0.125	0	0	0	0.125	0	0	0
C ₃	2	1	4	2	1	0	1	2	1	0	0	14	0.143	0.071	0.287	0.143	0.071	0	0.071	0.143	0.071	0	0
C ₄	0	1	2	3	2	1	0	1	0	0	0	10	0	0.1	0.2	0.3	0.2	0.1	0	0.1	0	0	0
C ₅	0	0	1	2	3	2	1	0	0	0	0	9	0	0	0.111	0.222	0.333	0.222	0.111	0	0	0	0
N ₆	0	0	0	1	2	3	2	1	0	0	1	10	0	0	0	0.1	0.2	0.3	0.2	0.1	0	0	0.1
C ₇	0	0	1	0	1	2	4	2	1	0	1	12	0	0	0.083	0	0.083	0.166	0.333	0.166	0.083	0	0.083
C ₈	0	1	2	1	0	1	2	4	1	3	1	16	0	0.063	0.125	0.063	0	0.063	0.125	0.25	0.063	0.188	0.063
C ₉	0	0	1	0	0	0	1	1	10	0	0	13	0	0	0.077	0	0	0	0.077	0.077	0.769	0	0
N ₁₀	0	0	0	0	0	0	0	3	0	9	0	12	0	0	0	0	0	0	0	0.25	0	0.75	0
C ₁₁	0	0	0	0	0	1	1	1	0	0	1	4	0	0	0	0	0	0.25	0.25	0.25	0	0	0.25

$\omega_A = 1$ and Δ is an odd number when A has
 R (rectus), E (entgegen), or a (axial) notation
 according to Cahn–Ingold–Prelog rules
 $= 0$ and Δ is an even number, if A does not
 have 3D specific environment
 $= -1$ and Δ is an odd number when A has
 S (sinister), Z (zusammen), or e (equatorial)
 notation according to Cahn–Ingold–
 Prelog rules

(8)

Thus, this 3D-chirality factor $\sin((\omega_A + 4\Delta)\pi/2)$ takes different values in order to codify specific stereochemical information such as chirality, Z/E isomerism, and so on. This factor therefore takes values in the following order $1 > 0 > -1$ for atoms that have specific 3D environments. The chemical idea here is not that the attraction of electrons by an atom depends on their chirality, due to experience shows that chirality does not change the electronegativities of atoms in the molecule in an isotropic environment in an observable way.⁶⁵ This correction has principally a mathematical means and must not be source of any misunderstanding. The present trigono-

metric 3D-chiral correction factor is invariant with respect to the selection of other chirality scales^{66–68} and gets ever the values 1, 0, and -1 for $R = E = a$, nonchiral = no Z/E isomerism involved = no a/e substitution, and $S = Z = e$ atoms, respectively.⁵⁶

A very interesting point is that the present 3D-chiral descriptor reduces to simple (2D) quadratic indices ones for molecules without specific 3D characteristics because $\sin(0 + 4\Delta)\pi/2 = 0$, being Δ zero or any even number. That is, when all the atoms in the molecule are not chiral, the TOMOCOMD-CARDD molecular descriptors do not change upon the introduction of this factor. This means that $*X = X$ and thus, $*q_k(x) = q_k(x)$.⁵⁶

2.2. Atom, atom-type, and total stochastic quadratic fingerprints

Note that the mathematical quadratic form's matrices, \mathbf{M}^k , are graph-theoretic electronic-structure models, like an 'extended Hückel' model. The \mathbf{M}^1 matrix considers all valence-bond electrons (σ - and π -networks) in one step and their power ($k = 0, 1, 2, 3, \dots$) can be considering as an interacting-electron chemical-network model in k step. This model can be seen as an intermediate between the quantitative quantum-mechanical Schrödinger equation and classical chemical bonding ideas.⁶⁹

Recently, our research group has also developed a new method based on the Markov chain theory, which has been successfully employed in QSPR and QSAR studies.^{70,71} This approach also describes changes in the electron (stochastic) distribution and vibrational decay with time throughout the molecular backbone using Markov chain formalism.

The present approach is based on a simple model for the intramolecular (stochastic) movement of all valence-bond electrons. Let us consider a hypothetical situation in which a set of atoms is free in space at an arbitrary initial time (t_0). In this time, the electrons are distributed around atom nucleus. Alternatively, these electrons can be distributed around cores in discrete intervals of time t_k . In this sense, the electron in an arbitrary atom i can move to other atoms at different discrete time periods t_k ($k = 0, 1, 2, 3, \dots$) throughout the chemical-bonding network.

The k th stochastic molecular pseudograph's atom adjacency matrix $[\mathbf{S}^k(G)]$ can be obtained from \mathbf{M}^k . Here, $\mathbf{S}^k(G) = \mathbf{S}^k = [{}^k s_{ij}]$, is a squared table of the order n (n = number of atoms) and the elements ${}^k s_{ij}$ are defined as follows:

$${}^k s_{ij} = \frac{{}^k a_{ij}}{{}^k \text{SUM}_i} = \frac{{}^k a_{ij}}{{}^k \delta_i} \quad (9)$$

where, ${}^k a_{ij}$ are the elements of the k th power of \mathbf{M} and the SUM of the i th row of \mathbf{M}^k are named the k -order vertex degree of atom i , ${}^k \delta_i$. The k th s_{ij} elements are the transition probabilities when the electrons move from atom i to j in the discrete time periods t_k . Note, that

k th element s_{ij} take into consideration the molecular topology in k step throughout the chemical-bonding (σ - and π -) network. For instance, the ${}^2 s_{ij}$ values can distinguish between hybrid states of atoms in bonds. In this sense, It can clearly be seen from Table 1 that electrons will have a higher probability of returning to the sp nitrogen ($N_{10} = 0.75$) than to the sp_2 nitrogen ($N_6 = 0.33$) in t_2 . A similar behavior can be observed among the different hybrid states of carbon atom in the molecule of 2-formyl-6-methyl-benzonitrile (see Table 1): Csp_3 ($C_{11} = 0.25$); Csp_2 ($C_2 = 0.625$); $\text{Csp}_{2\text{arom}}$ ($C_3 = 0.285$, $C_4 = 0.3$, $C_5 = 0.33$, $C_7 = 0.33$, $C_8 = 0.25$); and Csp ($C_9 = 0.769$). This is a logical result if the electronegativity scale of these hybrid states is taken into account. The k th total and local stochastic quadratic indices, ${}^s q_k(x)$ are calculated in the same way that the quadratic indices (nonstochastic), but using k th stochastic molecular pseudograph's atom adjacency matrix, $\mathbf{S}^k(G)$, like mathematical quadratic form's matrices.

3. Computational strategies and data sets

3.1. Computational methods

TOMOCOMD is an interactive program for molecular design and bioinformatics research, developed upon the base of a user-friendly philosophy.⁴³ It is composed of four subprograms: 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), each one of them allows to draw the structures and to calculate molecular 2D and 3D indices. In this paper, we outline salient features concerned with only the module CARDD and the calculation of nonstochastic and stochastic quadratic indices.

The calculation of total and local (both atom and atom-type) quadratic indices for any organic molecule was implemented in the TOMOCOMD-CARDD software.⁴³ The main steps for the application of this method in QSAR/QSPR and drug design can be briefly resumed as follows:

1. Draw the molecular pseudographs for each molecule of the data set, using the software drawing mode. This procedure is performed by a selection of the active atom symbol belonging to different groups of the periodic table.
2. Use appropriate weights in order to differentiate the molecular atoms. In this work, we used as atomic property the Paulin electronegativity for each kind of atom.⁶³
3. Compute the total and local quadratic indices of the molecular pseudograph's atom adjacency matrix. They can be carried out in the software calculation mode, where you can select the atomic properties and the family descriptor previously to calculate the molecular indices. This software generates a table in which the rows correspond to the compounds and

columns correspond to the total and local quadratic indices or other family molecular descriptors implemented in this program.

- Find a QSPR/QSAR equation by using mathematical techniques, such as multilinear regression analysis (MRA), neural networks (NN), linear discrimination analysis (LDA), and so on. That is to say, we can find a quantitative relation between a activity A and the quadratic indices having, for instance, the following appearance:

$$A = a_0q_0(x) + a_1q_1(x) + a_2q_2(x) + \cdots + a_kq_k(x) + c \quad (10)$$

where A is the measurement of the activity, $q_k(x)$ is the k th total quadratic indices, and the a_k 's are the coefficients obtained by regression analysis.

- Test the robustness and predictive power of the QSPR/QSAR equation by using internal and external cross-validation techniques.

The descriptors calculated were the following:

- $q_k(x)$ and $q_k^H(x)$ are the k th total quadratic indices not considering and considering H-atoms in the molecular pseudograph (G), respectively.
- $q_{kL}(x_E)$ and $q_{kL}^H(x_E)$ are the k th local (atom-type = heteroatoms: S, N, O) quadratic indices not considering and considering H-atoms in the molecular pseudograph (G), respectively.
- $q_{kL}^H(x_{E-H})$ are the k th local (atom-type = H-atoms bonding to heteroatoms: S, N, O) quadratic indices considering H-atoms in the molecular pseudograph (G).

The k th stochastic total [$^s q_k(x)$ and $^s q_k^H(x)$] and local [$^s q_{kL}(x_E)$, $^s q_{kL}^H(x_E)$, and $^s q_{kL}^H(x_{E-H})$] quadratic indices were also computed.

3.2. Data set selection

The general performance of the current method decisively depends on the selection of compounds for the training series used to build the classifier model. The most critical aspect of the construction of the training set is to warranty a great molecular diversity in this data set. With the purpose of guarantee this molecular diversity we have selected a data set composed by a great number of molecular entities, some of them reported as antibacterials^{72–74} and the rest with a series of other pharmacological uses.^{72,73}

In this study, we consider a general data set made up of 2030 compounds, 1006 with antibacterial properties and different action modes, and 1024 having other clinical uses (antivirals, sedative/hypnotics, diuretics, anticonvulsants, hemostatics, oral hypoglucemics, antihypertensives, antihelminthics, anticancer compounds, and so on). From these 2030 compounds, 1525 were chosen at random to forming the training set, being 754 of them actives (antibacterial) and 771 inactive ones. The great structural variability of the selected training data set makes it possible, not only the discovery of lead com-

pounds with determined mechanisms of antibacterial activity, but also with novel modes of action. It will be well-illustrated in this paper in a virtual experiment for lead generation.

The resting group composed of 252 antibacterial and 253 compounds with different biological properties was prepared as test data set for the validation of the models. These 505 compounds were never used in the development of the classification models. Finally, an external cross-validation set of 87 novel antimicrobial agents was also used in order to assess the predictive ability of the obtained classification models.

3.3. Chemometric analysis

Continuing from the previous section, we can try to develop a simple linear QSAR using TOMOCOMD-CARDD method using the general formula depicted in Eq. 10. The statistical analysis was carried out with the STATISTICA software.⁷⁵ The tolerance parameter (proportion of variance that is unique to the respective variable) used was the default value for minimum acceptable tolerance, which is 0.01. Forward stepwise was fixed as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as strategy for model selection. In its original form, the Occam's razor states that 'Numquam ponenda est pluritas sin necessitate', which can be translated as 'Entities should not be multiplied beyond necessity'.⁷⁶ In this case simplicity is loosely equated with the number of parameters in the model. If we understand predictive error to be the error rate for unseen examples, the Occam's razor can be stated for the selection of QSAR/QSPR models as ('QSAR/QSPR Occam's Razor'): Given two QSAR/QSPR models with the same predictive error, the simpler one should be preferred because simplicity is desirable in itself.⁷⁶ In this connection, we select the model with higher statistical signification but having as few parameters (a_k) as possible.

In Eq. 10 a_k are the coefficients of the classification function, determined by least square method as implemented in linear discriminant analysis (LDA) modulus of STATISTICA 99.⁷⁵ Forward stepwise was fixed as the strategy for variable selection.

Table 2. Global results of the classification of compounds in the training and prediction sets

Serie	Nonstochastic descriptors			Stochastic descriptors		
	% Correct	(–)	(+)	% Correct	(–)	(+)
<i>Training set</i>						
(–)	95.07	733	38	91.05	702	69
(+)	90.19	74	680	89.66	78	676
Total	92.66	807	718	90.36	780	745
<i>Test set</i>						
(–)	94.07	238	15	90.12	228	25
(+)	90.48	24	228	88.50	29	223
Total	92.28	262	243	89.31	257	248

Table 3. Names and classification (Eqs. 11 and 12) of some compounds in training set

Active compound names	$\Delta P\%$ ^a		Inactive compound names	$\Delta P\%$ ^a	
	Nonstochastic	Stochastic		Nonstochastic	Stochastic
Mefuralazine	95.13	92.29	Amantadine	−93.42	−87.95
Antibiotic B-5050-A	98.75	98.42	Litracen	−99.76	−95.95
Antibiotic 810 A1	99.42	95.19	Befuraline	−81.03	−72.98
Amikacin	99.85	99.82	Pizotifen	−98.11	−93.89
Mepartricin A	99.90	97.41	Fentanilo	−92.49	−85.72
Terramycin X	98.47	98.48	Elanzepine	−97.02	−90.24
Quatrimycin	95.49	97.48	Dimepranol	−94.20	−89.11
Texazone	44.11	32.12	Mezepine	−97.93	−94.72
Polymyxin B1	99.96	99.98	Crufomate	−79.01	−83.21
Amphotericin B	99.86	96.73	Ascaridole	−75.61	−81.66
Rifampicin	99.64	99.38	Actractil	−86.68	−85.79
Furaguanidine	62.03	40.70	B-arteether	−65.25	−67.02
Rifamycin	97.64	88.87	Oxolin	32.11	−26.69
Tetramycin	97.24	69.43	Nitroguanil	−46.17	−48.03
Globomycin	92.50	88.78	Atovaquone	−81.26	−57.26
Superciclin'Farmabion'	99.29	98.92	Nitronal	−94.18	−77.77
Allicin	−86.03	−79.83	Brometenamine	−97.11	−97.46
Streptomycin B	100.00	99.97	Nitrosorbide	−44.75	73.50
Disulformin	100.00	99.95	Stenopril	−97.91	−94.81
Glucosulfone	100.00	100.00	Oxazidione	−91.95	−69.80
Nifuraldezone	89.65	75.77	Anisindione	−87.92	−73.27
Neomycin C	99.82	99.86	Isocalcio'Erba'	−90.13	−93.56
Tetracycline	95.49	97.48	Etiron	−93.25	−93.45
Sulfamerazine	94.94	89.85	Thiophenobarbital	−86.74	−83.33
Vanepirim	99.90	98.36	Basthioryl	−97.72	−96.95
Meclocycline	97.34	98.66	Butamben	−91.73	−66.20
Antibiotic P-2563 2	92.35	80.87	Cyclopropane	−95.35	−95.27
Talmetoprim	75.45	69.45	Halothane	−76.81	−61.97
Apramycin	98.86	98.75	Agentit	−95.47	−84.13
Nitrofurantoin	79.36	67.18	Dipropamine	−93.43	−94.05
Aspoxicillin	89.58	90.83	Hemedin	−92.37	−94.44
Sulfathiadiazole	99.98	−21.85	Febensamin	−97.06	−96.90
Demecloxycline	98.11	98.62	Scyan	−94.26	−84.25
Epiroprim	79.55	78.37	Anticomman	−88.32	−98.03
Hetacillin	54.75	63.02	Carbromide	−77.13	−76.30
Spinulosin	74.73	5.75	Gliamilide	62.20	45.85
Lincomycin	61.64	46.33	Cystamine	−95.03	−88.28
Kanamycin sulfate	99.21	99.21	Paraldehyde	−93.85	−90.44
Nidroxyzone	69.20	84.33	Dextromoramide	−97.36	−90.17
Prazocillin	98.73	93.78	Aponal	−94.12	−92.27
Cefamandole	99.93	96.55	Chlorphenacemide	−87.73	−86.63
Septosil	98.52	77.26	Pytamine hydrochloride	−93.81	−89.59
Salazodine	99.85	98.92	Phenobarbital	−89.94	−90.61
Rifabutin	99.71	97.99	Primidone	−92.88	−90.23
Broxyquinoline	76.79	−42.45	Diciferron	−90.95	−91.01
Azotomycin	87.25	89.06	Naftazone	−69.16	−61.44
Chloroxine	82.98	−43.48	Tilidina	−97.41	−95.80
Griseofulvin	82.07	23.70	Clorfenoxamine	−94.29	−88.93
Amoxicillin	65.48	85.73	Moxaverine	−85.23	−77.51
Azoseptyl-K	99.10	97.63	Butaverine	−95.12	−94.96
Arsant	99.95	95.82	Phenacemide	−95.29	−93.43
Cefaclor	59.35	66.07	Ambucetamide	−87.89	−72.13
Doricin	99.00	91.32	Ethydine	−66.54	−85.04
Aspergillus acid	13.33	29.21	Tetramethrin	−87.41	−83.57
Dolamina	92.86	77.54	Stevaldil	−92.92	−96.05
Melarsenoxyd	98.84	89.44	Beclamide	−93.11	−94.28
Lasalocid B	84.37	53.42	Tolpropamine	−99.29	−97.39
Spectinomycin	88.38	72.97	Chinoin 103	−90.37	−86.30
Trimethoprim	54.58	47.79	Betaxolol	−80.08	−89.23
Carpetimycin A	90.82	88.89	Carazolol	−73.90	−64.68
Sulfametrole	95.52	93.62	Athotoin	−90.13	−85.10
Proethyl	58.68	83.70	Pilokarpin	−84.68	−81.28
Melarsen	97.56	95.74	Pheneturide	−95.37	−94.06
Thiozamin	62.11	90.04	Histapipendine	−96.58	−92.20

(continued on next page)

Table 3 (continued)

Active compound names	$\Delta P\%^a$		Inactive compound names	$\Delta P\%^a$	
	Nonstochastic	Stochastic		Nonstochastic	Stochastic
Cefazolin	99.17	99.28	Alprostadil	−42.14	−55.43
Imicillin	98.59	97.86	Papaverine	−64.05	−68.00
Astreonam	99.88	99.10	Phenoxybenzamine	−90.88	−86.53
Nifurtolone	79.90	89.58	Phenamazole	−81.91	−81.72
Chlorozotocin	55.65	90.48	Norephedrine	−91.43	−80.79
Brodinoprin	67.43	58.14	Methoxyphenamine	−95.37	−94.26
Ceftizoxime sodium	95.03	97.22	Metaxalone	−91.82	−91.70
Cinoquidox	76.34	56.24	Ferroglycine sulfate	−88.97	−93.02
Myxin	98.00	89.45	Mephesisin	−73.32	−83.03
Lenigron	87.01	62.00	Meprobamate	−97.99	−98.05
Anabial	60.93	41.63	Racefemine	−92.02	−91.97
Imipenem	11.98	20.05	Xylazine	−90.52	−90.97
Acrotiazol	99.74	83.42	Methophedrinum	−86.86	−88.75
Toyocamycin	95.65	93.94	Nandrolone	−85.17	−85.18
Rifordin	99.67	98.93	Estrone	−84.59	−80.39
Miran	93.34	68.48	Phenethylurea	−87.72	−74.01
Gluconiazide	89.55	78.68	Spiroplatin	−96.02	−90.35
Piridina'N'	99.20	95.37	Hadacidin	−80.84	−46.26
Dioxidine	56.91	67.63	Carbetimer	−89.67	−83.88
Dapsone	42.14	90.57	Tiosalicilic acid	−92.12	−53.60
Bemural	86.31	85.89	Tricloroetene	−64.56	−69.09
Reseptyl	93.68	69.62	Haloperide	−43.65	−75.46
Alfasol	99.67	87.83	Butanolium	−93.08	−93.38
Pyridinin	91.73	95.28	Dicarbene	−95.38	−92.64
Carbadox	98.24	85.89	Gemcadiol	−90.02	−93.98

^a $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$; where $P(\text{active})$ is the probability that the equations classify a compound as active. Conversely, $P(\text{inactive})$ is the probability that the models classify a compound as inactive. This value ($\Delta P\%$) takes positive values when $P(\text{active}) > P(\text{inactive})$ and negative otherwise. Therefore, when $\Delta P\%$ is positive (negative) the compound was classified as antibacterial (non-antibacterial).

The quality of the models were determined by examining Wilk's λ parameter (U -statistic), square Mahalanobis distance (D^2), Fisher ratio (F) and the corresponding p -level ($p(F)$) as well as the percentage of good classification in training and test sets. Models with a proportion between the number of cases and variables in the equation lower than 4 were rejected.

The Wilks' λ statistical helpful to value the total discrimination and can take values between 0 (perfect discrimination) and 1 (no discrimination). The D^2 indicates the separation of the respective groups. The statistical robustness and predictive power of the obtained model was assessed using an external prediction (test) set. The biological activity (antibacterial in this case) was codified by an indicator variable AMA (acronym of anti-microbial activity). This variable indicates either the presence of an active compound ($\text{AMA} = 1$) or an inactive compound ($\text{AMA} = -1$).

The classification of cases was performed by means of the posterior classification probabilities. This is the probability that the respective case belongs to a particular group (active or inactive) and it is proportional to the Mahalanobis distance from that group centroid. In closing, the posterior probability is the probability, based on our knowledge of the values of other variables, that the respective case belongs to a particular group. By using the models, one compound can be then classified as active, if $\Delta P\% > 0$, being $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$ or as inactive otherwise. $P(\text{Active})$ and $P(\text{inactive})$ are the probabilities that the equations classify a compound

as active and inactive, respectively. Finally, an external test set of 87 compounds was also used in order to assess the predictive ability of the obtained LDA models.

4. Results and discussion

4.1. Development of the classification models

The best discrimination functions obtained with nonstochastic and stochastic quadratic indices for the training set are given below, respectively:

$$\begin{aligned} \text{AMA} = & -3.6857 - 1.4689 \times 10^{-7} q_{10}^H(x) + 0.0003 q_4(x) \\ & + 0.0773 q_{0L}^H(x_E) - 0.0592 q_{1L}^H(x_E) \\ & - 0.0274 q_{2L}^H(x_E) + 0.0153 q_{3L}^H(x_E) \\ & - 0.6235 q_{0L}^H(x_{E-H}) + 0.4612 q_{1L}^H(x_{E-H}) \\ N = & 1525, \quad \lambda = 0.42, \quad D_2 = 5.47, \\ F(8.1516) = & 259.38, \quad p < 0.0001 \end{aligned} \quad (11)$$

$$\begin{aligned} \text{AMA} = & -3.4459 - 0.1896^s q_{12}(x) + 0.2191^s q_{15}(x) \\ & + 0.1195^s q_{3L}^H(x_E) - 0.1673^s q_{4L}^H(x_E) \\ & + 0.1199^s q_{15L}^H(x_E) + 0.9717^s q_{0L}^H(x_{E-H}) \\ & - 0.0164^s q_{1L}^H(x_{E-H}) - 0.9789^s q_{6L}^H(x_{E-H}) \\ & - 0.2814^s q_{13L}^H(x_{E-H}) \\ N = & 1525, \quad \lambda = 0.47, \quad D_2 = 4.54, \\ F(9.1515) = & 191.03, \quad p < 0.0001 \end{aligned} \quad (12)$$

where, N is the number of compounds, λ is Wilk's lambda, D^2 is the squared Mahalanobis distance, F is the Fisher ratio, and p -value is the significance level.

Eq. 11, which includes nonstochastic indices, classified correctly 92.66% of the compounds in the training set, misclassifying only 112 chemicals of a total of 1525. The percentage of false actives in this data set was only 2.49%, that is, 38 inactive compounds were classified as actives from 1525 cases. Conversely, 74 chemicals from the group of actives were misclassified as inactive ones (4.85% of misclassification).

In this set, model obtained with total and atom-type stochastic quadratic indices (Eq. 12) classified correctly

89.66% of antibacterial and 91.05% of inactive compounds, for a global good classification of 90.36%. Table 2 depicts the results of the classifications for both models in the training set.

The classification of all compounds in the complete training data set provides some assessment of the goodness of fit of the models, but it does not provide a thorough criterion of how the models can predict the biological properties of new compounds. To assess such predictive power, the use of a test set is essential.⁷⁷ In this sense, the activity of the chemicals in such set was predicted with the two obtained discrimination functions. The global good classification for this set was 92.28% (466/505) and 89.31% (451/505) using models

Table 4. Names and classification (Eqs. 11 and 12) of some compounds in test set

Active compound names	$\Delta P\%^a$		Inactive compound names	$\Delta P\%^a$	
	Nonstochastic	Stochastic		Nonstochastic	Stochastic
Tio-Urasin	97.81	76.16	Moroxidine	−80.67	−81.22
Nifuratrone	95.35	81.82	Triclofos	−84.62	−30.78
Solupront	97.99	87.29	Arecoline	−96.24	−95.77
Acefuralazine	97.06	84.75	Methenamine	−98.95	−97.07
Azidamfenicol	84.07	91.50	Valpromide	−92.49	−90.30
Bostrycin	85.01	92.22	Prorenone	−84.22	−88.65
Sulfapyrazine	94.05	87.53	Teofillina	−56.50	−81.86
Aspiculamycin	99.99	99.70	Metformin	−90.50	−98.30
Cinerubin A	99.75	99.92	Alarmin	−93.95	−89.06
Cefmetazole	98.84	98.91	Nicopholine	−76.72	−64.09
Everninomycin B	100.00	100.00	Pentaquinomethonium	−86.48	−72.48
Sulfaphenazole	88.67	86.86	Antafenite	−86.20	−81.12
Furoxacillin	98.24	93.84	Aligeron	−97.90	−92.91
Baludon	99.99	97.60	Oxaditon	−97.18	−92.20
Purpuromycin	99.48	99.37	Pyrantel tartrate	−83.66	−87.69
Diploicin	99.16	86.49	Cetovex	−91.15	−89.24
Antibiotic G-418	98.64	97.83	Fluoroxene	−82.55	−94.59
Rubrosal	99.53	98.15	Menthol	−90.93	−93.21
Ceftrizole	84.70	82.42	Penhexamine	−95.58	−97.22
Novobiocin	94.43	93.33	Bathyrin	−79.72	−87.61
Amicycline	93.43	97.28	Carbimazole	−91.34	−78.66
Nitrocyline	97.84	98.50	Perafenzine	−94.65	−73.61
Ribostamycin	98.03	98.16	Petidina	−97.14	−94.98
Flucloxacillin	98.74	94.75	Auxinutril	−73.67	−90.98
Seldomycin factor 1	97.80	97.31	Warfarin	−72.57	−42.50
Picloxydine	71.26	38.25	Estradiol	−86.27	−71.18
Habekacin	97.61	99.06	Acetylcholine	−96.63	−93.86
Chlortetracycline	98.55	98.66	Dipipanone	−96.49	−91.86
Racenomycin C	94.12	96.15	Aciperone	−75.26	−78.82
Hygromycin	99.27	96.34	Gobad	−86.10	−75.65
Streptomycin	99.88	99.01	Fenyltoloxamine	−97.75	−93.75
Natamycin	97.67	81.23	Bietamiverine	−93.21	−84.23
Monensin	97.05	50.90	Trastomin	−93.81	−94.51
Maridomycin 4	95.57	98.38	Octastine	−90.71	−86.41
Penimocycline	99.66	99.92	Propylhexedrine	−96.07	−97.54
Mocimycin	99.11	79.53	Trifluomeprazine	−78.80	−87.51
Maridomycin	98.45	98.87	Metiapine	−87.85	−78.73
Tylosin	99.11	96.66	Dimetilsulfoxide	−88.68	−83.93
Telomycin	85.89	88.18	Guanazole	−94.41	−98.31
Etamocycline	100.00	99.99	Norgamem	−87.82	−84.80
Vancomycin	100.00	100.00	Formetamate	−95.77	−94.11
Antibiotic LL-BM123 alpha	100.00	99.88	Bufalin	−66.34	−70.85

^a $\Delta P\% = [P(\text{active}) - P(\text{inactive})] \times 100$; where $P(\text{active})$ is the probability that the equations classify a compound as active. Conversely, $P(\text{inactive})$ is the probability that the models classify a compound as inactive. This value ($\Delta P\%$) takes positive values when $P(\text{active}) > P(\text{inactive})$ and negative otherwise. Therefore, when $\Delta P\%$ is positive (negative) the compound was classified as antibacterial (non-antibacterial).

11 and 12, respectively. The results of global classification of compounds in test set by these equations are also shown in Table 2.

In summary, the calculation of percentages of good classification in the training and external data sets permitted us to carry out the assessment of the models. In Tables 3 and 4, the results of classification using models 11 and 12 for some active and inactive compounds in training and test sets are shown. The complete set of compounds in these series, as well as their classification using both models is given as Supplementary data.

4.2. Comparison with other approaches for antimicrobial activity

In the last years, several in silico methods have been used to develop structure-based classification models on antimicrobial activity, which give rise to a good discrimination of this activity in large and heterogeneous series of organic compounds.^{36–41} However, due to differences in the composition of experimental data and chemometric methods used in carrying out the QSAR, it is not feasible to perform a comparison between the models reported in the literature for the selection of antibacterial agents. For these reason, 'strict' comparisons between the methodologies are not possible. Thus, the relative comparison will be based on the kind of method used for deriving the QSAR and their statistical parameter, the explored molecular descriptors, the number and diversity of chemical structural patterns contained in the data, the overall accuracy (%), and the validation

method used. Table 5 depicts the comparison between TOMOCOMD-CARDD method and others reported approaches for antimicrobial activity.

Firstly, TOMOCOMD-CARDD data set have more than 18 (17), 3 (4), 3 (3), 3 (4), 6 (4), and 5(4) times the number of chemicals (active compounds) with respect to models reported by García Domenech and de Julián-Ortiz,³⁶ Tomás-Vert et al.,³⁷ Mishra et al.,³⁸ Cronin et al.,³⁹ Molina et al.,⁴⁰ and Murcia-Soler et al.,⁴¹ respectively. However, all models significantly recognized the existence of active and inactive chemical groups.

The global good classification in the training set of TOMOCOMD-CARDD models (Eq. 11 = 92.66% and Eq. 12 = 90.36%) was higher than most of the reported LDA equations (see Table 5). Conversely, a connectivity function,³⁶ the BLR model³⁹ and a ANN model⁴¹ has shown an overall predictability of 94%, 94.7%, and 98.7, respectively; which seems to be bigger than the TOMOCOMD-CARDD functions predictability.

Nevertheless, it is remarkable that the TOMOCOMD-CARDD models were derived from training series 23 (1525/64), 2 (1525/661), and 5 (1525/305) times bigger than the series used by García Domenech and de Julián-Ortiz,³⁶ Cronin et al.,³⁹ and Murcia-Soler et al.,⁴¹ respectively.

Validation of the models is the other major bottleneck in QSAR.^{77,78} One of the most popular validation criteria is internal cross-validation (leave-one-out, leave-*n*-

Table 5. Comparison between TOMOCOMD-CARDD method and others approaches for antimicrobial activity

Models' features to be compared ^a	Structure-based classification models of antibacterial activity										
	Eq. 11	Eq. 12	1	2	3	4	5	6	7	8	9
<i>N</i> Total	2030	2030	111	111	664	596	661	661	352	433	433
<i>N</i> Antibacterials	1006	1006	60	60	249	307	249	249	219	217	217
Technique ^b	LDA	LDA	LDA	ANN	ANN	LDA	LDA	BLR	LDA	LDA	ANN
Wilks' λ (<i>U</i> -statistics)	0.42	0.47	0.28	—	—	0.57	NR	—	0.45	—	—
<i>F</i>	259.38	191.03	20.9	—	—	116.6	NR	—	48.2	—	—
<i>D</i> ²	5.47	4.54	NR	—	—	NR	NR	—	4.9	—	—
<i>p</i> -Level	0.00	0.00	0.00	—	—	NR	NR	—	0.00	—	—
Explored variables	75	75	16	16	62	NR	167	167	15	62	62
Variables in the model	8	9	7	16	62	3	6	6	7	6	62
<i>Training set</i>											
<i>N</i> Total	1525	1525	64	64	465	463	661	661	289	305	305
<i>N</i> Antibacterials	754	754	34	34	174	242	249	249	174	153	153
Accuracy (%)	92.66	90.36	94.0	89.0	NR	—	92.6	94.7	91.0	~85.7	~98.7
Families of drugs ^c	Broader range	Broader range	3	3	8	—	8	8	8	8	8
<i>Validation method</i>											
Validation method ^d	i	i	i	i	i	i	ii	ii	i	i	i
<i>N</i> Total	505	505	47	47	199	133	—	—	63	128	128
<i>N</i> Antibacterials	252	252	26	26	75	65	—	—	45	64	64
Predictability (%)	92.28	89.31	92	97.9	~95	84	93.6	94.3	89.0	~87.5	~91.4
Families of drugs ^c	Broader range	Broader range	3	3	8	—	—	—	5	6	6

^a Eqs. 11 and 12 are reported in this work, models 1 and 2 were reported by García Domenech and de Julián-Ortiz,³⁶ model 3 was reported by Tomás-Vert et al.,³⁷ model 4 was reported by Mishra et al.,³⁸ models 5 and 6 are after Cronin et al.,³⁹ model 7 was reported by Molina et al.,⁴⁰ and models 8 and 9 were reported by Murcia-Soler et al.⁴¹

^b LDA refers to linear discriminant analysis, ANN to artificial neural network, and BLR to binary logistic regression.

^c Only largely represented families were considered, for example, methods 1 and 2 used 3 in training quinolones, sulfonamides, and cephalosporins but add only diaminopyridine (1 compound), cephamicins (2), oxacephem (1), and sulfones (1) to predicting series.

^d Validation methods are: (i) test set, and (ii) leave-30%-out.

Table 6. Results of the virtual screening simulation of novel antimicrobial agents

Chemicals ^a	$\Delta P\%$ ^b		Chemicals ^a	$\Delta P\%$ ^b	
	Nonstochastic	Stochastic		Nonstochastic	Stochastic
1 T-3811	98.74	95.00	45	98.10	91.81
2 WQ 3034	99.97	98.98	46 KB 5246	89.80	64.56
3 WQ 2724	99.92	97.59	47 KB 5290	80.19	52.25
4 WQ 2743	99.91	97.44	48 KB 6600	94.20	81.67
5 KRQ 10196	95.00	87.43	49 KB 6625	88.48	68.02
6 KRQ 10099	69.65	36.54	50 A-255916	−24.80	20.70
7 KRQ 10018	90.01	66.79	51 A-270117	78.46	63.34
8 KRQ 10071	90.31	67.59	52 RU 79115	85.99	81.22
9 HMR 3647	95.27	88.11	53 Descladinosyl erythromycin	61.71	70.90
10 TE-810	74.95	37.07	54 ABT 773	84.38	65.77
11 TE-802	66.10	19.35	55 HMR 3562	99.48	98.48
12	72.17	45.84	56 HMR 3787	99.75	98.77
13	77.81	56.19	57	89.03	84.98
14 TEA-0769	89.25	86.54	58	72.02	45.07
15	97.53	86.56	59	89.98	71.53
16	96.94	83.69	60 CI 191,121	43.84	70.62
17	98.50	95.25	61 OCA-983	79.39	32.29
18	95.03	87.80	62	6.77	12.63
19	95.31	88.59	63 E-1010	64.80	74.53
20 PNU 101099	23.77	37.62	64 DK-35C	60.80	75.93
21 PNU 101850	77.05	−18.44	65	−18.94	70.07
22 Esperezolid	41.66	9.96	66	93.97	94.55
23 MC 02479	99.16	98.53	67 J 111, 225	−56.76	23.76
24 AS-924	97.41	91.38	68	70.69	88.50
25	83.49	71.70	69 LB 10827	99.99	99.99
26	83.57	78.21	70 MC 03971	99.98	99.93
27	73.80	72.13	71	3.73	−42.24
28	83.86	76.18	72	5.81	−51.94
29 PA 824	−16.48	−42.21	73	95.97	64.27
30 PA 1297	−39.68	−53.60	74 PNU 172576	79.85	28.77
31 PGE 711699	99.89	99.39	75 Bisbenzylamide eromomycin	100.00	100.00
32 PGE 7594630	99.92	99.57	76 Psammaplin A	99.58	98.48
33 SCH 27 899	100.00	100.00	77 HKI 9724037	98.72	99.60
34 WQ 3330	98.56	88.30	78	−7.54	75.79
35 WQ 2942	98.12	93.26	79 SEP 137199	−5.23	47.28
36 WQ 2756	99.65	97.73	80 SEP 32196	−37.23	82.90
37 WQ 2908	98.57	94.57	81 SEP 132617	−11.63	44.43
38 PGE 9262932	−2.19	10.19	82 KY-9	86.90	35.55
39 PGE 4175997	12.25	10.01	83 Ro 62-6091	64.86	77.89
40 PGE 9509 926	1.09	20.13	84 Ro 64-5781	94.53	87.01
41 NFSQ	98.94	98.09	85 VRC 483	−7.39	32.07
42	98.57	94.59	86 9567 567	1.81	−59.01
43	99.70	98.03	87 5522 293	55.45	1.33
44	98.09	90.32			

^a Chemicals 1–33 and 34–87 were taken from Refs. 89 and 90, respectively. The molecular structures of these compounds are illustrated in Table 7.

^b Classification of compounds by both models, Eqs. 11 and 12: $\Delta P\% = [P(\text{active}) - (\text{inactive})] \times 100$.

out, leave-30%-out and so on). Nevertheless, can be that exist a lack of correlation between the good results in internal cross-validation and the high predictive ability of QSAR models.^{77,78} Thus, the good high behavior in internal cross-validation appears to be necessary but not sufficient condition for the models to have a high predictive power. In this sense, Golbraikh and Tropsha emphasize that the predictive ability of a QSAR model can only be estimated using an external test set (external validation) of compounds that was not used for building the model and formulated a set of criteria for evaluation of predictive ability of QSAR model.^{77,78} With the Cronin et al.' model³⁹ exception (leave-30%-out), all the other reported models were successfully validated by means of external prediction series.

It is reasonable to expect some decrease in overall predictability of predicting sets with respect to training series for a simple reason; the model is developed to fit the points in training series, and therefore data points in predicting series are never used to develop it. In this case, the overall accuracy in test sets of TOMOCOMD-CARDD models (Eq. 11 = 92.28% and Eq. 12 = 89.31%) was higher than the rest of reported LDA equations (see Table 5). Only two nonlinear (ANN) models (Eqs. 23⁶ and 33⁷ in Table 5) have a bigger predictability but use a very much reduced number of active compounds (26 and 75, respectively) than the TOMOCOMD-CARDD approach (252 antibacterial agents).

Another remarkable problem, especially in the case of heterogeneous series of chemicals classification is the

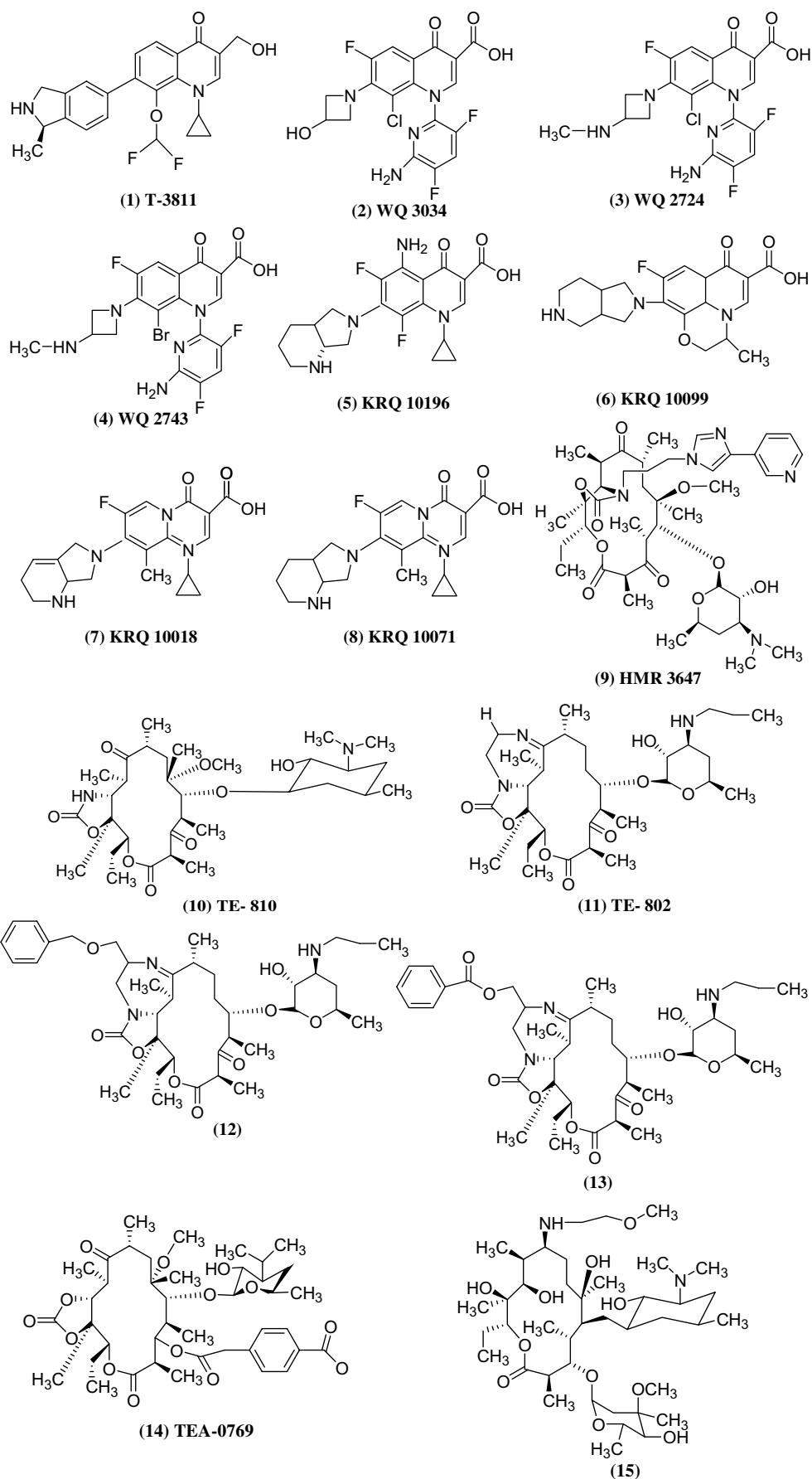
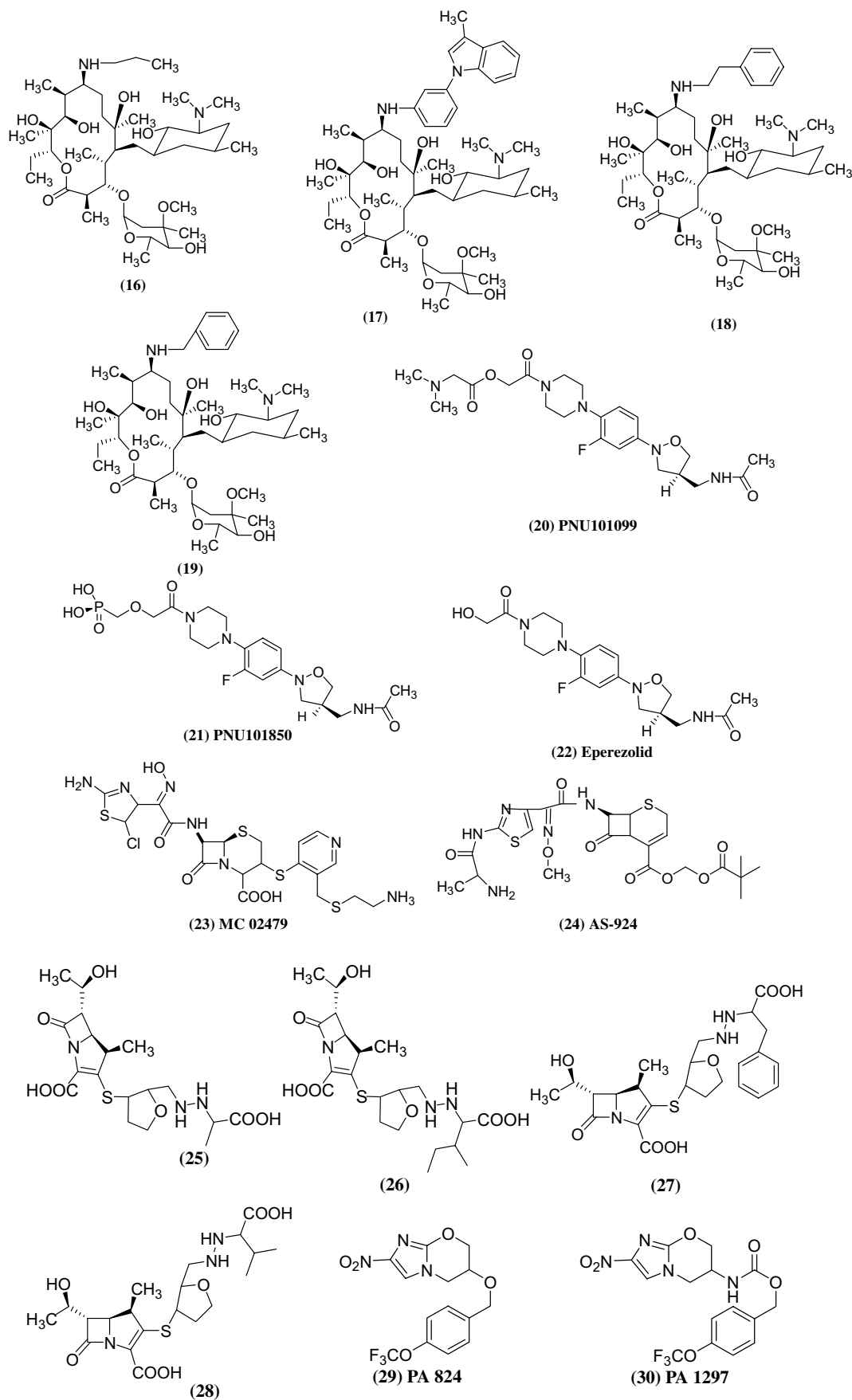
Table 7. Structure of new compounds reported in the antiinfective field with antibacterial activity

Table 7 (continued)



(continued on next page)

Table 7 (continued)

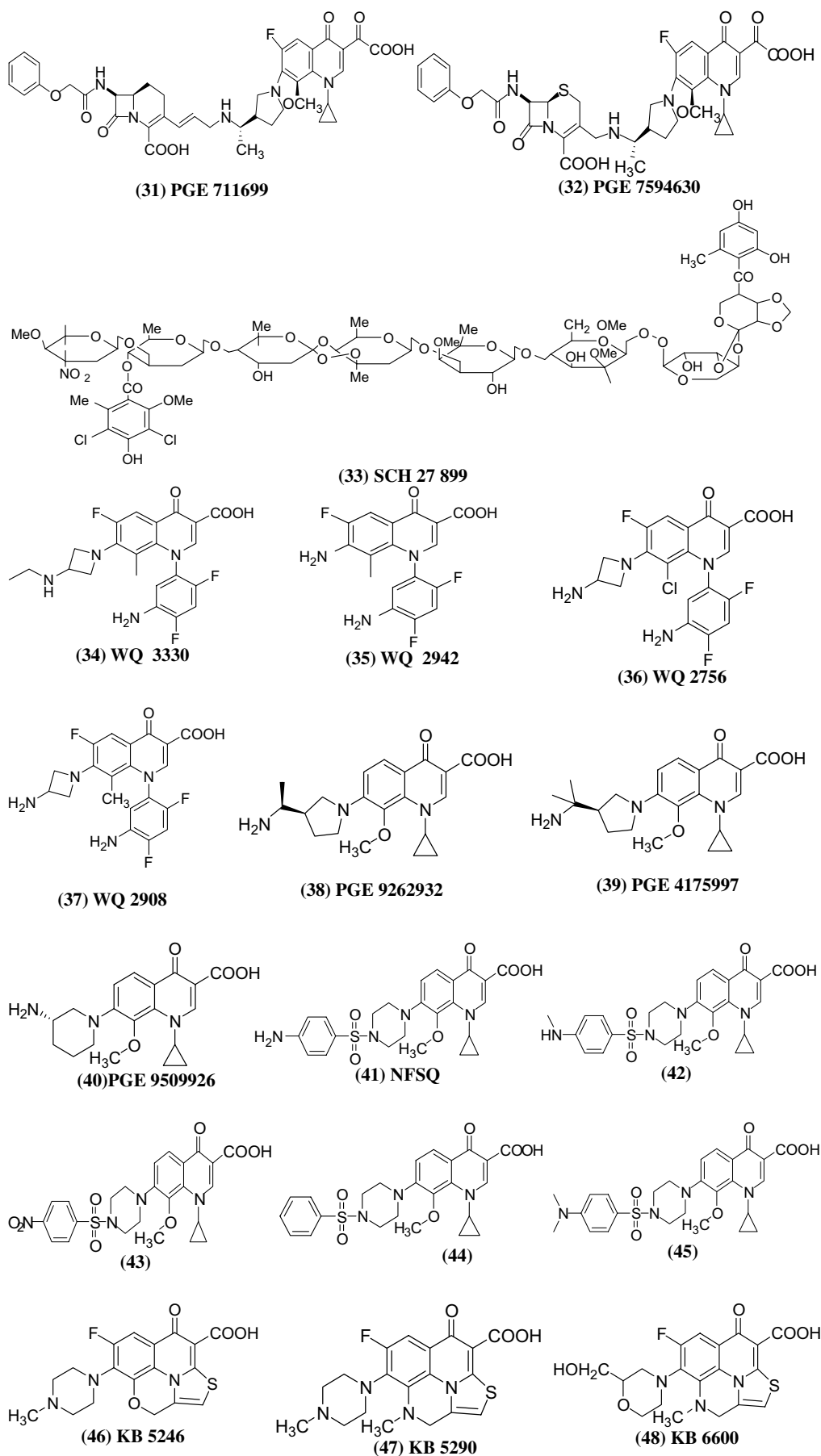
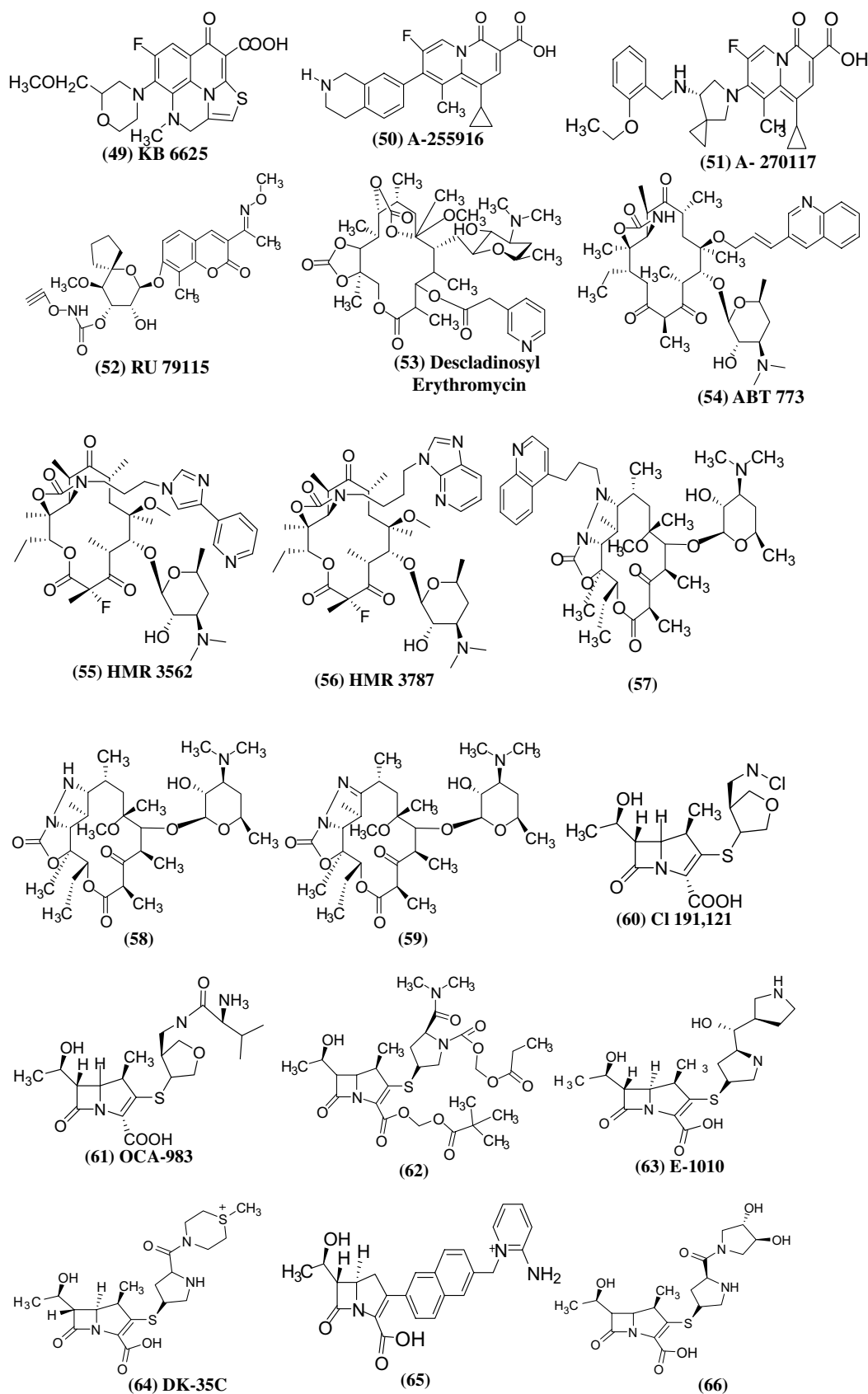
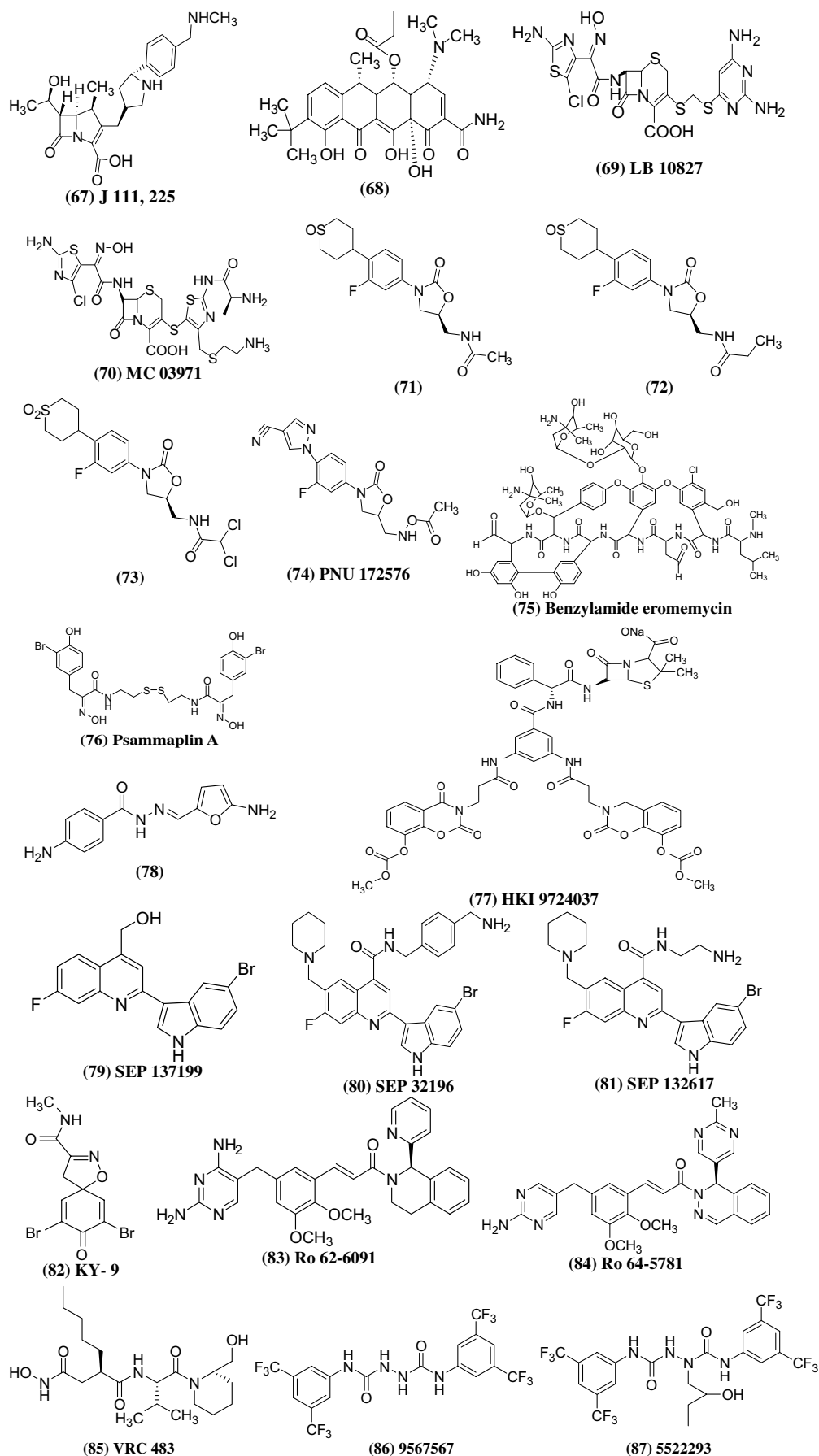


Table 7 (continued)



(continued on next page)

Table 7 (continued)



spectrum of structural patterns considered. Without doubts, the TOMOCOMD-CARDD models reported here considered a broader diversity of antimicrobial families (see Tables 3 and 4 in support data to obtain the complete list of (2030) chemicals used in training and prediction sets) in special if one takes into account that all previous studies add a few compounds of only 3–7 families to the predicting series.

4.3. Computational screening: a promising alternative for—in silico—design and/or identification of leads as antibacterial agents

Despite strong advances, the process of drug discovery is still an arduous task. In this sense, virtual screening (based on QSAR techniques) has emerged as an interesting alternative to high-throughput screening and an important drug-design tool.^{79–88} With the aim of testing the ability of our models to detecting new lead compounds with ‘unknown’ structures and action modes, we carried out a simulated virtual screening of chemicals that showed potent antimicrobial activity in experimental assays.^{89,90} To avoid the manipulation of large databases of compounds we have selected a series of new 87 compounds, previously reported in anti-infective field as potent antibacterial agents, that will be evaluate by TOMOCOMD-CARDD models (Eqs. 11 and 12) as antibacterial/non-antibacterial ones. The ability of the models to classify these compounds and their molecular structures is recorded in Tables 6 and 7, respectively.

As can be seen, both models (Eqs. 11 and 12) classify correctly most of the 87 selected compounds, showing an overall accuracy of 87.36% and 93.10%, respectively.

Some of these chemicals are new leads as antibacterial agents. That is to say, not one compound with this kind of structure was included in the training data set for developing models 11 and 12. In this sense, this in silico evaluation is equivalent to the discovery of new lead compounds using the developed models.

This result is in accordance with the character of the TOMOCOMD-CARDD approach, which permits to consider implicitly, through the calculation of nonstochastic and stochastic quadratic molecular descriptors, substructural and global features responsible for a specific activity. In this way, new lead compounds could be designed using the TOMOCOMD-CARDD method described in this paper.

5. Conclusion

This study has examined a large data set of compounds with a considerable structural variability that has been classified according to their antibacterial activity. In this sense, the collected data of antibacterial chemicals used in this study, results in an important tool not only for the theoretical research, but for the general scientific work in this area.

On the other hand, we have shown that TOMOCOMD-CARDD approach can be applied to generate useful structure-based quantitative models in order to account for antimicrobial activity of a broader range of molecular structural patterns. In a flexible way, this method permits a quick in silico discovery of new candidates to lead compounds making use of a minimum of resources. The simulated virtual screening of 87 new compounds reported in the anti-infective field with antimicrobial activities has proved the ability of our models for an adequate discrimination of new active compounds from inactive ones. Therefore, TOMOCOMD-CARDD method may be used as an efficient alternative to high-throughput screening of antimicrobial agents.

Acknowledgements

F.T. acknowledges financial support from the Spanish MCT (Plan Nacional I + D + I, project No. BQU2001-2935-C02-01) and Generalitat Valenciana (DGEUI INF01-051 and INFRA03-047, and OCYT GRU-POS03-173).

Supplementary data

The complete list of compounds used in training and prediction sets, as well as their posteriori classification according to models 11 and 12 are available free of charge via the internet at <http://www.sciencedirect.com>. Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.bmc.2005.02.015](https://doi.org/10.1016/j.bmc.2005.02.015).

References and notes

- Hooper, D. C. *Clin. Infect. Dis.* **2001**, *33*, 5157.
- Galimand, M.; Courvalin, P.; Lambert, T. *Antimicrob. Agents Chemother.* **2003**, *47*, 2565.
- Tenover, F. C. *Clin. Infect. Dis.* **2001**, *33*, S108.
- Leclercq, R.; Courvalin, P. *Antimicrob. Agents Chemother.* **2002**, *46*, 2727.
- Xiong, Y. Q.; Caillon, J.; Drugeon, H.; Potel, G.; Baron, D. *Antimicrob. Agents Chemother.* **1996**, *40*, 35.
- Collis, C. M.; Hall, R. M. *Antimicrob. Agents Chemother.* **1995**, *39*, 155.
- Maskell, J. P.; Sefton, A. M.; Hall, L. M. C. *Antimicrob. Agents Chemother.* **1997**, *41*, 2121.
- Tosini, F.; Visca, P.; Luzzi, I.; Dionisi, A. M.; Pezzella, C.; Petrucca, A.; Carattoli, A. *Antimicrob. Agents Chemother.* **1998**, *42*, 3053.
- Nagai, K.; Davies, T. A.; Jacobs, M. R.; Appelbaum, P. C. *Antimicrob. Agents Chemother.* **2002**, *46*, 1273.
- Murray, B. E. *N. Engl. J. Med.* **2000**, *342*, 710.
- Livermore, D. M. *Int. J. Anti-microb. Agents* **2000**, *16*, S3.
- Williams, R. J.; Heymann, D. L. *Science* **1998**, *279*, 1153.
- Appelbaum, P. C. *Clin. Infect. Dis.* **1992**, *15*, 77.
- Tenover, F. C.; Biddle, J. W.; Lancaster, M. V. *Emerg. Infect. Dis.* **2001**, *7*, 327.
- Asahi, Y.; Ubukata, K. *Antimicrob. Agents Chemother.* **1998**, *42*, 2267.
- Canu, A.; Malbrun, B.; Coquemont, M.; Davies, T. A.; Appelbaum, P. C.; Leclercq, R. *Antimicrob. Agents Chemother.* **2002**, *46*, 125.

17. Cetinkaya, Y.; Falk, P.; Mayhall, C. G. *Clin. Microb. Rev.* **2000**, *13*, 686.
18. Murray, B. E. *Emerg. Infect. Dis.* **1998**, *4*, 37.
19. Huycke, M. M.; Sahm, D. F.; Gilmore, M. S. *Emerg. Infect. Dis.* **1998**, *4*, 239.
20. Jung, F.; Delvare, C.; Boucherot, D.; Hamon, A. *J. Med. Chem.* **1991**, *34*, 1110.
21. Fung-Tomc, J. C.; Clark, J.; Minassian, B.; Pucci, M.; Tsai, Y. H.; Gradelski, E.; Lamb, L.; Medina, I.; Huczko, E.; Kolek, B.; Chaniewski, S.; Ferraro, C.; Washo, T.; Bonner, D. P. *Antimicrob. Agents Chemother.* **2002**, *46*, 971.
22. Macchia, M.; Menchini, E.; Orlandini, E.; Rossello, A.; Broccali, G.; Visconti, M. *Farmaco* **1995**, *50*, 713.
23. Choi, K. H.; Hong, J. S.; Kim, S. K.; Lee, D. K.; Yoon, S. J.; Choi, E. C. *J. Antimicrob. Chemother.* **1997**, *39*, 509.
24. Hagen, S. E.; Domagala, J. M.; Heifetz, C. L.; Johnson, J. *J. Med. Chem.* **1991**, *34*, 1155.
25. Domagala, J. M.; Bridges, A. J.; Culbertson, T. P.; Gambino, L.; Hagen, S. E.; Karrick, G. *J. Med. Chem.* **1991**, *34*, 1142.
26. Sum, P. E.; Petersen, P. *Bioorg. Med. Chem. Lett.* **1999**, *17*, 1459.
27. Chopra, I. *Curr. Opin. Pharmacol.* **2001**, *1*, 464.
28. Kus, C.; Göker, H.; Ayhan, G.; Ertan, R.; Altanlar, N.; Akin, A. *Farmaco* **1996**, *51*, 413.
29. Nagano, R.; Shibata, K.; Adachi, Y.; Imamura, H.; Hashizume, T.; Morishima, H. *Antimicrob. Agents Chemother.* **2000**, *44*, 489.
30. Capobianco, J. O.; Cao, Z.; Shortridge, V. D.; Ma, Z.; Flamm, R. K.; Zhong, P. *Antimicrob. Agents Chemother.* **2000**, *44*, 1562.
31. Choudhry, A. E.; Mandichak, T. L.; Broskey, J. P.; Egolf, R. W.; Kinsland, C.; Begley, T. P.; Seefeld, M. A.; Ku, T. W.; Brown, J. R.; Zalacain, M.; Ratnam, K. *Antimicrob. Agents Chemother.* **2003**, *47*, 2051.
32. Colca, J. R.; McDonal, W. G.; Waldon, D. J.; Thomasco, L. M.; Gadwood, R. C.; Lund, E. T.; Cavey, G. S.; Mathews, W. R.; Adams, L. D.; Cecil, E. T.; Pearson, J. D.; Bock, J. H.; Mott, J. E.; Shinabarger, D. L.; Xiong, L.; Mankin, A. S. *J. Biol. Chem.* **2003**, *278*, 21972.
33. Oliva, B.; Miller, K.; Caggiano, N.; O'Neill, A. J.; Cuny, G. D.; Hoemann, M. Z.; Hauske, J. R.; Chopra, I. *Antimicrob. Agents Chemother.* **2003**, *47*, 458.
34. Payne, D. J.; Miller, W. H.; Berry, V.; Brosky, J.; Burgess, W. J.; Chen, E.; DeWolf, W. E.; Fosberry, A. P.; Greenwood, R.; Head, M. S.; Heerding, D. A.; Janson, C. A.; Jaworski, D. D.; Keller, P. M.; Manley, P. J.; Moore, T. D.; Newlander, C.; Pearson, S.; Polizzi, B. J.; Qiu, X.; Rittenhouse, S. F.; Radosti, S.; Salyers, K. L.; Seefeld, M. A.; Smyth, M. G.; Takata, D. T.; Uzinskas, I. N.; Vaidya, K.; Wallis, N. G.; Winram, S. B.; Yuan, C. C. K.; Huffman, W. F. *Antimicrob. Agents Chemother.* **2002**, *46*, 3118.
35. Chopra, I.; Hodgson, J.; Metcalf, B.; Poste, G. *Antimicrob. Agents Chemother.* **1997**, *41*, 497.
36. García-Domenech, R.; de Julián-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445.
37. Tomás-Vert, F.; Pérez-Giménez, F.; Salabert-Salvador, Ma. T.; García-March, F. J.; Jaén-Oltra, J. *J. Mol. Struct. (Theochim)* **2000**, *504*, 249.
38. Mishra, R. K.; García-Domenech, R.; Galvez, J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 387.
39. 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.; Schüürmann, G. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 869.
40. Molina, E.; González-Díaz, H.; Pérez-González, M.; Rodríguez, E.; Uriarte, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 515.
41. Murcia-Soler, M.; Pérez-Giménez, F.; García-March, J.; Salabert-Salvador, Ma. T.; Díaz-Villanueva, W.; Castro-Bleda, M.; Villanueva-Pareja, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031.
42. McDonnell, G.; Russell, A. D. *Clin. Microbiol. Rev.* **1999**, *12*, 147.
43. Marrero-Ponce, Y.; Romero, V. TOMOCOMD software. Central University of Las Villas; 2002. TOMOCOMD (TOPOlogical MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version can be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es.
44. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; del Valle, A.; Torrens, F.; Castro, E. A. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 615–633.
45. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; Sánchez, A. M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 1005–1020.
46. Marrero-Ponce, Y.; Huesca-Guillen, A.; Ibarra-Velarde, F. *J. Theor. Chem. (THEOCHEM)* **2005**, *717*, 67–79.
47. Marrero-Ponce, Y.; Iyarreta-Veitia, M.; Montero-Torres, A.; Romero-Zaldivar, C.; Brandt, C. A.; Ávila, P. E.; Kirchgatter, K. *QSAR Comb. Sci.*, submitted for publication.
48. Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitia, M.; Mayón-Peréz, M.; García-Sánchez, R. *Bioorg. Med. Chem.* **2005**, *13*, 1293–1304.
49. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; Ofori, E.; Montero, L. A. *Int. J. Mol. Sci.* **2003**, *4*, 512.
50. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. A. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186.
51. Marrero-Ponce, Y.; Cabrera, M. A.; Romero-Zaldivar, V.; Bermejo, M.; Siverio, D.; Torrens, F. *Internet Electron. J. Mol. Des.*, in press.
52. Marrero-Ponce, Y. *Molecules* **2003**, *8*, 687.
53. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2010.
54. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Torrens, F.; Romero-Zaldivar, V.; Castro, E. *Molecules* **2004**, *9*, 1100–1123.
55. Marrero-Ponce, Y. *Bioorg. Med. Chem.* **2004**, *12*, 6351.
56. Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
57. Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC: Boca Raton, FL, 1992.
58. Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
59. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
60. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, 1999.
61. Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. *Int. J. Mol. Sci.* **2004**, *5*, 276 (See also CPS: physchem/0401004).

62. Marrero-Ponce, Y.; Medina, R.; Castro, E. A.; de Armas, R.; González, H.; Romero, V.; Torrens, F. *Molecules*, in press.
63. Pauling, L. *The Nature of Chemical Bond*; Cornell University Press: New York, 1939, pp 2–60.
64. Walker, P. D.; Mezey, P. G. *J. Am. Chem. Soc.* **1993**, *115*, 12423.
65. Eliel, E.; Wilen, S.; Mander, L. *Stereochemistry of Organic Compounds*; John Wiley and Sons, 1994.
66. Julián-Ortiz, J. V. de.; Alapont, C. de. G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. *J. Mol. Graphics Modell.* **1998**, *16*, 14.
67. Golbraikh, A.; Bonchev, D.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147.
68. González-Díaz, H.; Hernández-Sánchez, I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217–227.
69. Klein, D. J. *Internet Electron. J. Mol. Des.* **2003**, *2*, 814.
70. González-Díaz, H.; Marrero-Ponce, Y.; Hernández, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, U.; Castañedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; Pérez, M. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
71. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos de Armas, R. *Bull. Math. Biol.* **2004**, *66*, 1285.
72. Negwer, M. *Organic-Chemical Drugs and their Synonyms*; Akademie: Berlin, 1987.
73. The Merck Index. 12th ed.; Chapman and Hall; 1996.
74. Glasby, J. S. *Encyclopedia of Antibiotics*; Woodhouse: Manchester, 1978.
75. STATISTICA ver. 5.5, Statsoft, Inc; 1999.
76. Estrada, E.; Patlewicz, G. *Croat. Chem. Acta* **2004**, *77*, 203.
77. Wold, S.; Erikson, L. Statistical Validation of QSAR Results. Validation Tools. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995, pp 309–318.
78. Golbraikh, A.; Tropsha, A. Beware of q^2 ! *J. Mol. Graphics Modell.* **2002**, *20*, 269.
79. Drie, J. H. V.; Lajiness, M. S. *Drug Discovery Today* **1998**, *3*, 274.
80. Lajiness, M. Molecular Similarity-Methods for Selecting Compounds for Screening. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science: New York, 1990.
81. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. A. *J. Med. Chem.* **2000**, *43*, 1975.
82. Estrada, E.; Peña, A. *Bioorg. Med. Chem.* **2000**, *8*, 2755.
83. Estrada, E.; Peña, A.; García-Domenech, R. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 583.
84. Julián-Ortiz, J. V.; Gálvez, J.; Muñoz-Collado, C.; García-Domenech, R.; Gimeneo-Cardona, C. *J. Med. Chem.* **1999**, *42*, 3308.
85. Gozalbes, R.; Galvez, J.; Moreno, A.; Garcia-Domenech, R. *J. Pharm. Pharmacol.* **1999**, *51*, 111.
86. Gálvez, J.; Gomez-Lechón, M. J.; Garcia-Domenech, R.; Castell, J. V. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 2301.
87. Garcia-Domenech, R.; Garcia-March, F. J.; Soler, R.; Gálvez, J.; Antón-Fos, G. M.; Julián Ortiz, J. V. *Quant. Struct.-Act. Relat.* **1996**, *15*, 201.
88. González, H.; Olazabal, E.; Castañedo, N.; Hernández, I.; Morales, A.; Serrano, H. S.; González, J.; Ramos, R. *J. Mol. Mod.* **2002**, *8*, 237.
89. Bryskier, A. *Clin. Infect. Dis.* **1998**, *27*, 865.
90. Bryskier, A. *Clin. Infect. Dis.* **2000**, *31*, 1423.