Some New Trends in Chemical Graph Theory

ARTICLE *in* CHEMICAL REVIEWS · APRIL 2008

Impact Factor: 46.57 · DOI: 10.1021/cr0780006 · Source: PubMed

4 AUTHORS, INCLUDING:

Jorge Galvez
University of Valencia
186 PUBLICATIONS **2,640** CITATIONS

SEE PROFILE

J. Vicente de Julian-Ortiz
Fundación Centro de Innovación y Demostra…
60 PUBLICATIONS **1,366** CITATIONS

SEE PROFILE

Some New Trends in Chemical Graph Theory

Ramón García-Domenech,[†,‖] Jorge Gálvez,[†,⊥] Jesus V. de Julián-Ortiz,[‡,#] and Lionello Pogliani*,[§]

Unidad de Investigación de Diseño de Farmacos y Conectividad Molecular, Departamento de Química Fisica, Facultad de Farmacía, Universitat de València, 46100 Burjassot, València, Spain, Instituto de Tecnologia Quimica, CSIC-Universidad Politecnica de Valencia, Av. de los Naranjos s/n, 46022 València, Spain, and Dipartimento di Chimica, Università della Calabria, via P. Bucci 14/C, 87036 Rende (CS), Italy

Received April 19, 2007

Contents

* To whom correspondence should be addressed. E-mail: lionp@unical.it.
[†] Universitat de València.
[‡] CSIC-Universidad Politecnica de Valencia.
[§] Università della Calabria.
[‖] E-mail: ramon.garcia@uv.es.
[⊥] E-mail: jorge.galvez@uv.es.
[#] E-mail: jejuor@uv.es.

10.1021/cr0780006 CCC: $71.00 © 2008 American Chemical Society
Published on Web 02/27/2008

Ramón García-Domenech was born in Huelma, Jaen, Spain, and studied chemistry at the University of Granada, Spain, where he graduated in 1976. He obtained his doctorate in chemistry from the University of Valencia in 1981. He became Assistant Professor in 1977 at the University of Valencia, and since 1984, he has been Full Professor of Physical Chemistry at the University of Valencia, Spain. He coauthored more than 100 papers, two patents, and three books. His current research interest is focused on the design of new drugs by molecular topology.

Jorge Gálvez is Senior Professor of Physical Chemistry at the University of Valencia (Spain) with about 30 years experience in teaching and research. He is author or coauthor of more than 120 papers, in peer review journals, dealing with technetium radiopharmaceuticals and molecular topology (see http://www.uv.es/∼galvez/galvez.htm) as well as several EU and US patents. He is a Ph.D. in chemistry, the supervisor of radioactive facilities, and a member of the Academy of Medicine of Valencia (Spain) as well as of the International Academy of Mathematical Chemistry. His main research topic is the application of molecular topology to drug design. In 1985 he was the director of the first doctoral thesis in Spain dealing with the application of molecular topology to drug design.

1. Introduction

1.1. Short Introduction

This review consists of four parts, each treating a different aspect of chemical graph theory: (i) recent developments in molecular connectivity theory, (ii) drug design using a chemical graph method, (iii) a chemical graph formulation of chemical kinetics, and (iv) recent studies on biomacromolecules using chemical graph theory. This broad study aims to give the reader broad and updated information about recent chemical graph methods to widen the horizons of these methods, as already delineated in previous reviews on the subject. Not every recent aspect of chemical graph theory will be covered here, since this would entail writing a heavy book; nevertheless, this review together with the two cited reviews[1,2] on chemical graph theory will surely cover a great deal of the continually growing field of mathematical

J. V. de Julián-Ortiz graduated from the University of Valencia, Spain, in Chemical Sciences, speciality Biochemistry, and has been a Doctor in Pharmacy since 1997. He was a postdoctoral scholarship holder of the Spanish Ministry of Foreign Affaires and of the Romanian Ministry of Education, at the Polytechnic University of Bucharest (1998–1999). He was also a postdoctoral scholarship holder of the Ministry of Education and Science at the Institute of Computational Chemistry of the University of Girona (2002–2003), as well as a researcher of the “Red de Investigación de Centros de Enfermedades Tropicales” (Network of Research of Centers of Tropical Diseases)-Faculty of Pharmacy-University of Valencia (2003–2006). He held an Associate position at Molware SL, a molecular design company oriented to the Pharmaceutical Industry, and since January 2007 he has been a researcher in the I3P program at the Institute of Chemical Technology-Polytechnic University of Valencia-CSIC. He contributed more than 40 journal articles. His scientific interests covered the prediction of pharmacological properties of new compounds, drug design, virtual combinatorial chemistry, the application of the graph-theoretical indices to the construction of chemical structures (inverse QSAR), the obtaining of mathematical models able to generalize, and indices that consider three-dimensional structure and chirality. At the present time, his interests are centered on the modeling of crystalline media and the factors that influence the synthesis of the different zeolite structures.

Lionello Pogliani is Associate Professor in physical chemistry, at the University of Calabria, Italy. He graduated in Chemistry at the University of Firenze, Italy. He received his postdoctoral training at the department of Molecular Biology of the C. E. A. (Centre d’Etudes Atomiques) of Saclay, France, at the Physical Chemistry Institute of the Technical and Free University of Berlin, and at the Pharmaceutical Department of the University of California, San Francisco, CA. Here, he coauthored an experimental work, which was awarded with the *GM Neural Trauma Research Award*. He spent his sabbatical years at the Centro de Química-Física Molecular of the Technical University of Lisbon (Portugal) and at the Department of Physical Chemistry of the Faculty of Pharmacy of the University of Valencia-Burjassot (Spain). He contributed more than 150 papers in experimental, theoretical, and didactical fields of physical chemistry, including chapters in specialized books, and made more than 40 symposium presentations. Recently, he published a book on numbers 0, 1, 2, and 3. He is a member of the International Academy of Mathematical Chemistry.

chemistry. The recent developments in molecular connectivity theory treated here are those which have accomplished a graphical encoding of the core electrons of atoms with principal quantum number $n \geq 2$ as well as encoding suppressed hydrogen atoms by the use of a perturbation parameter instead of the introduction of any new graph theory concepts. The drug design methodology outlined here takes molecular connectivity theory as its starting ingredient in developing completely new descriptors, such as the charge descriptors and the geometrical topological indices, and then through a linear discriminant analysis produces design distribution diagrams for the discriminant functions. The application of graph theory to chemical kinetics treated here achieves not only the modeling of reaction networks but also the delineation of chemical kinetics equations. The last topic shows how graph theory can be advantageously applied in modeling biomacromolecules, even including their three-dimensional structure.

The reader will notice here and there throughout the four main sections of this review some repetitions and even discontinuities. These could not be avoided since the review attempts to offer a broad overview of what is happening in chemical graph theory. However, the entire review is held together by a common aim, i.e., recounting how properties and reactivities of molecules can be modeled with completely new concepts drawn from mathematical graph theory and, even more broadly, from topology.

1.2. Main Graph Concepts

Graph theory is a branch of mathematics that has to do with topology (or rubber-sheet geometry, which is concerned with the invariant properties of a rubber sheet) and combinatorics,[3,4] and it deals with the way objects are connected. It is centered on the concept of a graph. A *graph G* can be informally defined as a set of V vertices (or points) with a set of E edges (also connections, or lines) that connect these vertices; that is, $G = (V, E)$. More formally, a graph is a set of vertices, V, and a set of unordered pairs of elements of V, called edges, E. Thus a graph is determined by the set of vertices and by the set of edges joining the vertices and not by the particular appearance of the configuration. A graph obeys no Euclidean metrics. But, following Pythagoras' *everything is number*, graphs also have their numbers. The *degree of a vertex* in a graph is the number of edges that contain it. The order of a graph is the number p of its vertices. The *distance* between two vertices is the number of edges in the shortest path joining the two vertices. *Adjacent ve*rtices in a graph are vertices joined by an edge, while *adjacent edges* are edges having a vertex in common. A *walk* is an alternating succession of vertices and edges that starts with a vertex and ends with a vertex. A *path* is a walk in which no vertex occurs more than once. A *connected graph* has every pair of vertices joined by a path. A *tree* is a connected acyclic graph. End vertices are called *terminals*, and a tree with two terminals is called a *chain*. A tree with the maximum possible number of terminals given its number of vertices and edges is a *star*. A tree that includes a vertex which is different from the others, the *root*, is a *rooted tree*. A *cyclic graph* includes at least one walk that begins and ends at the same vertex and in which every vertex is visited no more than once. A *subgraph* G^s is obtained when edges and vertices are removed from a graph G without removing the end points of any unremoved edge. Two graphs are *isomorphic* if a one-to-one correspondence exists between the vertices and edges of the first graph and the corresponding ones of the second graph.

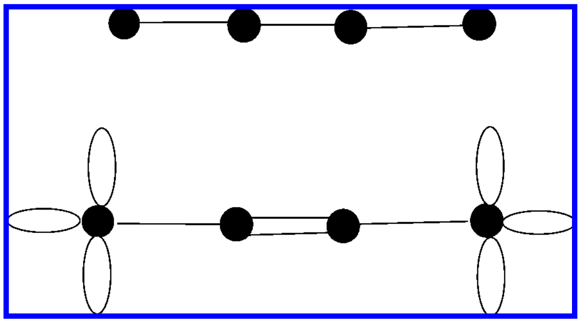


Figure 1. (A, top) The chemical graph of 1,2-difluoroethylene. (B, bottom) The chemical pseudograph of 1,2-difluoroethylene.

Chemical graph theory[5−16] applies graph theory to chemistry and centers its attention on the concept of a *chemical graph*, also known as a *molecular graph*, *structural graph*, or *constitutional graph*, all of which denote a graph where atoms and bonds are represented by vertices and edges, respectively. Actually, topological concepts have also been used in chemistry, and in fact the topology of a molecule rather than its geometry determines the form of the well-known Hückel molecular orbitals.[17] The degree of a vertex in a chemical graph is normally called its *valence*. Clearly, double bonds or lone-pair electrons cannot be described by a graph, so pseudographs or general graphs have also been used to represent molecules. A *pseudograph* (or general graph) $G' = (V, E')$ may contain *multiple edges* (E') between pairs of vertices and *self-connecti*ons (or loops), which are edges from a vertex to itself.[1,2] The degree of a vertex in a pseudograph is again the number of edges containing it, with the proviso that self-connections or loops contribute twice to its degree. Every graph is a pseudograph, but not every pseudograph is a graph. Some mathematicians reserve the term *simple graph* for a graph with no multiple edges and loops. A *chemical pseudograph* is a general graph where vertices and edges encode atoms and bonds, respectively. The degree of a vertex in a chemical pseudograph is normally called its valence. The pseudograph concept can encode multiple bonds and lone-pair electrons with multiple edges and loops, respectively.[1,18] Usually chemical graphs and pseudographs have depleted hydrogen atoms (also, hydrogen-suppressed graphs or HS graphs), which are simply molecular graphs from which hydrogens and their connecting bonds have been removed.

The chemical graph and the pseudograph of 1,2 difluoroethylene are shown in, respectively, the top and bottom of Figure 1. Note that the chemical graph in Figure 1 can also encode other molecules such as butane, dichloroethane, butadiene, etc. and that the general graph in Figure 1, even though more selective, can actually still encode any 1,2 XCH=CHX dihaloethylene, where X = F, Cl, Br, or I. A characteristic of chemical graphs and pseudographs is the impossibility for them to differentiate among different spatial isomers and especially between the *cis* and *trans* isomers around a double bond. This last problem has been partially solved in molecular connectivity theory by the concept of virtual rings.[1] The degree of the vertices in the graph at the top of Figure 1 is two for the interior vertices and one for the extreme vertices, while in the corresponding pseudograph it is three for the interior vertices and seven for the extreme vertices. Another concept from graph theory used in this study is that of a complete graph.[3,4,18,19] A *complete graph*,

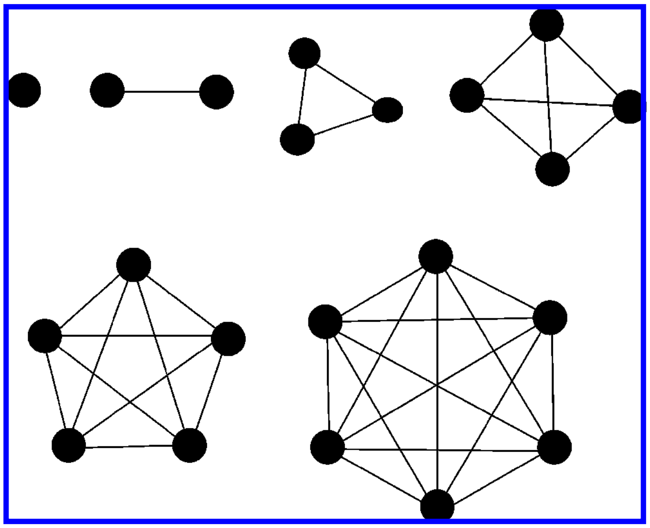


Figure 2. The K_1, K_2, K_3, K_4, K_5, and K_6 complete graphs.

K_p, of order p, is a graph where every pair of its vertices is adjacent. A complete graph is always r-regular; that is, it has all of its vertices with the same degree r, and indeed $r = p - 1$. Clearly, a complete graph is always regular, but the contrary is not true; that is, pure (nonbranched) cyclic graphs are regular.

Six complete graphs, K_1, K_2, K_3, K_4, K_5, and K_6, are shown in Figure 2. Here you can notice that K_1 is just a vertex, and this means that the vertices of graphs, as normally used in current studies, can be interpreted as K_1 complete graphs. A *null* K_0 complete graph has also been defined as a graph with no edges and vertices.[20] This null graph could be used to encode the depleted hydrogen atoms; that is, a H-suppressed chemical graph is just a set of K_0 and K_1 complete graphs and a set of edges connecting the K_1 graphs.

Graphs have been considered as two-dimensional objects even though important information about three-dimensional structure is implicit in the set of connections contained in a chemical graph. In fact, graphs containing only $-CH_2-$ clearly designate a cyclic molecule, and graphs containing quaternary ($>C<$) and tertiary ($-CH<$) carbons just as clearly describe a molecule with some degree of steric crowding. However, it is not inappropriate here to emphasize again what a graph is: a mathematical object that represents the structure of the various interconnections of a molecule. A graph is essentially a statement about vertices and edges and their relationships; it is not a physical representation of a molecule. Dimension in graph theory does not have the same meaning as the concept of dimension in physics. The claim that molecular graphs do not encode any dimensional information about molecular structure is irrelevant here. A more rigorous definition of a graph states that it is a one-dimensional complex made up of a set of zero-dimensional objects (vertices) and a set of one-dimensional objects (connections) together with a rule which assigns two distinct vertices to each connection.

For any kind of graph, it is possible to define an *adjacency matrix* and a *distance matrix*. Both of these matrices are symmetric. An adjacency matrix **A** has the element $a_{ij} = 1$ if vertices i and j are adjacent, i.e., have an edge in common, and otherwise $a_{ij} = 0$. Self-connections and multiple connections are excluded, i.e., $a_{ii} = 0$. The adjacency matrix of a pseudograph, **A′**, has the elements along the main diagonal, a_{ii}, that count the multiple connections and the self-connections (they count twice each).[1] The graph−pseudograph adjacency matrix can thus be written in a compact form in which the graph elements are $g_{i,j}$ and the pseudograph elements are $ps_{i,i}$ (the 4 × 4 matrix **M**). Matrix **M** can also be read as a prototype for a distance matrix which contains only graph elements, i.e., $ps_{i,i} = 0$, while the $g_{i,j}$ elements should be read as the corresponding distance between vertices i and j in a graph. The element d_{ij} in a distance matrix **D** equals the number of edges in the shortest path between vertices i and j. Thus, if in a cycle there are two paths not necessarily of the same length connecting vertices i and j, the meaningful distance is the smallest of these distances, and it is this distance which is entered as d_{ij}. As an example, matrix **A** is the adjacency matrix of the chemical graph of 1,2-difluoroethylene (Figure 1, top), matrix **A′** is the adjacency matrix for the chemical pseudograph of 1,2-difluoroethylene (Figure 1, bottom), and matrix **D** is the distance matrix of both the chemical and general graph of 1,2 difluoroethylene,

$$\mathbf{M} = \begin{pmatrix} ps_{1,1} & g_{1,2} & g_{1,3} & g_{1,4} \\ g_{2,1} & ps_{2,2} & g_{2,3} & g_{2,4} \\ g_{3,1} & g_{3,2} & ps_{3,3} & g_{3,4} \\ g_{4,1} & g_{4,2} & g_{4,3} & ps_{4,4} \end{pmatrix} \tag{1}$$

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \mathbf{A'} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 6 \end{pmatrix} \quad \boldsymbol{D} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

The sum of the elements either along a row or along a column in **A** and **A′** is the degree of a vertex of a chemical graph and pseudograph, respectively, normally called δ and δ^v (valence delta) in molecular connectivity theory.[13] From any chemical graph, and especially from the corresponding adjacency and distance matrices, it is possible to derive a set of *topological indices* or *graph-theoretical indices*.[1,21−25] The topological indices are numerical quantities which are based on certain topological features of a chemical graph, and they attempt to express numerically, in a direct manner, the topological information content for a given chemical compound. These indices are referred to as *graph invariants*, since isomorphic graphs possess identical topological indices. These indices may be used directly or in combination with other “ad hoc” indices as numerical descriptors to derive quantitative structure−property or structure−activity relationships (QSPR/QSAR).

There is another important category of graphs used in chemistry, the *reaction graphs*, which are applied in chemical kinetics and in computer-assisted organic synthesis.[7,26] Here, a graph encodes a reaction mechanism, where each vertex corresponds to a reactant, product, or intermediate, and the edges correspond to elementary steps. In order to summarize the successive intermediates in a multistep reaction mechanism, one depicts each intermediate by a vertex and each elementary reaction step by an edge. Reaction graphs use *undirected edges* (edges with no associated direction assigned, i.e., the normal edges), to encode reversible reaction steps, and *directed edges* (or arcs, edges with a direction, i.e., arrows), to encode irreversible reaction steps. This description will be expanded in a later section.

A less important category of graphs is that of the physicochemical graphs introduced to encode phase diagrams. In these graphs, colored vertices and edges encode phases and transitions between phases, respectively.[27] The interest in this type of graph soon faded, and an elaborate systematization of this subject was not achieved, Had the study of this type of graph continued and deepened, this field of chemical graph theory would by now have contributed new insights into problems concerning phase diagrams.

1.3. QSAR/QSPR Model Studies

The quantitative structure−activity relationship (QSAR) and quantative structure−property relationship (QSPR) inform us that for a structure (β) the property (P) or activity (A) follows. Thus, considering henceforth activities just as properties, we have the functional relationship $P = f(\beta)$. Throughout this review, by the term “structure” we mean “molecular structure”. The whole of science, and especially chemistry in its several different aspects (e.g., medicinal, organic, pharmacological, and physical), is exposed to a rapidly increasing flood of data. Thus, there is a need both to avoid drowning in this rising sea and to extract meaningful information from these data, possibly in a rather rapid and straightforward way. Mary Jo Nye[28] relates that Ernst Mach had already noted that, because our memory is limited, data must be reduced. He took the example of the kinetic law, $s = gt^2/2$. This is a simple rule of derivation by means of which from a given t we find the corresponding s. This rule replaces in a very complete and accurate manner a huge table of numbers in which for every falling time the space fallen through is reported. The rule given allows inferences, i.e., makes it possible to derive for any t the corresponding s, even if the corresponding measurements have not been performed. The expression $P = f(\beta)$ raises many questions: How should the structural factor β be quantified? What functional relationship should there be between P and β? How many data are needed to derive a meaningful relation? (Here by “meaningful relation” we mean the relation which is the simplest to derive and use.) How can the quality of the predictive relationship be checked easily? How stable is the derived relation?

The previous sections provide a clear answer to the first question because by molecular structure we here mean the corresponding graphs, i.e., the chemical graph, the general graph, and the complete graph if we are to encode the core characteristics of an atom. The information contained in these different types of graphs can be quantified with the aid of the corresponding adjacency and/or distance matrix, which allows the derivation of a set of invariants or descriptors. Todeschini and Consonni[16] defined a molecular descriptor as the final result of a logic-mathematical procedure, which transforms chemical information, encoded within a symbolic representation of a molecule, into useful numbers.

These invariants characterize in a unique way the chemical graph, and they do not depend on the numbering of the vertices of the graph. Practically, the dictum “if a structure, then a property” could be replaced by the dictum “for every chemical graph (cG), if it has a set of graph invariants or descriptors (I), then it has the property P”, as shown in shorthand notation in eq 2.

$$\{\mathrm{cG}\}(\{I\} \rightarrow P) \qquad (2)$$

In QSPR/QSAR, this rule has essentially a probabilistic character; that is, it is satisfied in all cases within certain limits. The generalization that this equation is valid for every property and activity allows its use for predictive purposes.

The best choice when looking for an unknown functional relationship is a low profile one, that is, a linear relationship: $P = aI + b$, where a and b are determined by a least-squares method. In fact, if statistics as a discipline has a message for the mathematical chemist, this message is that simple models work better than complex ones. It should not be forgotten that any function over a small domain can be approximated by a linear function and that a function such as $P = \mathrm{a}I^m + b$ is linear in a and b but is not linear in m. In some cases, a multiple relationship, $P = \mathrm{c}_0I_0 + \mathrm{c}_1I_1, c_2I_2 + ... = \sum_i c_iI_i$, where I_0 is the unitary invariant, $I_0 \cong 1$, works better. Now, if $\mathbf{P}$ is the (column) vector of the experimental data and $\mathbf{I}_1$, $\mathbf{I}_2$, etc. are the vectors of the graph-theoretical invariants, then the previous equation can also be written in a compact matrix form, i.e., as a dot product: $P = \mathbf{C}\cdot\mathbf{I}$, where $\mathbf{C}$ is the row correlation vector. The optimal invariants are normally chosen either by trying the entire combinatorial space described by the invariants or by the aid of a *greedy* algorithm, i.e., a forward selection technique which at each step just introduces the next best invariant. The nonlinear variables, e.g., m and n in relation, $P = c_0 + c_1I_1{}^m + \mathrm{c}_2I_2{}^n$, can be easily found by the aid of optimization procedures.

As for the number of data points needed to obtain a meaningful relationship, a simple rule of thumb applies: “*The more the better*”.[29] But even a small set of data can be meaningful. The problem of studying “*small*” samples may be critical in some situations, but it may be irrelevant in other cases. Actually, the sample size depends on the interest of researchers and the available data. In their fundamental work on the subject, Hansch and Leo[30] reported hundreds of regressions in which the samples had less than a dozen points.

The quality of the predictive equation in QSPR/QSAR studies is usually judged by a set of statistical parameters: r, or as some authors prefer r^2, which is called the correlation coefficient of the regression or the coefficient of determination, along with s, the standard deviation of the estimate, $c_\mathrm{i} \pm s_\mathrm{i}$, the regression parameters with their deviations, and F, the Fischer ratio. But this is not enough, and other methods should be used to check the model quality. Sometimes, some data (outliers) worsen the statistical quality of a predictive equation. In this case, it is good policy to leave them out of the model even if throwing away data, while making life easier, is intellectually unsatisfying. Plot methods, i.e., experimental vs predicted property plots, are good methods for checking the quality of the predictive relationship. Plot methods have recently received new attention because of a lack of precision that surfaced in many publications over a wide range of chemical fields.[31−33]

Normally, it is good policy in model studies to divide the entire set of data to be modeled into a *training set*, i.e., a set of vectors $\boldsymbol{P}_\mathbf{i}$, $\boldsymbol{I}_1$, $\boldsymbol{I}_2$, ..., $\boldsymbol{I}_n$ used to derive the fit-model equation, and an evaluation *set* of vectors $\boldsymbol{P}_j$, $\boldsymbol{I}_{n+1}$, $\boldsymbol{I}_{n+2}$, ..., $\boldsymbol{I}_{n+m}$, used to evaluate the $\boldsymbol{P}_\mathrm{j}$ values and thus to evaluate the reliability and predictive quality of the model equation fit found. This should also be done to aid in avoiding overfitting. Linear regressions can produce a good predictive relationship that predicts some data sets satisfactorily but that fails on other data because the original data have been overfitted. When there is a shortage of data and all data are needed to build the predictive equation, one of the validation methods most frequently used to check for overfitting of a predictive equation is the method called variously “leave-one-out”, “cross-validation”, or “jackknifing”. In this method, one piece of the training data (or two pieces for the second-order jackknife, and so on) is removed as if sliced out with a knife, the training is done on the remaining samples, and then the predictive equation so determined is used to predict the sliced-out data. This procedure is repeated until all of the data points have been left out and then predicted. Thus, for every data point the predictive equation has been validated. The prediction coefficient[34] q^2 is normally used to check the validity of the leave-one-out method. It is defined by $q^2 = (\mathrm{SD} - \mathrm{PRESS})/\mathrm{SD}$, where $\mathrm{SD} = \sum(y_i - \langle y \rangle)^2$ is the sum of the squared deviations of the observed values from their mean and where $\mathrm{PRESS} = \sum(y_i - y_{\mathrm{iloo}})^2$, in which y_{iloo} is a predicted value of the property under study when the prediction has been made by the leave-one-out method. Some authors accept any q^2 value greater 0.50 as a satisfactory result even though a higher value would be preferred. It is also good policy to keep the number of descriptors (n) well below the number of properties to be modeled, as care should be taken to avoid chance correlations that may occur whenever there are more variables than the actual number of observations. Preferably a “wealthy” predictive equation should have the ratio of the number of observations, i.e., the points to be fitted, to the number of variables as large as possible. For the addition of a new descriptor, a decreasing Fischer ratio F with all other statistics increasing in quality is a clear sign that the new descriptor is useless.

The multiple linear predictive relationship can be unstable; that is, its regression parameters may show chaotic behavior under addition or deletion of a single invariant because of interrelated invariants. Further, the value of the deviation of the regression parameter, s_i, can overshadow the value of the regression parameter itself, rendering it statistically meaningless and the predictive relationship unreliable. For the first problem, the solution is either to work with a single descriptor or to work with orthogonalized descriptors. For the second problem, the orthogonalization procedure in some cases helps in decreasing s_i, even if it needs to be redone every time a new compound is inserted or deleted in the set of compounds to be modeled. Nevertheless, even if s_i is large, the predictive character of the original equation is not endangered.[35−38] It should be emphasized here that with multilinear relations, where small rounding effects may be magnified into consistent errors in predicted values, the number of signficant figures in the coefficients returned by the regression procedure is sometimes critical for an optimal prediction.[1]

1.4. Plot Methods

Normally, methods which involve the use of plots are in general usually referred to as “graphical methods”. Nevertheless, to avoid any misunderstanding with “graph methods”, which are methods that involve the use of graphs, the phrase “plot methods” will be used in this field of study. Despite some misconceptions about plot methods, the recent awakening of interest in statistical methods in QSAR/QSPR methodology[31−33,40−45] has emphasized the critical value of plots in detecting the statistical quality of a model, as has already been described in a previous work.[1] The conclusions that can be reached from the cited studies can be summarized in a few points: (i) the need for constant use of an evaluation set, (ii) simple fit models are better than complex fit models, a fact that may help in avoiding overfitting, (iii) the exclusion of highly correlated descriptors is not always good policy, (iv) plot methods should always be included in any model study, and (v) the observed vs calculated, i.e., y vs y_{calc}, and the calculated vs observed plots are not symmetric, because their slopes a (y vs y_{calc}) and b (y_{calc} vs y), respectively, are related to the correlation coefficient, r, by the fundamental relation of the least-squares method, $ab = r^2$. Furthermore, $a = 1$ and $b = r^2$; that is, the $y_{\mathrm{calc}}|y$ plot has slope r^2 and intercept different from zero, while (vi) the $y|y_{\mathrm{calc}}$ plot disposes the points around the bisector of the first and third quadrants.[46−48] This asymmetry is maintained in the corresponding residual, $D = y - y_{\mathrm{calc}}$, plots, where (vii) the residual $D|y$ plot bears a regression line which passes through the center of mass and has a negative slope of $-r_s^2$ (where $r_s^2 = 1 - b = 1 - r^2$) and an intercept equal to the intercept of the $y_{\mathrm{calc}}|y$ plot, while the points in the $D|y_{\mathrm{calc}}$ residual plot are not correlated and are symmetrically scattered around the zero baseline. This means that the asymmetry of the two plots $y|y_{\mathrm{calc}}$ and $y_{\mathrm{calc}}|y$ is also reflected in the asymmetry of the $D|y_{\mathrm{calc}}$ and $D|y$ residual plots. This means that plot methods (i.e., the use of the four types of plots, $y|y_{\mathrm{calc}}$, $y_{\mathrm{calc}}|y$, $D|y$, and $\mathrm{D}|y_{\mathrm{calc}}$ plots, and of their characteristics) are an effective aid in detecting anomalies in models and thus in validating models.

The importance of the plot methods is further exemplified by the following case, first detected by Anscombe[48] and reworked in ref 33, where the descriptor vector $\mathbf{d} = (10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)$ can describe the following three properties, with the model eq $\mathbf{y}(\mathrm{A, B, C}) = 0.50\mathbf{d} + 3.00$, in exactly the same way, i.e., with $r^2 = 0.667$, $s = 1.24$, $F = 18$, $N = 12$,

$$\mathbf{y}(\mathrm{A}) = (8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68)$$

$$\mathbf{y}(\mathrm{B}) = (9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74)$$

$$\mathbf{y}(\mathrm{C}) = (7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73)$$

If we regress the experimental vs calculated values, the $y = y_{\mathrm{calc}}$ line is obtained, up to three decimal figures. If we now look at the corresponding plots, shown in Figure 3, we notice at once that something suspicious is going on, especially in the last two descriptions. Clearly, only the first plot represents a rational model of our property.

2. New Trends in Molecular Connectivity

2.1. Mathematical Tools

2.1.1. Complete Graphs and the Core Electron Representation

As can be seen from Figure 1, the pseudograph encoding 1,2-difluoroethylene could also encode 1,2-dichloroethylene, 1,2-dibromoethylene, and 1,2-diiodoethylene. It is thus evident that even when the concept of a pseudograph is used, which allows a rather faithful encoding of a molecule if all its atoms are second row atoms, and with which multiple bonds and nonbonding electrons can easily be encoded, there is still a problem with the encoding of the core electrons of atoms. Now, for atoms with principal quantum number $n \geq 2$, the contribution of the core electrons was taken into account with the aid of complete K_p graphs (see Figure 2).[19,49−54] The complete graph representation for the core electrons together with the pseudograph representation of a

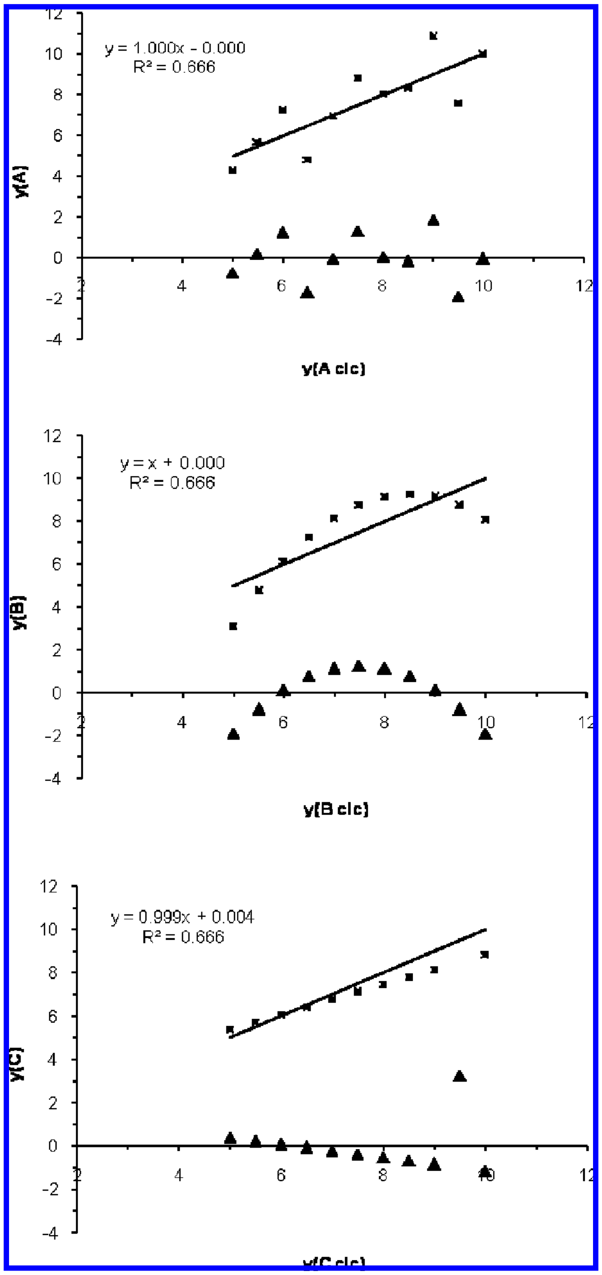


Figure 3. Plot of three different properties modeled with a unique descriptor.

molecule solves one of the main problems of chemical graph theory, i.e., the differentiation of diatomic molecules that up until now were represented by the same molecular simple graph and also the same general graph.

In Figure 4 the chemical pseudograph plus complete graph of 1,2-dichloroethylene (top) and 1,2-dibromo ethylene (bottom) are shown. The two carbons are, as always, represented with K_1 complete graphs, while the two Cl and Br atoms are represented with a K_3 and a K_5 graph, respectively. Up to now, odd-complete graphs, i.e., those with $p = 2k - 1$ for $k = 1, 2, 3, ...$, have usually shown a good model quality in QSAR/QSPR studies, while sequential complete graphs, i.e., those with $p = k$ for $k = 1, 2, 3, ...$, have rarely shown a good model quality. Instead, even-complete graphs, i.e., those with $p = 2k$ for $k = 1, 2, 3, ...$, give rise to a series of conceptual difficulties, as they would oblige us to redefine all those hydrogen-suppressed (HS) chemical general graphs made up of second row atoms, whose vertices are represented with K_1 complete graphs. In fact, once the K_0 graph has been discarded for obvious reasons, a K_2 graph with two vertices and an edge should be introduced for the second row atoms, which would invalidate all the chemical graphs used so far to encode molecules with second row atoms. It should be underlined that the dimensions of the halogen vertices in Figure 4 have no metrical meaning. This is just a zoom of the halogen vertex, and the dimensions are as meaningless as the dimensions of the K_1 vertex. The circle enclosing the complete graph also has no meaning; it is just a frame for the complete graph. The depleted hydrogen atoms of a chemical graph could easily be encoded with the already cited null graph, K_0, with no points and no edges, and with the intriguing property of a negative regularity,[20] but we will come back to this problem of the depleted hydrogen atoms. The algorithm for the degree of a vertex of a chemical pseudograph-complete graph, δ^ν, should now be able to calculate all contributions from a general plus complete graph. The algorithm presented in eq 3, which is centered on the two key parameters of complete graphs, p and r, and on a key parameter of a pseudograph, $\delta^\nu(ps)$, has been proposed and successfully tested[49−54]

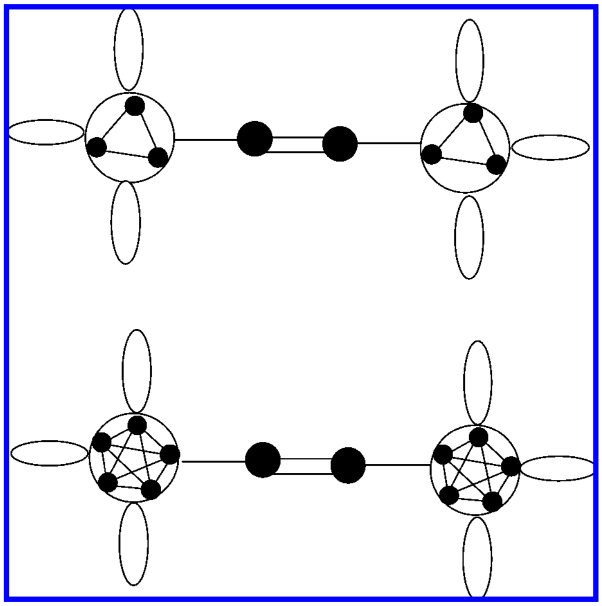


Figure 4. The chemical pseudograph plus complete graph of 1,2-dichoro- (top) and -dibromoethylene (bottom). The core electrons of the carbon atom are encoded by a K_1 graph, while the core electrons of Cl and Br are encoded by K_3 and K_5 graphs, respectively.

$$\delta^\nu = q\delta^\nu(ps)/(pr + 1) \qquad (3)$$

The parameter $\delta^\nu(ps)$ is the vertex degree of an atom in a pseudograph, and it can be obtained from the adjacency matrix of the chemical pseudograph. The parameter q equals 1 or p; the parameter pr equals the sum of all vertex degrees in complete graphs, and it equals twice the number of its connections. The great majority of model studies which used complete graphs to encode the core electrons showed a clear preference for the odd-complete graphs ($p = 1, 3, 5, 7, ...$) with some interesting exceptions in which sequential complete graphs ($p = 1, 2, 3, 4, ...$) also showed good model quality. The parameter q, which has fixed values, might be used as an optimizing parameter, something like Randić's variable index,[55] but at an atomic level rather than at a molecular level. Throughout the properties studied, $q = 1$

Table 1. Electronegativity and Atomic Radii (pm) of 30 Elements Belonging to the Main Groups 1A−7A and to the Periods $n = 2-6$[a]

	1A	2A	3A	4A	5A	6A	7A
2 ($\mathbf{K_1}$)	Li (3) 1.0; 152	Be (4) 1.5; 112	B (5) 2.0; 98	C (6) 2.5; 91	N (7) 3.0; 92	O (8) 3.5; 73	F (9) 4.0; 72
3 ($\mathbf{K_3}$)	Na (11) 0.9; 186	Mg (12) 1.2; 160	Al (13) 1.5; 143	Si (14) 1.8; 132	P (15) 2.1; 128	S (16) 2.5; 127	Cl (17) 3.0; 99
4 ($\mathbf{K_5}$)	K (19) 0.8; 227	Ca (20) 1.0; 197	Ga (31) 1.6; 135	Ge (32) 1.8; 137	As (33) 2.0; 139	Se (34) 2.4; 140	Br (35) 2.8; 114
5 ($\mathbf{K_7}$)	Rb (37) 0.8; 248	Sr (38) 1.0; 215	In (49) 1.7; 166	Sn (50) 1.8; 162	Sb (51) 1.9; 159	Te (52) 2.1; 160	I (53) 2.5; 133
6 ($\mathbf{K_9}$)	Cs (55) 0.7; 265	Ba (56) 0.9; 222	Tl (81) 1.8; 171	Pb (82) 1.9; 175	Bi (83) 1.9; 170	Po (84) 2.0; 164	At (85) 2.2; 142

[a] For each atom, in parentheses is the atomic number, Z. In parentheses, in the first column (boldfaced), is the dorresponding type of odd complete graph.

or p, with the consequence that four representations of basis indices are possible: (i) for $q = 1$, $p =$ odd, i.e., a K_p-(p-odd) representation; (ii) for $q = 1$, $p =$ sequential, a K_p-(p-seq) representation; (iii) for $q = p$, $p =$ odd, a K_p-(pp-odd) representation; (iv) for $q = p$, $p =$ sequential, a K_p-(pp-seq) representation. For the depleted hydrogen atoms in hydrogen-suppressed (HS) chemical pseudographs, it could be assumed that $q = p = 0$. Now, the full pseudograph-complete graph adjacency matrix (which can be asymmetric in some cases) for a chemical graph with four vertices has the general form shown in eq 4.

$$\mathbf{A} = \begin{pmatrix} k_1 & 0 & 0 & & \\ 0 & k_2 & 0 & 0 & k_3 \end{pmatrix} \begin{pmatrix} ps_{1,1} & g_{1,2} & g_{1,3} \\ g_{2,1} & ps_{2,2} & g_{2,3} \\ g_{3,1} & g_{3,2} & ps_{3,3} \end{pmatrix} \quad (4)$$

In eq 4, $k_i = q/(pr + 1)_{Kpi}$; that is, this value depends on the type of complete graph chosen for the given vertex-atom. For hydrogen-suppressed 1-Cl,2-Br-ethylene, Cl^1-C^2═C^3-Br^4, assuming $q = 1$, and p odd, after multiplication, we have $ps_{1,1} = 6/7$, $g_{1,2} = 1/7$ (for Cl), and $ps_{4,4} = 6/21$, $g_{4,3} = 1/21$ (for Br), while for the remaining carbon atoms we have $g_{2,1} = g_{2,2} = g_{2,3} = g_{3,2} = g_{3,3} = g_{3,4} = 1$, with all other elements being zero. The final matrix for this compound is no longer symmetric, as is shown in eq 5.

$$\mathbf{A} = \begin{pmatrix} 6/7 & 1/7 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1/21 & 6/21 \end{pmatrix} \quad (5)$$

Clearly, the **A** matrix for the compounds in Figure 1 is symmetric. The choice $q = 1$ or p has its rationale in the fact that for $q = p$ the δ^v values are rather similar to the $\delta^v = (2/n)^2\delta^v(ps)$ values introduced within the frame of the electrotopological state, E_S, concept.[1,57] The δ^v values are rather similar when $q = 1$ to the values obtained with the $\delta^v = \delta^v(ps)/(Z - Z^v - 1)$ algorithm, which was used to derive the valence molecular connectivity indices for any kind of atom.[13] For second-row atoms, we always have $\delta^v = \delta^v(ps)$, as it should be.

In Table 1 the electronegativity values taken from ref 56 for the atoms of the 1A−7A groups are collected together with their atomic number for principal quantum number $n = 2-6$. Let us now see how the $\delta^v(K_p)$ and $\delta^v(K_{pp})$ with p odd relate to the electronegativity of the corresponding atoms.

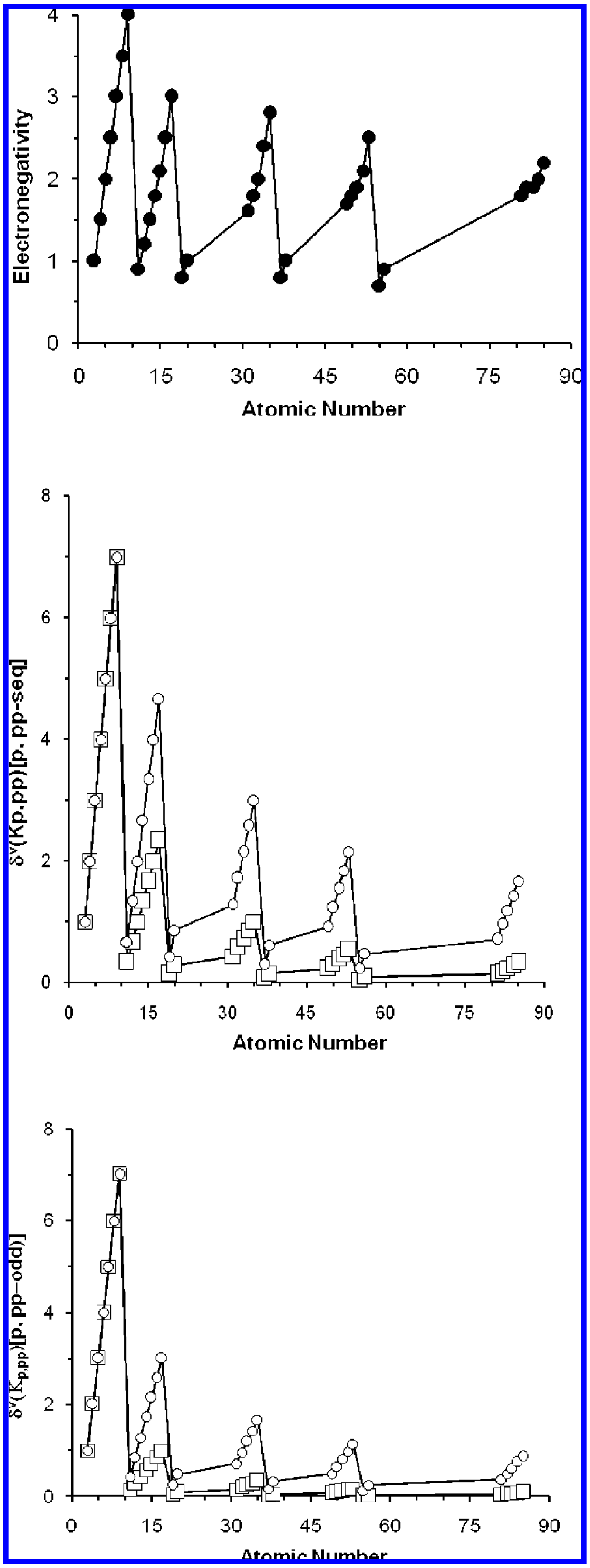


Figure 5. (Top) Electronegativity values of atoms belonging to groups 1A−7A as a function of the atomic number. (Middle) The $\delta^v[K_p(p\text{-seq})]$ (□) and $\delta^v[K_p(pp\text{-seq})]$ (○) values vs the atomic number. (Bottom) The $\delta^v[K_p(p\text{-odd})]$ (□) and $\delta^v[K_p(pp\text{-odd})]$ (○) values vs the atomic number.

Figure 5, top, shows the electronegativity values of these atoms as a function of their atomic number, while Figure 5, middle, shows plots of the $\delta^v[K_p(p\text{-seq})]$ values together with the higher $\delta^\nu[K_p(pp\text{-seq})]$ values vs their atomic number. Figure 5, bottom, shows instead plots of the $\delta^\nu[K_p(p\text{-odd})]$ values and the higher $\delta^\nu[\mathrm{K}_p(pp\text{-odd})]$ values vs the corresponding atomic number. The $\delta^\nu(K_p)$ values, either sequential or odd, coincide for $n = 2$, since $\delta^\nu = \delta^\nu(ps)$, and they differ consistently from the electronegativity values even though their trend is quite similar to the trend of the electronegatiivty values. For $n = 3$, the best agreement with the electronegativity is shown by the $\delta^\nu[K_p(pp\text{-odd})]$ values, while, for $n > 3$, the best agreement with electronegativity is shown by the $\delta^\nu[K_p(pp\text{-seq})]$ values. A detailed analysis of these figures shows also that the $K_p(pp\text{-seq})$ and $K_p(p\text{-seq})$ values are somewhat closer to each other than the corresponding $K_p(pp\text{-odd})$ and $K_p(p\text{-odd})$ values, which means that the odd complete case guarantees a better resolution between the p-odd and pp-odd values.

2.1.2. Hydrogen Perturbation

Now, algorithm 3 as well as the previous algorithms of molecular connectivity does not encode the bonded hydrogen atoms and does not allow differentiation between p- and $\delta^\nu(ps)$-similar atoms, which differ only in the bonded number of hydrogen atoms. Consider the first atom of the compounds {LiF, BeHF, BH_2F, CH_3F}: These all have the same δ^ν, and the same holds for {BeF_2, BHF_2, CH_2F_2}, {BF_3, CHF_3}, and correspondingly for higher row atoms, i.e., for vertices with $p > 1$. The depleted hydrogens are responsible for this degeneracy in δ^ν. In fact, the HS-graphs and general graphs within these sets of compounds are equivalent. To circumvent this degeneracy, the hydrogen contribution will be introduced into δ^ν as a perturbation parameter, which avoids the introduction of new graph concepts. This will be tested with different properties of different classes of compounds which have different values for the ratio n_H/n_{ht} of the number of hydrogen atoms, n_H, to the number of heteroatoms, n_{ht} (The subscript ht stands for "hetero"). It should be remarked that, normally, for the class of alkanes, the algorithm of eq 3 simplifies into $\delta^\nu = \delta^\nu(ps) = \delta$, as the chemical general graphs simplify into the simple chemical graphs. Attempts to quantify hydrogen atoms have already been made by Kier and Hall,[57] and recently a modified connectivity index was proposed in which the contribution of the hydrogen atoms to the overall connectivity index was parametrized as $n_H/6$.[58] The guidelines for the new δ^ν algorithm are[53,54,59] (i) the new δ^ν should not contradict the δ^ν of eq 3 for compounds with no hydrogens, (ii) the new values for δ^ν should include a contribution from the bonded hydrogen atoms which decreases with a decreasing number of bonded hydrogen atoms, (iii) to preserve the good results of algorithm 3, this contribution should be minimal (i.e., the resulting δ^ν should not be affected in a significant way by the hydrogen atoms), (iv) this contribution should decrease in importance with increasing p (i.e., the dependence on p should equal the one given in eq 3), and last but not least, (v) the new algorithm should not introduce any new graph concept relative to algorithm 3. With all these guidelines in mind, we will define the fractional parameter f_δ based solely on graph concepts as shown in eq 6.

$$f_\delta = [\delta^\nu_{\mathrm{m}}(ps) - \delta^\nu(ps)]/\delta^\nu_{\mathrm{m}}(ps) = 1 - \delta^\nu(ps)/\delta^\nu_{\mathrm{m}}(ps) \qquad (6)$$

Here, $\delta^\nu_{\mathrm{m}}(ps)$ is the maximal $\delta^\nu(ps)$ value an heteroatom can have in a chemical HS-pseudograph when all bonded hydrogens are substituted by heteroatoms. A closer look at f_δ reveals that it is given by the ratio $f_\delta = n_H/\delta^\nu_{\mathrm{m}}(ps)$, where n_H equals the number of hydrogen atoms bonded to a heteroatom. The parameter f_δ will obey all the given guidelines, since it can be considered as a kind of perturbation to the δ^ν of eq 7.

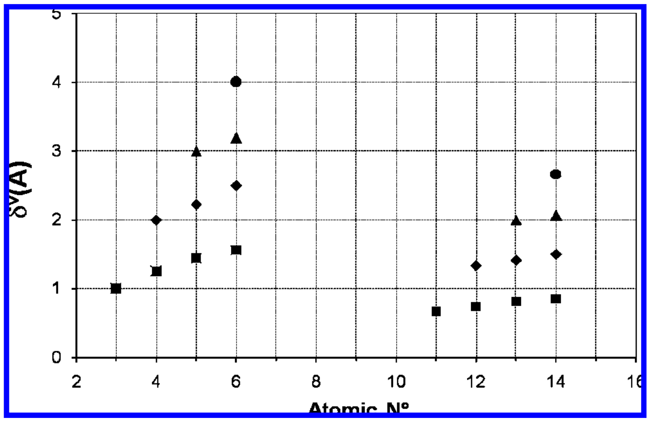


Figure 6. δ^ν values obtained with eq 7 vs the atomic number for atom A throughout eight sets of compounds, which differ among them in the number of hydrogen atoms bonded to A.

$$\delta^\nu = \frac{(q + f_\delta^{\,n})\delta^\nu(ps)}{(pr + 1)} \qquad (7)$$

Notice (i) that, for completely substituted carbon atoms (or heteroatoms), $f_\delta = 0$, since $\delta^\nu_{\mathrm{m}}(ps) = \delta^\nu(ps)$ (or $n_H = 0$), and eq 3 is retrieved and (ii) that, for saturated hydrocarbons, δ and δ^ν are no longer equal, even if $\delta^\nu(ps) = \delta$, but instead since $p = 1$ we have $\delta^\nu = (1 + f_\delta^{\mathrm{n}})\delta$. Clearly, for quaternary carbons where $f_\delta = 0$, $\delta^\nu = \delta$ is retrieved. The exponent n (here $n = 2$, 4, 6, and 8) is not an optimization parameter, since it has a constant value for each property studied. It actually detects the importance of the hydrogen atoms throughout the model of a property or activity of a class of compounds; that is, the higher the n values, the lower the perturbation and the lower the importance of the hydrogen atoms. The element $k_i = q/(pr + 1)_{Kpi}$ of the matrix in eq 4 now has to be modified into $k_i = (q + f_\delta^{\mathrm{n}})/(pr + 1)_{Kpi}$. This algorithm requires knowledge of two pseudograph adjacency matrices, **A**, for each compound: the normal adjacency ps-matrix for $\delta^\nu(ps)$ and the adjacency ps-matrix for the corresponding fully substituted compound for $\delta^\nu_{\mathrm{m}}(ps)$ to be used to derive f_δ^{n}. For example, the adjacency ps-matrix for the HS-pseudograph of CH_3-NH_2 is used to derive $\delta^\nu(ps)$, while the adjacency ps-matrix for the pseudograph of CX_3-NX_2 is used to derive $\delta^\nu{}_m(ps)$. Here X is a kind of dummy monovalent heteroatom, and this dummy compound will not be used for any further calculation.

δ^ν as given by eq 7 has been plotted vs the atomic number of the first atom (A) for the following four sets of compounds (Figure 6, left side), for which $q = p = 1$, and with f_δ^2—{LiF, BeHF, BH_2F, CH_3F} (■), {BeF_2, BHF_2, CH_2F_2} (◆), {BF_3, CHF_3} (▲), {CF_4} (●)—and also for the following four sets of compounds (Figure 6, right side, where $q = p = 2$)—{NaCl, MgHCl, AlH_2Cl, SiH_3Cl} (■), {$MgCl_2$, $AlHCl_2$, SiH_2Cl_2} (◆), {$AlCl_3$, $SiHCl_3$} (▲), {SiF_4} (●). Note that guarantee i holds for compounds LiF, BeF_2, BF_3, and CF_4 and for the corresponding second-row compounds as it should, while guidelines ii and iii hold for the other compounds. The $\delta^\nu(A)$ value of A in a set is never larger than the $\delta^\nu(A)$ value in the next set of compounds, and within a set it increases with an increasing number of bonded hydrogens, while for equal A atoms it increases with an increasing number of fluorine atoms. The difference between adjacent $\delta^\nu(A)$ values within a set of compounds, s, $\Delta\delta = \delta^\nu(A^s{}_{i+1}) - \delta^\nu(A^s{}_i)$, decreases, as is evident in the first and second set, in keeping with the decreasing importance of the bonded hydrogen atoms. The difference between $\delta^\nu(A)$ values, $\Delta\delta' = \delta^\nu(A_i{}^{s+1}) - \delta^\nu(A_i{}^s)$, throughout the given sets of compounds is nearly constant, in keeping with the constant contribution of the entering fluorine atom. Equation 7 guarantees that, with increasing p, $\delta^\nu(A)$ decreases in any case (guideline iv), as can be seen on the right side of Figure 6. Thus, all valence χ^ν values obtained with eq 7 for δ^ν are able to differentiate among compounds that have atoms with $\delta^\nu(ps) \neq \delta^\nu{}_m(ps)$. As we do not have a complete set of meaningful physicochemical property values to check the present conjecture, we regress the new ${}^1\chi^\nu(7)$ values obtained with eq 7 (with $f_\delta{}^2$) vs the old ${}^1\chi^\nu(3)$ values obtained with eq 3 for the first five sets of compounds to determine if they diverge markedly from each other. The linear regression so obtained is rather good, as is indicated in eq 8.

$$
{}^1\chi^\nu(7) = 1.121\ (\pm 0.091)\ {}^1\chi^\nu(3) - 0.091\ (\pm 0.033){:}\quad F = 340,\quad r^2 = 0.977,\quad s = 0.03,\quad N = 10 \tag{8}
$$

Repeating the same calculations for the sets of the third row compounds, we obtain the results in eq 9.

$$
{}^1\chi^\nu(7) = 1.070\ (\pm 0.034)\ {}^1\chi^\nu(3) - 0.078\ (\pm 0.028){:}\quad F = 966,\quad r^2 = 0.992,\quad s = 0.02,\quad N = 10 \tag{9}
$$

This means that the old ${}^1\chi^\nu(1)$ values cannot at all be considered inconsistent values. Let us see how eq 7 (with $f_\delta{}^2$) and eq 3 affect four C−C bond types in hydrogen-suppressed pseudographs of alkanes ($q = p = 1$),

$$
\mathrm{C{-}C}{:}\quad \delta^\nu(3) = \delta^\nu(ps) = \delta = 1,\quad \delta^\nu{}_m(ps) = 4,\quad \delta^\nu(f_\delta{}^2) = 1.5625 \tag{10a}
$$

$$
\mathrm{{-}C{-}C{-}}{:}\quad \delta^\nu(3) = \delta^\nu(ps) = \delta = 2,\quad \delta^\nu{}_m(ps) = 4,\quad \delta^\nu(f_\delta{}^2) = 2.5 \tag{10b}
$$

$$
\mathrm{{>}C{-}C{<}}{:}\quad \delta^\nu(3) = \delta^\nu(ps) = \delta = 3,\quad \delta^\nu{}_m(ps) = 4,\quad \delta^\nu(f_\delta{}^2) = 3.1875 \tag{10c}
$$

$$
{\rightarrow}\mathrm{C{-}C}{\leftarrow}{:}\quad \delta^\nu(3) = \delta^\nu(ps) = \delta = 4,\quad \delta^\nu{}_m(ps) = 4,\quad \delta^\nu(f_\delta{}^2) = 4 \tag{10d}
$$

Here δ is the main molecular connectivity parameter that can be obtained from the simple HS-graph, which, in this case, coincides with the HS general graph. The relation shown in eq 10e between $\delta^\nu(7)$, from eq 7, and $\delta^\nu(3)$, from eq 3, confirms that the $\delta^\nu(3)$ values are not that inconsistent.

$$
\delta^\nu(7) = 0.8\ (\pm 0.03)\ \delta^\nu(3) + 0.813\ (\pm 0.091),\quad F = 585,\quad r^2 = 0.997,\quad s = 0.07,\quad N = 4 \tag{10e}
$$

2.1.3. Molecular Connectivity and Pseudoconnectivity Basis Indices and their Duals

To avoid a huge combinatorial problem, only a restricted set of molecular connectivity basis indices, $\{\beta\}$, will be considered here for model purposes. It should be remembered that with a δ^ν and a $\delta^\nu(ps)$ two different types of valence indices could actually be derived, with one being based on δ^ν and the other on $\delta^\nu(ps)$. The actual set of valence connectivity indices will only be based on δ^ν. Our model set $\{\beta\} = \{\{\chi\}, \{\psi\}, \{\beta_\mathrm{d}\}\}$ is composed of three subsets: $\{\chi\}$, a collection of eight molecular connectivity basis indices, $\{\psi\}$, a collection of eight molecular pseudoconnectivity basis indices, and $\{\beta_\mathrm{d}\}$, a collection of twelve dual connectivity and pseudoconnectivity basis indices.[13,60−62] These will all be classified as molecular connectivity basis indices. It should be noted that a small molecule with differing heteroatoms, such as acetic acid, can have more than forty molecular connectivity basis indices. The cited subsets of basis indices are given in eqs 11, and their definitions are collected into pairs according to their formal similarity in eqs 12−15.

$$
\{\chi\} = \{D, {}^0\chi, {}^1\chi, \chi_t, D^v, {}^0\chi^v, {}^1\chi^v, \chi^v{}_t\},\ \{\psi\} = \{{}^S\psi_I, {}^0\psi_I, {}^1\psi_I, {}^T\psi_I, {}^S\psi_E, {}^0\psi_E, {}^1\psi_E, {}^T\psi_E\}
$$

$$
\{\beta_d\} = \{{}^0\chi_d, {}^1\chi_d, {}^1\chi_s, {}^0\chi^v{}_d, {}^1\chi^v{}_d, {}^1\chi^v{}_s, {}^0\psi_{Id}, {}^1\psi_{Id}, {}^1\psi_{Is}, {}^0\psi_{Ed}, {}^1\psi_{Ed}, {}^1\psi_{Es}\} \tag{11}
$$

$$
D = \textstyle\sum_i \delta_i \qquad {}^S\psi_I = \sum_i I_i \tag{12}
$$

$$
{}^0\chi = \textstyle\sum_i (\delta_i)^{-0.5} \qquad {}^0\psi_I = \sum_i (I_i)^{-0.5} \tag{13}
$$

$$
{}^1\chi = \textstyle\sum (\delta_i\delta_j)^{-0.5} \qquad {}^1\psi_I = \sum (I_iI_j)^{-0.5} \tag{14}
$$

$$
\chi_t = (\textstyle\prod \delta_i)^{-0.5} \qquad {}^T\psi_I = (\prod I_i)^{-0.5} \tag{15}
$$

Index χ_t (and $\chi^\nu{}_t$) is the total molecular connectivity index, and it has its ψ counterpart in the total molecular pseudoconnectivity index, ${}^T\psi_I$ (and ${}^T\psi_E$). The sums in eqs 12 and 13, as well as the products ($\prod$) in the two eqs 15, are taken over all vertices of the hydrogen-suppressed chemical graph. The sums in the two eqs 14 are over all edges of the chemical graph, i.e., σ bonds in the molecule. Replacing δ with δ^ν, the subset of valence χ^ν indices, $\{D^\nu, {}^0\chi^\nu, {}^1\chi^\nu, \chi^\nu{}_t\}$, are obtained. Replacing, instead, I_i with S_i, the ψ_E subset $\{{}^S\psi_E, {}^0\psi_E, {}^1\psi_E, {}^T\psi_E\}$ is obtained. Superscripts S and T stand for sum and total, while the other sub- and superscripts follow the established nomenclature for χ indices.[1,13] Basis ψ indices are indirectly related to the δ^ν defined in eq 7 through the I-state (ψ_I subset) and S-state (ψ_E subset) indices,[57] as shown in eq 16.

$$
I = (\delta^\nu + 1)/\delta,\quad S = I + \textstyle\sum \Delta I,\quad \text{with}\quad \Delta I = (I_i - I_j)/r^2{}_{ij} \tag{16}
$$

In eq 16, r_{ij} counts the atoms in the minimum path length separating two atoms, i and j, and is equal to one plus the graph distance, $d_{ij} + 1$. $\sum\Delta I$ incorporates the information about the influence of the remainder of the molecular environment, and since it can be negative, S can also be negative. Since some atoms have $S < 0$, to avoid imaginary ψ_E values, every S value of our classes of compounds has been rescaled. In fact, the S values for 20 metal halides, MeX (Table 3), have been rescaled to the S value of Ba in BaF_2, where $S[\mathrm{Ba}(\mathrm{BaF_2})] = -3.083$. The S values of 54 organic compounds in Table 4 have, instead, been rescaled to the S value of Si in SiF_4, $S[\mathrm{Si}(\mathrm{SiF_4})] = -6.611$, while the S values of 25 chlorofluorocarbons (Table 5) and 34 halomethanes (Table 9) have been rescaled to the S value of the carbon atom in CF_4, i.e., $S[C(CF_4) = -5.5$. The rescaling procedure, which is also done to avoid too small or too large S_i values, has a minor influence on the quality of the modeling.[61] The dual basis indices are defined in the following by eqs 17−19.[62]

$$^0\chi_d = (-0.5)^N\prod\nolimits_i(\delta_i) \qquad ^0\psi_{Id} = (-0.5)^N\prod(I_i) \quad (17)$$

$$^1\chi_d = (-0.5)^{(N+\mu-1)}\prod(\delta_i + \delta_j)$$
$$\psi_{Id} = (-0.5)^{(N+\mu-1)}\prod(I_i + I_j) \quad (18)$$

$$^1\chi_s = \prod(\delta_i + \delta_j)^{-0.5} \qquad ^1\psi_{Is} = \prod(I_i + I_j)^{-0.5} \quad (19)$$

The corresponding χ^v valence dual indices and ψ_E dual indices are obtained by replacing δ by δ^v and I_i by S_i in these expressions. The exponent μ in eqs 18 is the cyclomatic number, which indicates the number of cycles in a chemical graph, and it is equal to the minimum number of edges which must be removed in order to convert the (poly)cyclic graph into an acyclic subgraph. Dual indices except for $^1\chi_s$ and $^1\psi_{Is}$ can be negative.

Alkali halides and all those compounds made up of two connected atoms have deceptively simple graphs, i.e., •−•. In this case, nonvalence χ indices are useless, since they are the same for all these compounds. Further, $^1\chi^v \cong \chi^v{}_t$, $^1\psi_I = {}^T\psi_I$, and $^1\psi_E = {}^T\psi_E$, while the zeroth- and first-order dual indices obey the relations given in eqs 20.

$$^0\chi_d = (-0.5)^2(\delta_1{\cdot}\delta_2) = 0.25(\chi_t)^{-2};$$
$$^1\chi_d = (-0.5)(\delta_1 + \delta_2) = -0.5D$$
$$^0\chi^v{}_d = 0.25(\chi^v{}_t)^{-2}; \quad ^0\psi_{Id} = 0.25(^T\psi_I)^{-2};$$
$$^0\psi_{Ed} = 0.25(^T\psi_E)^{-2}$$
$$^1\chi^v{}_d = -0.5D^v; \quad ^1\psi_{Id} = -0.5{}^S\psi_I; \quad ^1\psi_{Ed} = -0.5{}^S\psi_E \quad (20)$$

The first two nonvalence indices are useless, while the last three can be described by their parent nondual index, from which they differ only by a constant (−0.5). For the first-order "soft" dual indices of MeX, the relations in eq 21, of which the first is useless, hold.

$$^1\chi_s = (\delta_1 + \delta_2)^{-1/2} = D^{-1/2}; \quad ^1\chi^v{}_s = (D^v)^{-1/2};$$
$$^1\psi_{Is} = (^S\psi_I)^{-1/2}, \quad ^1\psi_{Es} = (^S\psi_E)^{-1/2} \quad (21)$$

2.1.4. Higher-Order Terms

With the basis indices, it is possible to use a trial-and-error procedure to construct a series of higher-order descriptors known as (i) molecular connectivity terms,[1,63] denoted by $X = f(\chi)$, (ii) molecular pseudoconnectivity terms, denoted by $Y = f(\psi)$, and (iii) higher-order terms, denoted by $Z = f(X,Y,\beta_d)$ and $Z' = f(Z,\beta_d)$. All of these will be called molecular connectivity terms. The X and Y terms can be as convoluted as the rational function of eq 19, even though they are usually simpler.

$$T = [a(\beta_1)^m + b(\beta_2)^n]^q/[c(\beta_3)^o + d(\beta_1)^p]^r \quad (22)$$

Here, for $\beta = \chi$, we have $T = X$, and for $\beta = \psi$, we have $T = Y$. a through d and m through r are parameters that depend on the property being modeled. Actually, it seldom happens that all of these parameters are different from zero or one.

2.1.5. Variable Molecular Connectivity Index

A quite useful molecular connectivity index which introduces a variable vertex degree in a chemical graph has recently been introduced and used in QSPR/QSAR studies.[64−70] This variable is the connectivity index, $^1\chi^f$, where the left-superscript 1 has the usual meaning given by eq 14 and where the right-superscript f means that the new index is now a function of a single variable $f(x)$, and is defined by eq 23.

$$^1\chi^f = \sum[(\delta_i + x)(\delta_j + x)]^{-0.5} \quad (23)$$

Actually, although many types of variable indices were proposed, we will be concerned here only with the variable molecular connectivity index. The definition given is valid for the homoatomic HS chemical graphs; that is, it encodes atoms of the same type. This definition makes it possible to define, following the rules for dual indices, a previously never used, rather "easy" dual variable index, i.e., $^1\chi^f{}_d = (-0.5)^{(N+\mu-1)}\prod(x\delta_i + x\delta_j)$. For heteroatomic HS chemical graphs, the formally similar but more general definition given in eq 24, in which y_H stands for the variable contribution of the heteroatom, should be used.

$$^1\chi^f = \sum[(\delta_i + x)(\delta_j + y_H)]^{-0.5} \quad (24)$$

The corresponding dual index would be $^1\chi^f{}_d = (-0.5)^{(N+\mu-1)}\prod(x\delta_i + y_H\delta_j)$. Here, each type of heteoatom, y_H, will be described by a different variable; that is, for organic molecules with carbons only, $y_H = x$, while if in addition other atom types are present, then $y_H = y$, w, z, etc. for each additional type. Algorithm 24 has also been proposed for structures containing cycles, where y_H is the contribution of carbon in a cycle. The adjacency matrix in this case is rather similar to the adjacency matrix of a chemical pseudograph, but in this case, the elements along the main diagonal are unknown and, thus, can be optimized to better adapt them to the data. This is the interesting advantage of this *variable* descriptor, which clearly depends on the optimized x or y_H value (or w or z). To avoid possibly negative radicands, we should always have $(\delta_i + x)(\delta_j + x) > 0$. Actually, the case where both terms under the radicand are negative and thus give a positive radicand has until now not been treated. If $x > 0$, then$^1\chi^f \le {}^1\chi$, while, for $x \le 0$, we have instead $^1\chi^f \ge {}^1\chi$. For the $(\delta_i + x)(\delta_j + y_H)$ radicand, rather similar reasoning is valid. Figure 7 shows the changes of $^1\chi^f$ with $x \ge 0$ and $x < 0$ for the hydrogen-suppressed chemical graph of ethane (CH_3-CH_3, ▲) and neopentane [$(CH_3)_4$, ■]. The top figure shows how, for $x = -1$ (CH_3-CH_3, ▲) and $x = -1$ and -4 [$C(CH_3)_4$, ■], $^1\chi^f \rightarrow \infty$, and how, for small x values, the $^1\chi^f$ values for the two cases do not superpose. The bottom figure shows that superposition, for large x values, is apparent, as can be resolved at the second and third decimal digits, which are important in multilinear models. Similar reasoning is valid for $(\delta_i + x)(\delta_j + y_H)$.

We would like to mention here some other graph-theoretical molecular indices which have been developed recently, to wit: the modified molecular connectivity indices *mMCI*, the eccentric connectivity index *Xu*, and a new atom-type-based index, the *AI* index.[71−74] Concerning these last indices, plot methods show that some of the data are clustered and some residuals are heteroscedastic; that is, they grow with growing values of the property, while for some other data a clear nonrandom sigmoidal pattern can be detected. The model equation in this case should probably be aug-

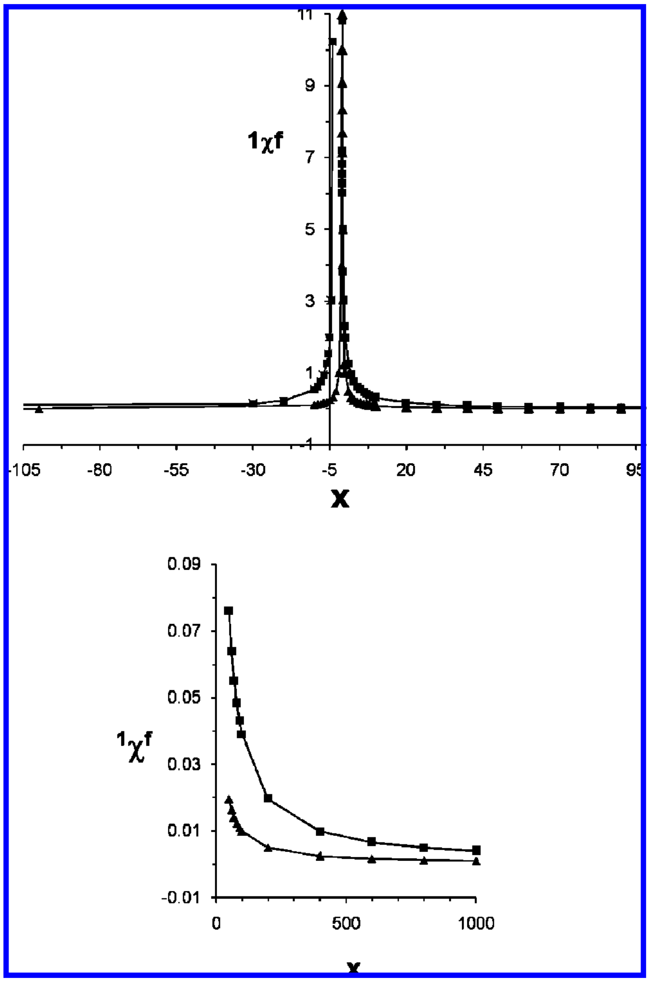


Figure 7. Variable connectivity index, $^1\chi^f$, vs x for the chemical graphs of ethane (triangles) and trimethylmethane (squares). (Bottom) Portion of the same figure for $100 \leq x \leq 1000$.

mented with higher-order components. These are characteristics that cannot be detected by leave-one-out methods[31−33]

2.2. Modeling with the Complete Graph and the Hydrogen-Perturbed Algorithms

2.2.1. Binding, $\Delta H_L°$, and Lattice, $\Delta H_G°$, Enthalpies and the Polarizability of Metal Halides

Metal halides offer a good opportunity to check the model qualities of the algorithm based on complete graphs for the core electrons, i.e., of eq 7, which, with no hydrogen atoms, simplifies into eq 3, since $f_\delta = 0$.[76] The subsets of MC basis indices for metal halides are (see the method section) as follows: $\{D^v, {}^0\chi^v, {}^1\chi^v\}$, $\{{}^S\psi_I, {}^0\psi_I, {}^1\psi_I, {}^S\psi_E, {}^0\psi_E, {}^1\psi_E\}$, and $\{{}^0\chi^v{}_d, {}^0\psi_{Id}, {}^0\psi_{Ed}, {}^1\chi^v{}_s, {}^1\psi_{Is}, {}^1\psi_{Es}\}$. The most affected set is the set of connectivity indices χ, which, due to the high degeneracy of the χ values, loses all nonvalence indices. The other subsets of basis indices are nearly unaffected by internal degeneracy, even if some indices do disappear because of redundancy with other indices. The experimental lattice and binding enthalpy values shown in Table 2 are taken from ref 75. This model[76] shows that it is possible to adhere to the guideline that the similar properties of a common set of compounds should be described with a common descriptor.

2.2.1.1. Binding Enthalpy (BE). Two descriptions give good quality for the binding enthalpy: the K_p-(p-odd) description gives the best higher-order Z and Z′ terms, while the K_p-(pp-odd) gives the best single-basis-index descriptor as well as the best two-basis-index descriptor. These K_p-(pp-odd) single-basis-index/two-basis-index pairs are also the best K_p-(pp-odd) descriptors for the lattice enthalpy. The binding enthalpy can be used as a training set for the binding and lattice enthalpy. In fact, the main descriptor for the binding enthalpy is also a good descriptor for the lattice enthalpy.

Table 2. Experimental, $\Delta H_L°$, $\Delta H_G°$, and Calculated, $^{ppo}\Delta H_L°$(C), and $^{po}\Delta H_G°$(C), Lattice and Binding Enthalpies, Respectively, at 298 K[a]

MeX	$\Delta H_L°$	$^{ppo}\Delta H_L°$(C)	$\Delta H_G°$	$^{po}\Delta H_G°$(C)
LiF	1046	1047	769	766
NaF	928	988	687	673
KF	826	829	588	614
RbF	792	785	563	572
CsF	756	756	566	545
LiCl	861	891	641	654
NaCl	787	788	559	565
KCl	717	733	503	516
RbCl	692	702	476	483
CsCl	668	682	475	463
LiBr	817	806	614	603
NaBr	751	724	539	521
KBr	689	681	474	480
RbBr	665	658	453	454
CsBr	649	643	440	437
LiI	761	762	570	571
NaI	703	690	505	493
KI	648	653	449	455
RbI	629	634	421	432
CsI	610	622	418	417

[a] All energies are in kJ mol^{-1}.

$q = 1$, K_p-(p-odd) description

$$Z'_{po}(B) = [{}^{po}Z(B) + 0.02({}^0\chi^v{}_d)]^{0.5}:$$
$$F = 1058, \quad r = 0.991, \quad s = 12, \quad N = 20, \quad \boldsymbol{u} = (33, 10)$$
$$\boldsymbol{C} = (1342.54, -252.262)$$

where $\boldsymbol{u}$ is the utility vector of $\boldsymbol{C}$, with $u_i = c_i/s_1$, and where

$$X_{po}(B) = (D^v)^{0.5}/[D^v + 12({}^0\chi^v)^{1.2}]^{0.3}:$$
$$F = 797, \quad r = 0.989, \quad s = 14, \quad N = 20$$

$$Y_{po}(B) = ({}^0\psi_E)^3:$$
$$F = 104, \quad r = 0.924, \quad s = 37, \quad N = 20$$

$$Z_{po}(B) = [X_{po}(B) + 0.2Y_{po}(B)]^{1.3}:$$
$$F = 962, \quad r = 0.991, \quad s = 13, \quad N = 20, \quad \boldsymbol{u} = (31, 9.8)$$
$$\boldsymbol{C} = (1168.20, 131.102)$$

$q = p$, K_p-(pp-odd) description

$$\{{}^0\psi_I\}: \quad F = 174, \quad r = 0.952, \quad s = 29, \quad N = 20$$

$$^{ppo}\{{}^0\psi_I, {}^1\psi_I\}: \quad F = 230, \quad r = 0.982, \quad s = 19, \quad N = 20$$

The calculated binding enthalpies, i.e., the $^{po}\Delta H_G°(C)$ of Table 2, have been obtained with the $^{po}Z'(B)$ term. Actually, the $^{po}Z(B)$ term with fewer adjustable parameters is more than enough to derive meaningful $^{po}\Delta H_G°(C)$ values. The maximum deviation from the experimental value is shown by KF, which deviates by 4.4%.

2.2.1.2. Lattice Enthalpy. The K_p-(p-odd) description shows the best two-index combination, while the K_p-(pp-

odd) description has the best X term while the good Y and higher-order terms do not show an improvement over the X term. The lattice enthalpy cannot be used as a training set for both types of enthalpies, since the resulting descriptor is unsatisfactory as an evaluation descriptor.

$q = 1$, K_p-(p-odd) description

$$^{po}\{^0\chi^v, {}^1\psi_E\}:\ F = 220,\quad r = 0.981,\ s = 22.2,\ N = 20$$

$q = p$, K_p-(pp-odd) description. The best single-basis-index/two-basis-index pairs are the same basis descriptors used for the binding enthalpy, i.e., $\{^0\psi_I\}$, with $r = 0.951$, and $\{^0\psi_I, {}^1\psi_I\}$ with $r = 0.978$. Nevertheless, the best descriptor for the lattice enthalpy is

$$X_{ppo}(L) = (D^v)^{0.4}/[0.8D^v + 30(^0\chi^v)^{1.2}]^{0.8}:$$
$$F = 1209,\quad r = 0.993,\quad s = 13.5,\quad N = 20,\quad \boldsymbol{u} = (35, 70)$$

$$\boldsymbol{C} = (5335.84, 514.953)$$

The $\Delta H_L°$ values calculated with the aid of the $^{ppo}X(L)$ term, i.e., $^{ppo}\Delta H_L°(C)$, are shown in Table 2. The only evident anomaly is the calculated value for LiCl, which differs by 5% from the experimental value. The origin of this anomaly is far from evident, since both the Li and Cl atoms give rise to more than decent residuals in other compounds.

2.2.1.3. Binding and Lattice Enthalpy: A Unique Description. The $Z'_{po}(B)$ term of the binding enthalpy is a good descriptor for the lattice enthalpy as well, as we can see from these statistics,

$$\Delta H_L° \text{ with } Z'_{po}(B):\quad F = 1010,\quad r = 0.991,\quad s = 15,\quad N = 20,\quad \boldsymbol{u} = (32, 5.9),\quad C = (1568.41, -170.545)$$

This interesting result lets us achieve the goal of modeling similar properties of a class of compounds (here the two enthalpy sets) with a common optimal descriptor. The experimental vs calculated plot for the lattice and binding enthalpies as well as the corresponding residual plot (bottom) is shown in Figure 8 for a total of 40 points. Here, the calculated values have been obtained with this $Z'_{po}(B)$ term. The maximum percent deviation is that for LiI, for which the enthalpy deviates by 3.9% from its experimental value.

2.2.1.4. Polarizability of Metal Halides. The polarizabilities of the metal halides whose lattice and binding enthalpies have just been described are collected in Table 3. Three sets of different polarizabilities obtained from ref 75 are collected in Table 3 and analyzed here. The first set contains the ionic polarizabilities calculated by adding the free ion values, α_i, to give the values shown in the second column. The second set contains the polarizabilities for the crystal, α_c, calculated from the refractive index according to $(n^2 - 1)/(n^2 + 2) = 4\pi\Sigma\alpha/3V$, where V is the volume per ion pair, to give the values shown in the fourth column. Finally, the third set contains the polarizabilities α_d for diatomics calculated from the equation $\Sigma\alpha = R_0^2(R_0 - 0.2082\mu_{\text{exp}})$, where R_0 is the experimental internuclear MeX distance and μ_{exp} is its experimental dipole moment, to give the values shown in the sixth column.

The first set, the ionic polarizability, will be used as a training set, while the other two sets will be modeled with the best descriptor for the ionic polarizability, in accord with the principle of modeling the same type of property of a common set of compounds with the same descriptor.[76] The ionic polarizability is the most difficult property to model. The overall model of these three properties with the given set of basis indices lets us detect the surprising fact that only within a K_p-(pp-seq) representation for the core electrons is it possible to find good descriptors for this property. Further, no good terms of any type could be found. The pseudoconnectivity index, $^1\psi_I$, plays the major role in this model.

2.2.1.4.1. Ion Polarizability, α_i,

$$^{pps}\{^1\psi_I\}:\quad F = 214,\quad r = 0.960,\quad s = 0.6,\quad N = 20,\quad \boldsymbol{u} = (15, 4),\quad \boldsymbol{C} = (28.7817, -4.62735)$$

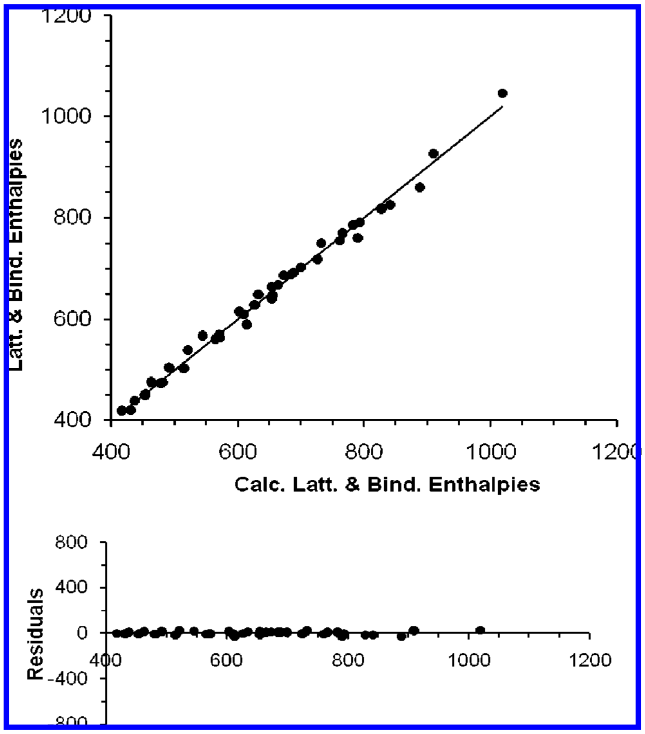


Figure 8. Experimental vs calculated (Calc) plot of the lattice and binding enthalpies ($N = 40$) for 20 metal halides together with the corresponding residual plot (bottom).

Table 3. Polarizabilities (in Å^3) of the Free Ion, α_i, of the Crystal, α_c, and of the Diatomic, α_d, and the Corresponding Calculated (C) Values, with K_p-Based Descriptors, for 20 Alkali Halides

MeX	α_i	α_i(C)	α_c	α_c(C)	α_d	α_d(C)
LiF	2.84	2.57	0.92	1.02	0.63	0.70
NaF	2.96	3.25	1.16	1.74	0.87	1.42
KF	3.60	3.89	2.01	2.40	1.85	2.08
RbF	4.16	4.27	2.58	2.80	2.57	2.48
CsF	5.16	4.52	3.61	3.05	3.91	2.73
LiCl	4.40	3.92	2.97	2.43	2.23	2.11
NaCl	4.52	4.74	3.29	3.28	2.75	2.96
KCl	5.16	5.49	4.17	4.07	3.82	3.75
RbCl	5.71	5.95	4.80	4.54	4.70	4.22
CsCl	6.71	6.24	5.89	4.85	6.33	4.53
LiBr	5.59	5.55	4.16	4.13	3.14	3.81
NaBr	5.71	6.52	4.42	5.14	3.81	4.82
KBr	6.35	7.41	5.35	6.07	4.88	5.75
RbBr	6.91	7.96	6.01	6.64	5.93	6.32
CsBr	7.90	8.31	7.00	7.00	7.74	6.68
LiI	8.31	6.83	6.22	5.47	4.85	5.15
NaI	8.43	7.93	6.54	6.60	5.84	6.29
KI	9.07	8.93	7.47	7.65	6.94	7.33
RbI	9.63	9.55	8.15	8.29	7.94	7.97
CsI	10.63	9.94	9.15	8.70	8.75	8.38

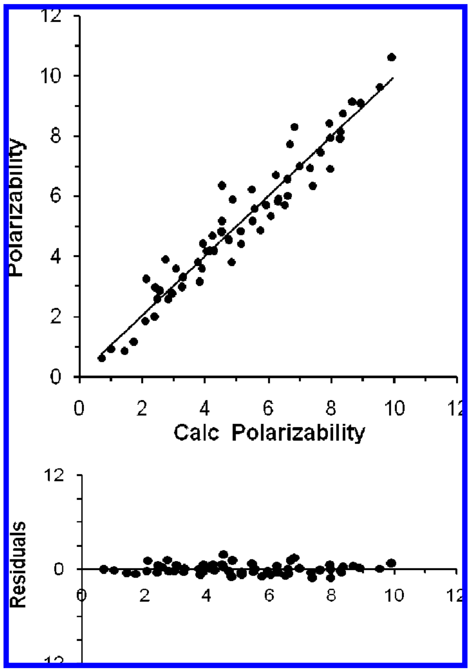


Figure 9. (Top) Experimental vs calculated (Calc) plot of three different types of polarizability ($N = 60$) for 20 metal halides. (Bottom) The corresponding residual plot.

This descriptor has been used to calculate the $\alpha_i(C)$ values shown in the third column of Table 3. From now on, to model the remaining two sets of polarizability values, we will use this basis index obtained with a K_p-(*pp*-seq) description.

2.2.1.4.2. Crystal Polarizability, α_c. The $\{^1\psi_I\}$ index shows even here the best statistical quality,

$$^{pps}\{^1\psi_I\}: \quad F = 373, \quad r = 0.977, \quad s = 0.5, \quad N = 20, \quad \boldsymbol{u} = (19, 6), \quad \boldsymbol{C} = (29.9731, -6.46902)$$

The quite interesting looking calculated $\alpha_c(C)$ values are given in the fifth column of Table 3.

2.2.1.4.3. Diatomic Polarizability, α_d. The $^1\psi_I$ index is not the best index here, but its statistical quality is not too different from the statistics of the best basis index here, which is $\{^0\psi_I\}$, with $F = 275$, $r = 0.969$, and $s = 0.6$,

$$^{pps}\{^1\psi_I\}: \quad F = 182, r = 0.954, s = 0.7, N = 20, \quad \boldsymbol{u} = (13, 6), \boldsymbol{C} = (29.9980, -6.79786)$$

The calculated $\alpha_d(C)$ values are collected in the seventh column of Table 3. The quite positive plot of the three different sets of polarizability values versus their corresponding calculated values is shown in Figure 9 together with their residual plot (bottom).

2.2.2. Polarizability of Organic Compounds

For this property, for which values from ref 77 are shown in Table 4, the hydrogen contribution must be included, which means that $f_\delta \neq 0$ for the valence delta should be tested. For these compounds, the ratio between the number of hydrogen atoms and the number of other atoms (heteroatoms, ht) is $n_H/n_{ht} = 1.2$. The optimal model for the polarizability α with molecular connectivity indices derived with a δ^v calculated first with $f_\delta = 0$,[51] $q = 1$, and $p =$ odd $= 1, 3, 5, ...$, i.e., with a K_p(*p*-odd) representation, which is here the best representation for the core electrons, is achieved with the following two types of descriptors,

$$^{po}\{^1\chi, D^v, {}^0\chi^v\}, \quad F = 1021, \quad r = 0.992, \quad s = 0.50, \quad N = 54$$

$$X'_{po}(\alpha) = (3^0\chi^v + 1.2^1\chi + 0.01^0\chi^v{}_d), \quad F = 2013, \quad r = 0987, \quad s = 0.6, \quad N = 54$$

A slightly improved model is obtained for $f_\delta \neq 0$ (ref 59) and $n = 8$ with the X' term. The K_p(*p*-odd) representation

Table 4. Experimental, α, and Computed, α(C) (with the Present Method) and α(M) (with the MM3 Method), Molecular Polarizability of Organic Compounds in Units of Å^3 [a]

compd	α	α(C)	α(M)	compd	α	α(C)	α(MM3)
*ethane	4.48	4.36	4.49	*acetaldehyde	4.59	4.75	3.91
propane	6.38	6.18	6.52	*acetone	6.39	6.88	6.09
neopentane	10.20	10.22	10.7	F-methane	2.62	3.17	2.82
*cyclopropane	5.50	5.21	6.1	*triF-methane	2.81	5.01	3.02
cyclopentane	9.15	9.00	10.28	tetraF-methane	2.92	3.32	2.96
cyclohexane	11.00	10.90	12.21	Cl-methane	4.55	4.46	4.41
*ethylene	4.12	3.33	3.5	*DiCl-methane	6.82	6.37	6.11
propene	6.26	5.39	5.54	triCl-methane	8.53	8.46	7.78
*2Me-propene	8.29	7.49	7.6	tetraCl-methane	10.51	10.61	9.37
**trans*-2-butene	8.49	7.54	7.59	*Br-methane	5.61	5.99	6.16
cyclohexene	10.70	10.37	11.26	*DiBr-methane	8.68	9.43	9.85
butadiene	7.87	6.52	7.32	triBr-methane	11.84	13.05	13.63
*benzene	9.92	9.36	10.91	*I-methane	7.59	7.55	8.06
toluene	12.30	11.39	12.99	di-I-methane	12.90	12.56	13.52
hexamethylbenzene	22.63	22.35	23.61	tri-I-methane	18.04	17.73	19.16
*acetylene	3.50	2.80	2.39	*CH_2=CCl_2	7.83	7.69	6.92
propyne	4.68	4.95	4.33	*cis*-CHCl=CHCl	7.78	7.73	6.93
$C(CCH)_4$	12.19	12.61	10.15	disilane	11.10	10.92	10.61
*allene	5.00	4.70	4.63	*formamide	4.08	3.94	3.32
methanol	3.32	3.31	3.35	*acetamide	5.67	6.12	5.52
ethanol	5.11	5.11	5.37	acetonitrile	4.48	4.67	3.56
*2-propanol	6.97	7.12	7.43	*propionitrile	6.24	6.60	5.64
cyclohexanol	11.56	11.87	13.1	pivalonitrile	9.59	10.69	9.76
dimethyl ether	5.24	5.55	5.48	benzylcyanide	11.97	12.41	11.82
**p*-dioxane	8.60	9.72	10.35	*triCl-acetonitrile	10.42	10.98	8.98
methylamine	3.59	3.58	4.02	*pyridine	9.92	9.15	9.39
formaldehyde	2.45	2.72	1.74	thiophene	9.00	8.70	10.24

[a] Asterisks indicate left-out compounds (evaluation set).

for the core electrons remains as the best representation for all of the chosen n values,

$$^{f8,po}\{^1\chi, D^v, {}^0\chi^v\},\quad F = 1018,\quad r = 0.992,\quad s = 0.51,\quad N = 54$$

$$^{f8}X'_{po}(\alpha) = (3{}^0\chi^v + 1.2{}^1\chi + 0.01{}^0\chi^v{}_d),\quad F = 2061,\quad r = 0.988,\quad s = 0.61,\quad N = 54,\quad \boldsymbol{u} = (46, 2.3)$$

$$\boldsymbol{C} = (0.69688, -0.46547)$$

With the correlation vector, $\boldsymbol{C}$, of $^{f8}X'_{po}(\alpha)$, the calculated values $\alpha(C)$ shown in the corresponding column in Table 4 have been obtained. In Table 4, the α(MM3) values obtained with the MM3 methods[77] are also included. Now, leaving out the 23 compounds marked with an asterisk in Table 4, i.e., every third compound, makes it possible to avoid changing the value of $n_H/n_{ht} = 1.2$. The new optimal model with the new set of training points for $f_\delta = 0$ and $f_\delta \neq 0$ is, respectively,

$$^{po}\{^1\chi, D^\nu, {}^0\chi^\nu\}:\quad F = 768,\quad r = 0.994,\quad s = 0.5,\quad N = 31$$

$$X'_{po}(\alpha) = (3{}^0\chi^\nu + 1.4{}^1\chi + 0.01{}^0\chi^\nu{}_d)^{0.9}:\quad F = 2157,\quad r = 0.993,\quad s = 0.5,\quad N = 31$$

$$^{f8,po}\{^1\chi, D^\nu, {}^0\chi^\nu\}:\quad F = 796,\quad r = 0.994,\quad s = 0.5,\quad N = 31$$

$$^{f8}X'_{po}(\alpha) = (3{}^0\chi^\nu + 1.3{}^1\chi + 0.01{}^0\chi^\nu{}_d):\quad F = 2329,\quad r = 0.994,\quad s = 0.5,\quad N = 31,\quad \boldsymbol{u} = (48, 2.6)$$

$$C = [0.69365, -0.55008]$$

Here, lowering the number of compounds increases a bit the divergence between $f_\delta = 0$ and $f_\delta \neq 0$ in favor of the $f_\delta \neq 0$ case. The number of data points N appears to be important for the importance of the proton perturbation. The calculated values of the training (N = 31) plus evaluation (N = 23) sets obtained with $^{f8}X'_{po}(\alpha$, for N = 31) and with its regression vector, $\boldsymbol{C}$, are very similar. Actually, the two descriptors are very similar, as are also their statistics for N = 54. Figure 10 enables comparison of the $^{f8}X'_{po}(\alpha$, N = 31) values (top) with the MM3 calculated values (bottom). The comparison underlines the good quality of the model.

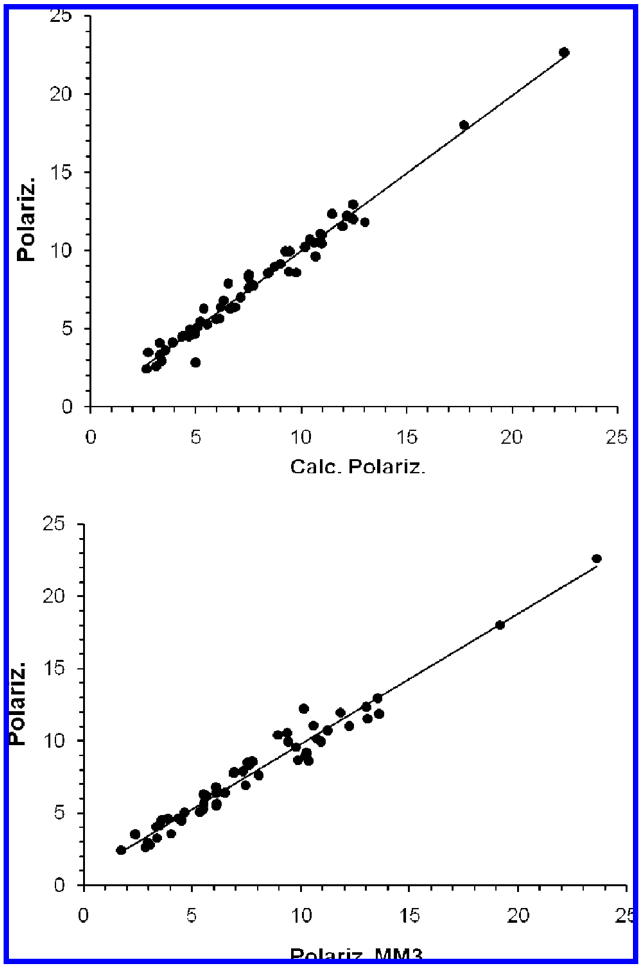


Figure 10. Plot of the experimental vs calculated values for 31 (training set) and 23 (evaluations set) polarizability values of organic compounds, and the corresponding plot of the MM3 calculated values (bottom).

2.2.3. Partition Coefficients of Halogenated Organic Compounds

The partition coefficients (log P) of 25 halocompounds in six different media are collected in Table 5 (These data are taken from ref 78; but see also ref 54). This property has recently been studied by other authors with different approaches.[79-81] The halogenated compounds studied have $n_H/n_{ht} = 0.5$. Here again, as was done with the metal halides, we will try to model similar properties of a common set of compounds with similar descriptors. Here, the K_p-(p-seq) and K_p-(pp-seq) descriptions do not show any optimal model quality at all, while the K_p-(p-odd) and K_p-(pp-odd) descriptions show, as already seen with the enthalpy of metal halides, the best model qualities. Three subgraph indices will be used throughout the log P model: $D_F = \sum_i\delta_F$, $D_{Cl} = \sum_i\delta_{Cl}$, and $D_{Br} = \sum_i\delta_{Br}$, for the F, Cl, and Br atoms, respectively. Since the liver partition coefficient is the set of values which is described best among all of the sets of log P values, the descriptor for log P(liver) will be used to model all of the other five sets of log P values.

All the log P values in different media will now be modeled with the best linear combination of basis indices (LCBI) derived from the liver case, which is also the best descriptor for the saline and blood cases. The model requires a K_p-(p-odd) representation (with q = 1) of the core electrons. With no hydrogen perturbation, i.e., with $f_\delta = 0$,[54] the best LCBI model for log P(liver), in which the best single basis index $^1\chi^\nu$ does not contribute, is

$$^{po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}:\quad F = 52,\quad r = 0.955,\quad s = 0.2,\quad N = 25,\quad \boldsymbol{u} = (7.9, 9.5, 4.8, 3.2, 5.1)$$

$$\boldsymbol{C} = (-2.4135, 2.74399, 0.70903, 0.22563, 1.10583)$$

If we consider as strongly correlated only those indices with $r > 0.98$,[82] then the four basis indices of $\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ are all poorly correlated, since $r(^0\chi, {}^0\psi_I) = 0.94$, $r(^0\chi, D_F) = 0.37$, $r(^0\chi, D_{Cl}) = 0.61$, $r(^0\psi_I, D_F) = 0.08$, $r(^0\psi_I, D_{Cl}) = 0.74$, and $r(D_F, D_{Cl}) = 0.37$. The $F - r - s - q^2$ values for

Table 5. Liquid and Rat Tissue Air Partition Coefficient (log *P*) in Different Media at 37 °C, for 25 Halocompounds

molecule	saline	olive oil	blood	liver	liver-10-out (clc)	muscle	fat
CH_3Cl	−0.056	0.933	0.393	0.540	0.574	−0.013	1.130
*CH_2Cl_2	0.775	2.117	1.288	1.152	0.993	0.899	2.079
$^+CHCl_3$	0.529	2.604	1.318	1.324	1.168	1.143	2.307
CCl_4	−0.456	2.573	0.655	1.152	1.257	0.657	2.555
$CH_2=CHCl$	−0.367	1.387	0.225	0.204	0.173	0.342	1.301
$CCl_2=CH_2$	−0.456	1.808	0.699	0.645	0.576	0.312	1.836
*$CHCl=CHCl$(cis)	0.512	2.444	1.334	1.185	0.965	0.785	2.356
$^+CHCl=CHCl$(tr)	0.149	2.250	0.981	0.952	0.965	0.547	2.170
$CCl_2=CHCl$	−0.081	2.743	1.340	1.435	1.382	1.004	2.744
$CCl_2=CCl_2$	−0.102	3.329	1.276	1.847	1.814	1.301	3.214
CH_3-CH_2Cl	0.037	1.590	0.611	0.558	0.480	0.508	1.587
*$CHCl_2-CH_3$	0.389	2.270	1.049	1.033	0.911	0.709	2.215
$^+CH_2Cl-CH_2Cl$	1.057	2.563	1.483	1.553	1.477	1.369	2.537
CCl_3-CH_3	−0.125	2.470	0.760	0.934	1.058	0.498	2.420
$CHCl_2-CH_2Cl$	1.124	3.249	1.763	1.863	1.861	1.360	3.158
$CHCl_2-CHCl_2$	1.369	3.803	2.152	2.292	2.279	2.004	3.576
*CCl_3-CH_2Cl	0.548	3.429	1.620	1.945	2.008	1.597	3.332
$^+CH_2F_2$	0.117	0.678	0.204	0.439	0.530	0.158	0.155
CH_2FCl	0.489	1.348	0.706	0.537	0.599	0.391	1.188
CH_2BrCl	0.937	2.558	1.618	1.465	1.356	1.045	2.512
CH_2Br_2	1.158	2.981	1.870	1.833	1.763	1.607	2.899
*$CF_3-CHClBr$	−0.301	2.297	0.721	0.882	0.600	0.649	2.260
$^+CH_2=CHBr$	−0.357	1.748	0.607	0.522	0.249	0.354	1.692
CH_2Br-CH_2Cl	0.950	2.755	1.722	1.631	1.745	1.405	2.982
CF_3-CH_2Cl	−0.377	1.380	0.104	0.265	0.289	0.090	1.326

[a] Asterisk and superscript plus indicate left out compounds in two different leave-out methods.

Table 6. $F/r/s/q^2$ Values Due to $^{po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ for $f_\delta^2 = 0$ and $\neq 0$, and $N = 25$

f_δ^2	saline	blood	liver	oil	muscle	fat
0	$26/0.917/0.2_5/0.765$	$38/0.941/0.2_5/0.828$	52/0.955/0.2/0.850	99/0.976/0.2/0.935	41/0.944/0.2/0.838	150/0.984/0.1/0.965
≠0	19/0.888/0.3/0.689	53/0.956/0.2/0.873	$81/0.970/0.1_5/0.902$	219/0.989/0.1/0.964	48/0.951/0.2/0.862	283/0.991/0.10.968

Table 7. $F/r/s$ Values Due to $^{po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$ for $f_\delta^2 = 0$ and $\neq 0$, and $N = 15$

f_δ^2	saline	blood	liver	oil	muscle	fat
0	11/0.905/0.3	19/0.941/0.3	27/0.957/0.2	47/0.975/0.2	21/0.944/0.2	88/0.986/0.2
≠0	9.1/0.886/0.4	25/0.954/0.2	48/0.975/0.2	$96/0.987/0.1_5$	29/0.960/0.2	281/0.996/0.1

both zero and nonzero f_δ^2 are given in Table 6. Normally the introduction of f_δ^2 improves the model,[59] especially for the blood, oil, liver, and fat cases. The model quality begins to decrease for $n \geq 4$, except for the saline case, which has $F = 25$ at $n = 4$.

Let us now leave-10-out (with + and * in Table 5 but keeping $n_H/n_{ht} = 0.56$) and work with a training set of 15 compounds. For this training choice, the best descriptor for the saline, blood, and liver case is again the best overall descriptor, which for both the zero-*H*-perturbation ($f_\delta^n = 0$) and the nonzero-*H*-perturbation ($f_\delta^2 \neq 0$) is again $^{po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$, with the statistics shown in Table 7. Excluding the fat case, where the $f_\delta^2 = 0$ and $f_\delta^2 \neq 0$ descriptions diverge with the $f_\delta^2 \neq 0$ case being favored as *N* gets smaller, the other log *P* cases show practically no consistent changes with *N*, an interesting, albeit expected, result with n_H/n_{ht} = constant, but not one which always occurs. Figure 11 shows the model from 80 training points plus 20 external validated points (i.e., the five left-out points marked with an asterisk in Table 5 for each of four sets of log *P*) obtained with the correlation coefficients for the four indices $^{f2,po}\{^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$, which are here collected into a vector form. In these vectors, the corresponding utility is given in parentheses after each component index.

$$\boldsymbol{C}(\mathrm{Bl}) = [-2.81203\ (8.2),\ 3.15516\ (9.9),\ 0.74932\ (4.7),\ 0.10050\ (1.2),\ 1.68496\ (5.8)]$$

$$\boldsymbol{C}(\mathrm{Liv}) = [-2.66129\ (8.7),\ 2.99186\ (11),\ 0.77664\ (5.5),\ 0.19568\ (2.7),\ 1.47851\ (5.7)]$$

$$\boldsymbol{C}(\mathrm{Oil}) = [-2.08605\ (8.3),\ 2.85394\ (12),\ 0.37152\ (3.2),\ 0.03685\ (0.6),\ 1.37054\ (6.5)]$$

$$\boldsymbol{C}(\mathrm{Fat}) = [-1.78911\ (8.6),\ 2.66236\ (14),\ 0.14340\ (1.5),\ -\ 0.07069\ (1.4),\ 1.05050\ (6.0)]$$

The K_p-(*pp*-odd) description with $q = p$ of this property with a combination of basis indices is rather unsatisfactory. This desription shows, instead, an optimal *X* term for the liver case and for other three cases with $f_\delta = 0$, $f_\delta^6 \neq 0$, and $N = 15$. This term is actually also a good descriptor for all other cases. Its statistics in the different media are shown in Table 8

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

S_F and S_{Cl} have been calculated using eq 16; when several

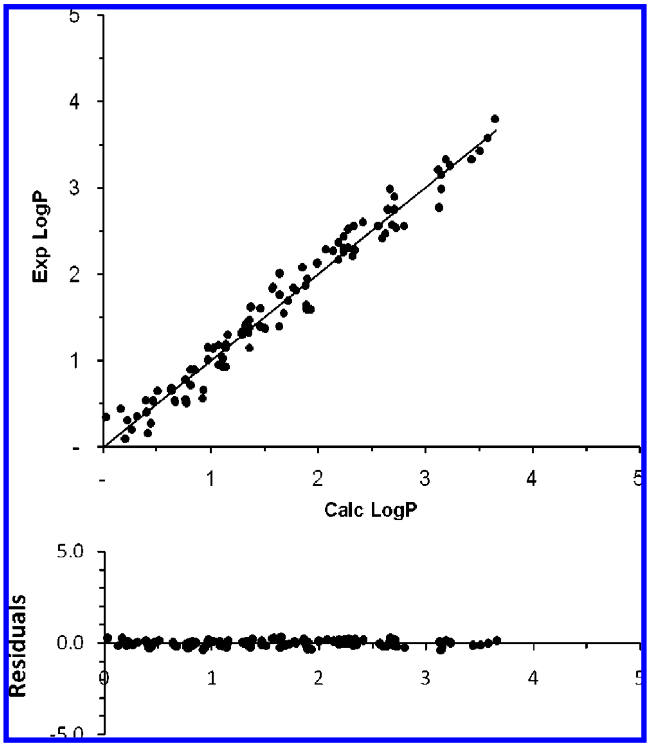


Figure 11. (Top) Model of 80 training and 20 evaluated points of log P in four different media (blood, liver, oil, and muscle), with the optimal combination ${}^{f2}\{{}^0\chi, {}^0\psi_I, D_F, D_{Cl}\}$, which has been obtained from the liver case only. (Bottom) The corresponding residual plot.

halogen atoms are present, the average of the S values of each halogen atom has been used. Regardless of the media, the exclusion of *cis*-1,2-dichloroethylene normally improves the model.[54,59] This improvement is in fairly good accord with a proposed graph-theoretical representation of *cis*−*trans* isomerism.[1,83] The statistics for this term with or without hydrogen perturbation are rather similar. Let us leave out the 10 previous compounds, i.e., those with * and + in Table 5, and redo the model of this training set of 15 compounds. For the new training set for both the $f_\delta = 0$ and $f_\delta{}^8$ perturbation, the best term is the following term, for which statistics are shown in Table 8 and also which for $f_\delta{}^8$ is slightly different in this form than the previous term.

$${}^{f8}X_{ppo}(L)_{n=15} = [3D_{Cl} + 2.5(D_{Br}) + 0.6S_F - 0.06S_{Cl} - 0.3({}^0\chi)^{1.8}]^{0.8}({}^1\chi)^{3.1}/(D^v)^{1.9}$$

$$F = 388,\ r = 0.978,\ s = 0.1,\ N = 20,\ \boldsymbol{u} = (20, 2.2),\ \boldsymbol{C} = (42509, 0.12397)$$

Except in the liver case, in which the hydrogen perturbation is clearly dependent on N, since it increases for $N = 15$, the statistics are rather unaffected by the perturbation. Thus, excluding the liver case, the X term is not affected by the hydrogen perturbation. In the fifth column of Table 5, the training ($N = 15$) and evaluation ($N = 10$) calculated log P(liver) values have been obtained with the correlation vector of ${}^{f8}X_{ppo}$(Liv), $\boldsymbol{C} = (8.80259, -0.2952)$, with the utility of the regression parameters being $\boldsymbol{u} = (32, 6.0)$. The values found are quite satisfactory.

2.2.4. Three Properties of Halomethanes, CH_nX_{4-n}

The molar refraction R_m, boiling point BP, and parachor Pa for halomethanes are collected in Table 9.[50,84] For these physical properties, the K_p-(p-odd) representation for the core electrons gives the best results. For R_m and BP, we have the same n_H/n_{ht} and N values, but N changes for Pa.

2.2.4.1. Molar refraction, R_m. Here, $n_H/n_{ht} = 0.2$. A previous model of this property[50,51] with $f_\delta = 0$ had the following optimal descriptors,

$${}^{po}\{D, {}^0\psi_I\}:\quad F = 3836,\quad r = 0.9980,\quad s = 0.6,\quad N = 34$$

$$Z'_{po}(R_m) = (X + 1.3Y + 0.3{}^1\psi_{IS}):\quad F = 22805,\quad r = 0.9993,\quad s = 0.3,\quad N = 34$$

Here, $X_{po} = ({}^0\chi^v + \chi_t)^{0.6}$ ($F = 1092$), and $Y_{po} = (0.97{}^0\psi_I - 1.15{}^0\psi_E)^{1.2}$ ($F = 2959$). With $f_\delta \neq 0$,[59] no better combination of indices could be found, but instead, a Z term with improved quality was found:

$${}^{f8}Z_{po}(R_m) = ({}^{f8}X_{po} + {}^{f8}Y_{po})^{0.9}:\ F = 52939,\quad r = 0.9997,\quad s = 0.2,\quad N = 34,\quad \boldsymbol{u} = (52, 84)$$

$$\boldsymbol{C} = (12.9475, -11.9537)$$

Here, ${}^{f8}X_{po} = [0.7{}^0\chi^v + 0.9(\chi_t)^{0.8}]^{0.5}$ ($F = 1132$), and ${}^{f8}Y_{po} = |(0.87{}^0\psi_I - 1.1{}^0\psi_E)|^{1.2}$ ($F = 3422$), with vertical bars here denoting the absolute value. For $n = 6, 4, 2$, the statistics of the Z term decrease slowly but steadily. If we leave out the 15 compounds marked with an asterisk in Table 9 but keep $n_H/n_{ht} = 0.2$, then for $N = 19$, the optimal combination is an $f_\delta = 0$ combination not bettered by any $f_\delta \neq 0$ combination, and the advantage of the ${}^{f8}Z$ term shrinks, as indicated by ${}^{f8}F/{}^{f0}F = 1.05$, with the same r and s:

$${}^{f8,po}\{D, {}^1\psi_I\}:\quad F = 2586,\quad r = 0.9985,\quad s = 0.5,\quad N = 19$$

$${}^{f8}Z_{po}(R_m) = (X + 2Y):\quad F = 16723,\quad r = 0.9995,\quad s = 0.3,\quad N = 19,\quad \boldsymbol{u} = (129, 30)$$

$$\boldsymbol{C} = (6.24058\ (129), -6.22364\ (30))$$

Here, ${}^{f8}X_{po} = [0.7{}^0\chi^v + 0.9(\chi_t)^{0.8}]^{0.7}$, and ${}^{f8}Y_{po} = |(0.83{}^0\psi_I - {}^0\psi_E)|$. Thus, lowering N lowers the importance of the hydrogen perturbation.

2.2.4.2. Boiling Points, BP. Here $n_H/n_{ht} = 0.2$. The best description[50,51] with no f_δ contribution was

$${}^{po}\{{}^0\chi, {}^0\psi_I, {}^S\psi_E\}:\quad F = 4938,\quad r = 0.999,\quad s = 3.6,\quad N = 34$$

Table 8. *F*/*r*/*s* Statistics of ${}^{fn}X_{ppo}$(Liv) for $f_\delta = 0$ and $f_\delta{}^n \neq 0$, and for $N = 25$, and $N = 15$

$f_\delta{}^n$	N	saline	blood	liver	oil	muscle	fat
$f_\delta = 0$	25	18/0.66/0.4	106/0.906/0.2	580/0.981/0.1	181/0.942/0.3	157/0.934/0.2	155/0.953/0.3
$f_\delta{}^6 \neq 0$	25	17/0.66/0.4	105/0.906/0.2_5	598/0.981/0.1	194/0.945/0.3	161/0.935/0.2	158/0.934/0.3
$f_\delta = 0$	15	13/0.70/0.5	74/0.922/0.3	780/0.992/0.1	142/0.957/0.3	99/0.940/0.2	174/0.965/0.2
$f_\delta{}^8 \neq 0$	15	13/0.71/0.5	84/0.930/0.2	1033/0.994/0.1	133/0.954/0.3	102/0.942/0.2	175/0.965/0.2

Table 9. Molar Refraction, R_m/cm^3·mol^{-1}, Boiling Points, BP/K, and Parachor, σ, of Halomethanes

CH_nX_{4-n}[a]	R_m/cm^3·mol^{-1}	BP/K	σ
*CH_3F (+)	6.7	195	78
**CH_2F_2	6.6	221	87
CHF_3	6.5	189	96
*CF_4	6.4	144	105
*CH_3Cl	11.7	249	
CH_2Cl_2	16.6	313	
*$CHCl_3$	21.5	335	
*CCl_4	26.4	349	221
**$CFCl_3$	21.4	297	192
CF_2Cl_2	16.4	243	163
$CHFCl_2$	16.4	282	
*CHF_2Cl	11.5	233	
CF_3Cl	11.4	192	134
*CH_2FCl	11.6	264	
CH_3Br	14.6	277	
CH_2Br_2	22.4	370	
*$CHBr_3$	30.2	422	
CBr_4	38.0	462	277
CF_3Br	14.3	214	
CF_2Br_2	22.2	298	
*$CFBr_3$	30.1	381	
CCl_3Br	29.3	378	
*CCl_2Br_2	32.2	408	
$CClBr_3$	35.1	433	
CH_2FBr	14.5	291	
*CH_2ClBr	19.5	342	
$CHFClBr$	19.4	309	
CHF_2Br	14.4	259	
*$CHFBr_2$	22.3	338	
$CHCl_2Br$	24.4	361	
$CHClBr_2$	27.3	394	
*$CFCl_2Br$	24.3	326	
*$CFClBr_2$	27.2	353	
*CF_2ClBr	19.3	269	

[a] An asterisk indicates left-out values. For compounds with ** and (+), see text.

With an $f_\delta{}^6$ contribution,[59] the model improves a bit,

$$^{f6,po}\{^0\chi,\ {}^0\psi_I,\ {}^S\psi_E\}: \quad F = 5426, \quad r = 0.9991, \quad s = 3.5, \quad N = 34, \quad \boldsymbol{u} = (22, 43, 8.1, 84)$$

$$\boldsymbol{C} = (-163.432, 189.053, 2.57254, 269.697)$$

Let us again omit the 15 marked compounds (*) of Table 9, which avoids changes in n_H/n_{ht}, as was already done for R_m. With a training set of only 19 compounds, the $f_\delta = 0$ and $f_\delta \neq 0$ optimal descriptors are, respectively,

$$^{po}\{^0\chi,\ {}^0\psi_I,\ {}^S\psi_E\}: \quad F = 1990, \quad r = 0.9987, \quad s = 4.3, \quad N = 19$$

$$^{f6,po}\{^0\chi,\ {}^0\psi_I,\ {}^S\psi_E\}: \quad F = 2046, \quad r = 0.9988, \quad s = 4.2, \quad N = 19, \quad \boldsymbol{u} = (13, 23, 4.6, 54)$$

$$\boldsymbol{C} = (-166.262, 189.922, 2.67257, 273.529)$$

The $f_\delta \neq 0$ and $f_\delta = 0$ descriptions are in this case much closer to each other, which is to say decreasing N decreases the divergence between these descriptors. The different behaviors of the hydrogen perturbation for these two properties between compounds having the same %H shows that $f_\delta{}^n$ is property dependent, a fact already detected with the different behaviors of log$P$ in different media.

2.2.4.3. Parachor, Pa. In this case, $n_H/n_{ht} = 0.15$, and the best $f_\delta = 0^{50}$ descriptor was

$$^{po}\{D,\ {}^1\psi_I\}: \quad F = 15653, \quad r = 0.99990, \quad s = 1.1, \quad q^2 = 0.9993, \quad N = 9$$

With an $f_\delta{}^6$ contribution,[59] the description improves to

$$^{f6}\{D,{}^1\psi_I\}: \quad F = 17885, \quad r = 0.99992, \quad s = 1.0, \quad q^2 = 0.9993, \quad N = 9$$

2.2.4.4. Composite [R$_m$ + BP + Pa] Property. The $^{f8}\{D,{}^1\psi_I\}$ and $^{f8}\{D,\ {}^0\chi,{}^1\psi_I\}$ descriptors, which share two basis indices, are the best descriptors for all these three properties together,[50,51,59]

$$R_m: \quad {}^{f8}\{D,\ {}^1\psi_I\}: \quad F = 3655, \quad r = 0.998, \quad s = 0.6, \quad N = 34$$

$$\text{BP}: \quad {}^{f8}\{D,\ {}^0\chi,{}^1\psi_I\}: \quad F = 2386, \quad r = 0.998, \quad s = 5.2_5, \quad N = 34$$

$$\text{Pa}: \quad {}^{f8}\{D,\ {}^1\psi_I\}: \quad F = 17014, \quad r = 0.99991, \quad s = 1.0, \quad N = 9$$

The corresponding best $f_\delta = 0$ composite [R_m, BP, Pa] description is

$$R_m: \quad \{D,\ {}^0\psi_I\}: \quad F = 3836, \quad r = 0.998, \quad s = 0.6, \quad N = 34$$

$$\text{BP}: \quad \{D,\ {}^0\chi,\ {}^0\psi_I\}: \quad F = 2547, \quad r = 0.998, \quad s = 5.2, \quad N = 34$$

$$\text{Pa}: \quad \{D,\ {}^0\psi_I\}: \quad F = 6673, \quad r = 0.9998, \quad s = 1.7, N = 9$$

The two descriptions are nearly equivalent. If we leave out the compounds marked with an asterisk in Table 9 so that the %H does not change, then from the resulting training set we obtain the following model qualities for the best $f_\delta{}^n \neq 0$ ($n = 8$) and $f_\delta = 0$ cases, respectively, for which the types of descriptors do not change.

$$f_\delta{}^8 \neq 0: \quad R_m: \quad F = 2498, \quad r = 0.9984, \quad s = 0.5, \quad N = 19$$

$$\text{BP}: \quad F = 1647, \quad r = 0.9984, \quad s = 4.7, \quad N = 19$$

$$\text{Pa}: \quad F = 15481, \quad r = 0.99995, \quad s = 0.9, \quad N = 6$$

$$f_\delta = 0: \quad R_m: \quad F = 2122, \quad r = 0.9981, \quad s = 0.6, \quad N = 19$$

$$\text{BP}: \quad F = 1095, \quad r = 0.9977, \quad s = 5.8, \quad N = 19$$

$$\text{Pa}: \quad F = 19699, \quad r = 0.99996, \quad s = 0.8, \quad N = 6$$

Globally, the two descriptions diverge with a negative divergence for Pa, and usually the $f_\delta{}^8 \neq 0$ description improves over the $f_\delta = 0$ description. The N-dependence of the hydrogen perturbation is more convoluted than simple arguments based on %H seem to indicate. Since D and $^0\chi$ are strongly correlated, as indicated by a correlation of $r =$

Table 10. Boiling Points, T_b/K, of Primary Amines and Alcohols

amine	T_b/K	alcohol[a]	T_b/K
CH_3-	256.6_5	$(CH_3)_2CH-$	355.5_5
CH_3CH_2-	290.1_5	$CH_3CH_2CH_2-$	370.2_5
$(CH_3)_2CH-$	307.1_5	$CH_3CH_2C(CH_3)_2-$	375.4_5
$CH_3CH_2CH_2-$	322.1_5	$CH_3CH(CH_3)CH_2-$	381.2_5
$CH_3CH_2CH(CH_3)-$	336.1_5	$CH_3(CH_2)_3-$	390.7_5
$CH_3CH(CH_3)CH_2-$	341.1_5	$CH_3CH_2CH_2CH(CH_3)-$	392.0_5
$CH_3CH_2CH_2CH_2-$	350.9_5	$CH_3C(CH_3)_2CH(CH_3)-$	393.5_5
$CH_3CH_2C(CH_3)_2-$	351.1_5	$CH_3(CH_2)_2C(CH_3)_2-$	396.1_5
$(CH_3CH_2)_2CH-$	364.1_5	$CH_3CH(CH_3)CH(CH_3CH_2)-$	400.6_5
$CH_3CH_2CH_2CH(CH_3)-$	365.1_5	$CH_3CH_2CH(CH_3)CH_2-$	402.0_5
$CH_3CH(CH_3)CH_2CH_2-$	368.1_5	$CH_3CH(CH_3)CH_2CH_2-$	405.1_5
$CH_3C(CH_3)_2CH(CH_3)-$	375.1_5	$CH_3CH(CH_3)CH_2CH(CH_3)-$*	406.1_5
$CH_3(CH_2)_4-$	377.5_5	$(CH_3CH_2)_2C(CH_3)-$	409.1_5
$CH_3(CH_2)_3CH(CH_3)-$	390.6_5	$CH_3CH_2C(CH_3)_2CH_2-$	409.8_5
$CH_3(CH_2)_5-$	403.1_5	$CH_3(CH_2)_4-$	411.1_5
$CH_3CH_2CH(CH_3)CH_2CH(CH_3)-$	406.6_5	$(CH_3CH(CH_3))_2CH-$	413.1_5
$CH_3(CH_2)_4CH(CH_3)-$	415.1_5	$(CH_3CH_2)_3C-$	415.1_5
$CH_3(CH_2)_6-$	430.0_5	$CH_3CH(CH_3)CH(CH_3)CH_2-$	418.1_5
$CH_3(CH_2)_7-$	449.1_5	$CH_3CH_2CH_2CH(CH_3)CH_2-$	421.1_5
$CH_3(CH_2)_8-$	465.1_5	$CH_3CH_2CH(CH_3)CH_2CH(CH_3)-$*	432.9_5
$CH_3(CH_2)_9-$	490.1_5	$(CH_3CH_2)_2(CH_3)C-$	434.1_5
		$(CH_3(CH_2)_3)(CH_3CH_2)(CH_3)C-$	436.1_5
		$CH_3(CH_2)_6-$	449.9_5
		$CH_3(CH_2)_5C(CH_3)_2-$	451.1_5
		$(CH_3CH_2CH_2)_2(CH_3CH_2)C-$	455.1_5
		$CH_3CH(CH_3)CH_2(CH_2)_4-$	461.1_5
		$CH_3(CH_2)_7-$	467.5_5

[a] Asterisks indicate two compounds that had (wrongly) the same set of indices in a previous work.

0.998, it might be advantageous to orthogonalize them with a stepwise orthogonalization procedure.[85−87]

2.2.5. Boiling Points of Amines and Alcohols

The experimental values are collected in Table 10 and are taken from ref 61. For amines and alcohols, K_p has $p = 1$. For $f_\delta = 0$, the two classes of compounds show different model quality.[51] The amines yield a high quality model, and the alcohols yield a lower quality model. The results for $f_\delta^n \neq 0$[59,88] are deceiving if each class is treated separately, but the results are very interesting if both classes are modeled together. In that case, the percent of hydrogen atoms, with $n_H/n_{ht} = 2.1$, is the highest among those for all the properties examined here. For $f_\delta = 0$, we obtain the following best single-, two-index, and term descriptors,

$$\{{}^1\chi\}:\quad F = 284,\quad r = 0.928,\quad s = 17.9,\quad N = 48$$

$$\{D, \chi_t\}:\quad F = 165,\quad r = 0.938,\quad s = 16.7,\quad N = 48$$

$$X(Am + Al) = [(\chi_t)^{0.3} + 0.05\,{}^1\chi^v]:\quad F = 383,\quad r = 0.946,\quad s = 15.7,\quad N = 48$$

Improved single- and two-index descriptors are obtained for $f_\delta^2 \neq 0$, and for $f_\delta^8 \neq 0$, an improved term is obtained:

$$^{f2}\{D^v\}:\quad F = 484,\quad r = 0.956,\quad s = 14.1,\quad N = 48$$

$$^{f2}\{D^v, {}^0\chi^v\}:\quad F = 389,\quad r = 0.972,\quad s = 11.3,\quad N = 48$$

$$^{f8}X(Am + Al) = [(\chi_t^{\ v})^{0.2} + 0.03\,{}^1\chi^v]:\quad F = 1150,\quad r = 0.981,\quad s = 9.4,\quad N = 48,\quad \boldsymbol{u} = (34, 58)$$

$$\boldsymbol{C} = (-768.436, 951.991)$$

The improvement achieved is substantial, especially for the $^{f8}X(Am + Al)$ term.

To determine whether the improvement in the results is due only to the increase in N from the separate class sizes of 21 and 27 to the combined class size of 48, every second compound starting with the first one in both classes of amines and alcohols in Table 10 (1°, 3°, 5°, ...) is excluded from the model. This gives 10 training amines and 13 training alcohols, and n_H/n_{ht} does not change. The best descriptors for $f_\delta \neq 0$ and $f_\delta = 0$ for these $N = 23$ training points are then

$$^{f2}\{D^v\}:\quad F = 212,\quad r = 0.954,\quad s = 13.3;\quad \{\chi_t\}:\quad F = 131,\quad r = 0.928,\quad s = 16.5,\quad N = 23$$

$$^{f2}\{D^v, {}^0\chi^v\}:\quad F = 172,\quad r = 0.972,\quad s = 10.7;\quad \{{}^1\chi, {}^1\chi^v\}:\quad F = 104,\quad r = 0.955,\quad s = 13.5,\quad N = 23$$

$$^{f8}X'(Am + Al) = [(\chi_t^{\ v})^{0.2} + 0.03\,{}^1\chi^v]^{1.4}:\quad F = 693,\quad r = 0.985,\quad s = 7.6,\quad N = 23,\quad \boldsymbol{u} = (26, 55)$$

$$\boldsymbol{C} = (-599.922, 777.008)$$

$$X'(Am + Al) = [(\chi_t)^{0.4} + 0.08\,{}^1\chi^v]^{0.6}:\quad F = 192,\quad r = 0.949,\quad s = 13.9,\quad N = 23$$

The importance of the hydrogen perturbation for these three descriptors determined with a smaller training set is practically the same, and the optimal term is only slightly different from the one previously determined with the larger $N = 48$ set. Thus, in this mixed class of compounds, changes in N do not affect the importance and type of the hydrogen perturbation. The model of the 23 training points plus the

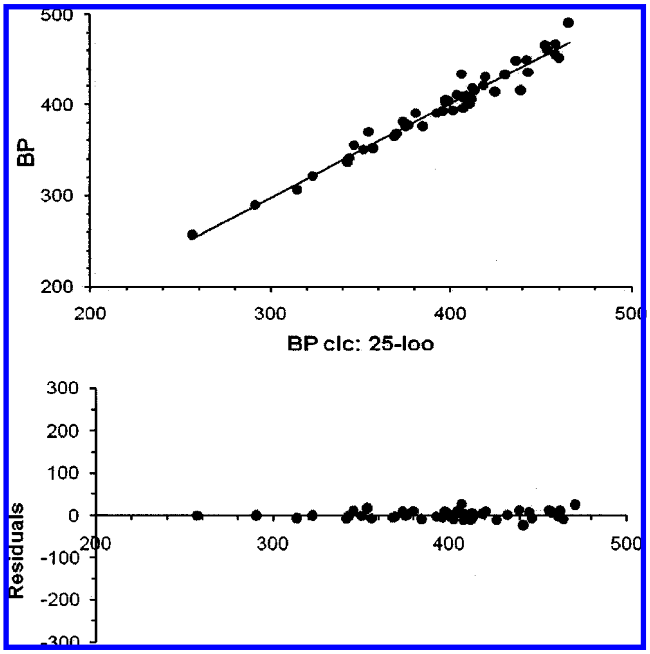


Figure 12. Model of a training set of 23 boiling points and an evaluation set of 25 boiling points of alcohols and amines; at the bottom is the corresponding residual plot.

25 evaluating points with $^{f8}X'(Am + Al)$ and its ***C*** vector is hardly different from the model of the entire set of 48 points considered as a training set (vide supra). The model is displayed in Figure 12. In fact, this last term for $N = 48$ has $F = 1143$, $r = 0.980$, $s = 9.4$.

2.3. Model of Properties with the Variable Descriptor

The values taken from ref 69 for the ether toxicity, pC, of 21 ethers are collected in Table 11. Because an oxygen atom is present, algorithm 16 was used with $y_H = y_O$. The statistics corresponding to four different (x, y_O) values are shown in the following together with the magnitudes of the contributing bond terms (*cbt*) to $^1\chi^f$ for the CH_3-CH_2 bond, where the variability is contributed only by x.

Table 11. Ether Toxicity, p*C*, Computed Ether Toxicity, p*C*$_{clc}$, with x = 1000, and y_O = −1, and the Corresponding Residuals, res

ether	p*C*	p*C*$_{clc}$	res
dimethyl	1.43	1.37	0.06
methyl ethyl	1.74	1.89	−0.15
diethyl	2.22	2.32	−0.1
methyl isopropyl	2.26	2.32	−0.06
methyl propyl	2.45	2.32	0.13
ethyl propyl	2.60	2.65	−0.05
ethyl isopropyl	2.60	2.65	−0.05
methyl butyl	2.70	2.66	0.04
methyl isobutyl	2.79	2.66	0.13
methyl *sec*-butyl	2.79	2.66	0.13
methyl *tert*-butyl	2.79	2.65	0.14
methyl pentyl	2.88	2.89	−0.01
ethyl butyl	2.82	2.89	−0.07
ethyl isobutyl	2.82	2.89	−0.07
ethyl *sec*-butyl	2.85	2.88	−0.03
ethyl *tert*-butyl	2.92	2.88	0.04
dipropyl	2.79	3.02	−0.03
propyl isobutyl	2.82	3.02	−0.20
diisopropyl	2.82	2.89	−0.07
ethyl pentyl	3.00	2.88	0.12
ethyl *tert*-butyl	3.15	2.88	−0.27

$$(0,0):\quad F = 93,\quad r = 0.955,\quad s = 0.13,\quad N = 21,\quad cbt = 0.70711$$

$$(10, -1):\quad F = 140,\quad r = 969,\quad s = 0.11,\quad N = 21,\quad cbt = 0.08704$$

$$(100, -1):\quad F = 176,\quad r = 975,\quad s = 0.10,\quad N = 21,\quad cbt = 0.00985$$

$$(1000, -1):\quad F = 178,\quad r = 0.976,\quad s = 0.10,\quad N = 21,\quad cbt = 0.0009985$$

The statistical improvement caused by the (x, y_O) values is evident. The computed values with $(x, y_O) = (1000, -1)$ are shown in Table 11, together with the corresponding residuals. The *cbt* values give an idea of the influence of x on the connectivity value: as x increases with $y_O = -1$, the influence of the carbon−carbon connectivity decreases relative to that of the carbon−oxygen connectivity, which becomes the preponderant connectivity describing the whole molecule.

Table 12. Boiling Points, bp/°C, of Cycloalkanes and Alkylcycloalkanes, Calculated Boiling Points, bp$_{clc}$, and Their Residuals, res, Calculated with x = −0.25 and y = −0.45

compd	bp	bp$_{clc}$	res	compd	bp	bp$_{clc}$	res
c3	−32.8	−26.8	−6.0	c4	12.6	12.0	0.6
M-c3	0.7	−0.2	0.9	c5	49.3	48.3	1.0
1,1MM-c3	20.6	20.2	0.4	1,2MM-c3	32.6	27.1	5.4
E-c3	35.9	35.7	0.2	M-c4	36.3	37.0	−0.7
c6	80.7	82.3	−1.6	1,1,2MMM-c3	54.0	47.5	6.5
E-c4	70.8	70.6	0.2	M-c5	71.8	71.7	0.1
c7	118.4	114.0	4.4	1,1MM-c5	87.5	89.6	−2.1
1,2MM-c5	95.7	95.6	0.1	E-c5	103.5	103.1	0.4
M-c6	100.9	104.1	−3.2	C8	149.0	143.3	5.7
1,1,2MMM-c5	114.0	113.3	0.7	1,1,3MMM-c5	104.9	110.9	−6.0
P-c5	131.0	129.8	1.2	iP-c5	126.5	123.1	3.4
1,1MM-c6	119.6	120.7	−1.1	1,2MM-c6	126.7	126.2	0.5
1,3MM-c6	122.3	124.8	−2.5	1,4MM-c6	121.9	124.8	−2.9
E-c6	131.9	133.2	−1.3	M-c7	134.0	134.2	−0.2
B-c5	156.6	154.8	1.8	1,2MP-c5	149.5	150.5	−1.0
1,1,3MMM-c6	136.6	140.5	−3.9	1,2,4MMM-c6	144.8	145.8	−1.0
1,3,5MMM-c6	139.5	144.4	−4.9	1,4ME-c6	150.8	152.3	−1.5
P-c6	156.7	157.9	−1.2	iP-c6	154.8	151.7	3.1
1,4MiP-c6	171.3	169.7	1.6	B-c6	180.9	180.7	−0.2
sB-c6	179.3	177.4	1.9	iB-c6	171.3	174.2	−2.9
tB-c6	171.5	165.9	5.6	1,2EE-c5	150.6	152.7	−2.1

The cycloalkane and alkylcycloalkane boiling points, bp, taken from ref 68, together with the corresponding computed values, bp$_{clc}$ and their residuals are collected in Table 12. To look into the role of the carbon atoms of the cyclic part and the carbon atoms of the acyclic part of an alkylcycloalkane molecule, the variable parameter x (see eq 24) was associated with the cyclic atoms and the variable parameter $y_H = y$ was associated with the acyclic atoms. The calculated boiling point values have been obtained with the pair of values $(x, y) = (-0.25, -0.45)$. The model of these boiling points is based on a quadratic equation in $^1\chi^f$, i.e., $bp = c_1{}^1\chi^f + c_2({}^1\chi^f)^2 + c_0$, which gives rise to the following two sets of statistics with $(x,y) = 0,0$ and $(x,y) = (-0.25, -0.45)$ as indicated.

$$(0, 0):\quad F = 3480,\quad r = 0.9972,\quad s = 4.1,\quad N = 42;\quad (-0.25, -0.45):\quad F = 6372,\quad r = 0.9985,\quad s = 3.0,\quad N = 42$$

These and the statistics for the previous property make obvious the improvement of the model achieved by using ${}^1\chi^f = f(x)$.

3. From Graph Invariants to Lead Design

3.1. Mathematical Tools

Throughout this section, somewhat different molecular connectivity indices will be introduced. In order to outline the particular QSPR techniques used with this methodology, descriptors will be defined before explaining the modeling tools applied with them. Diverse statistical and molecular techniques will be sketched here.

3.1.1. Descriptors

The following types of indices, which have been mainly used in this research, are described in increasing order of complexity.

3.1.1.1. Discrete Invariants. These are natural numbers calculated from what chemists understand qualitatively as the chemical structure. *N* is the number of non-hydrogen atoms, i.e., the number of molecular graph vertices.[89−91] V_k, where *k* is 3 or 4, is the number of vertices of degree *k*, i.e., the number of atoms having *k* bonds, σ or π, to non-hydrogen atoms.[91] PR_k for *k* between 0 and 3 is the number of pairs of ramifications at distance *k*, i.e., the number of pairs of single branches at distance *k* in terms of bonds.[91] *L* is the length, i.e., the maximum distance between non-hydrogen atoms measured in bonds, and is thus the diameter of the molecular graph defined as $\max(d_{ij})$.[91] *W* is the Wiener number, i.e., the sum of the distances between any two non-hydrogen atoms measured in bonds.[92−95]

3.1.1.2. Connectivity Indices. Throughout the present section, the connectivity indices defined as in eq 25 will be used.[13,21] Some of them are slightly different from the previously defined indices. The connectivity index of order k[21] may be derived from the adjacency matrix and is normally written as ${}^k\chi_t$. The order *k* is between 0 and 4 and is the number of connected non-hydrogen atoms which appear in a given substructure.

$$ {}^k\chi_t = \sum_{j=1}^{{}^kn_t} \left(\prod_{i \in S_j} \delta_i \right)^{-1/2} \tag{25a}$$

In eq 25a, δ_i is the number of simple (i.e., σ) bonds of the atom *i* to non-hydrogen atoms, S_j represents the *j*th substructure of order *k* and type *t*, and ${}^kn_t$ is the total number of subgraphs of order *k* and type *t* that can be identified in the molecular structure. The types used are path (p), cluster (c), and path-cluster (pc). Following the concepts defined in the Introduction, section A, a subgraph of type p is formed by a path, a subgraph of type c is formed by a star, while the pc subgraph can be defined as every tree which is neither a path nor a star. Alternatively, a pc subgraph is any tree containing at least a star and a path.

As an example, Table 13 displays all the p, c, and pc subgraphs found in a simple molecular structure.

The use of the valence delta, δ^ν, instead of δ enables the encoding of π and lone-pair electrons[13] in the form given in eq 25b.

$$ {}^k\chi_t^{\ \nu} = \sum_{j=1}^{{}^kn_t} \left(\prod_{i \in S_j} \delta_i^{\ \nu} \right)^{-1/2} \tag{25b}$$

Table 13. Types of Subgraphs Present in the 2-Methylpropanol Structure

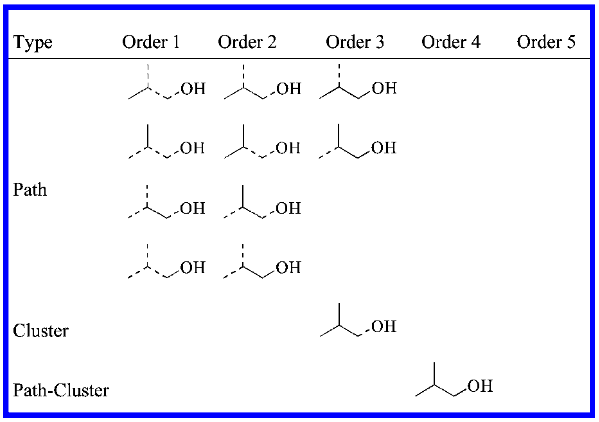


Table 14. Values of δ^ν for the Different Heteroatoms Present in the Listed Groups

group	δ^ν	group	δ^ν
NH_4^+	1	H_3O^+	3
NH_3	2	H_2O	4
$-NH_2$	3	$-OH$	5
$-NH-$	4	$-O-$	6
$=NH$	4	$=O$	6
$-N-$	5	O (nitro)	6
$=N-$	5	O (carboxyl)	6
$=N^+=$ (azide)	4	$-F$	7
$=N^-$ (azide)	6	$-Cl$	0.690
$-N=$ (nitro)	6	$-Br$	0.254
$-S-$	1.33	$-I$	0.15
$=S$	0.99	$=P-$	0.560
S ($-SO_2-$)	2.67	P(5)	2.22

Here, δ^ν is just the degree of a vertex in a pseudograph (the $\delta^\nu(ps)$ of section 2), and in this context, the old definition, $\delta^\nu = Z^\nu/(Z - Z^\nu - 1)$, for the $\delta_i^{\ \nu}$ of higher row atoms holds.[13] The values listed in Table 14 and used in this approach were adopted because of their general performance.

3.1.1.3. Topological Charge Indices (TCI). The topological charge indices G_k and J_k of order $k = 1-5$ are defined for a given graph by eq 26,[90] in which *N* is the number of non-hydrogen atoms and $c_{ij} = m_{ij} - m_{ji}$ is the charge term between vertices *i* and *j*. δ represents here the *Krönecker* delta symbol, i.e., if $\alpha = b$, then $d(a,b) = 1$, and if $\alpha \neq b$ then $d(a,b) = 0$, and finally, d_{ij} is the topological distance between vertices *i* and *j*.

$$ G_k = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |c_{ji}| \delta(k,d_{ij}), \quad \text{and} \quad J_k = \frac{G_k}{N-1} \tag{26}$$

The variables m_{ij} are the elements of the $N \times N$ matrix $\mathbf{M}$ obtained as the product of two matrices, i.e., $\mathbf{M} = \mathbf{A}\cdot\mathbf{Q}$. The elements of $\mathbf{M}$ expanded in terms of the elements of $\mathbf{A}$ and $\mathbf{Q}$ are given in eq 27.

$$ m_{ij} = \sum_{h=1}^{N} a_{ih} q_{hj} \tag{27}$$

$\mathbf{A}$ is the *adjacency* matrix in which elements a_{ih} are 0 if either $i = h$ or *i* is not linked to *h*; 1 if *i* is linked to *h* by a single bond; 1.5 if *i* is linked to *h* by an aromatic bond; 2 if *i* is linked to *h* by a double bond; and 3 if *i* is linked to *h* by a triple bond. $\mathbf{Q}$ is the inverse squared distance or Coulombian matrix. Its elements, q_{hj}, are 0 if $h = j$ and otherwise $q_{hj} = 1/d_{hj}^2$, where d_{hj} is the topological distance between vertices h and j.

Thus, G_k represents the overall sum of the c_{ij} charge terms for every pair of vertices i and j separated by a topological distance k. The valence topological charge indices G_k^v and J_k^v are defined in a similar way, but using A^v, the electronegativity-modified adjacency matrix, instead of A. The elements of A and A^v are identical except for the main diagonal, where A has zeroes and A^v the corresponding Pauling electronegativity values, EN, weighed for $EN(\mathrm{Cl}) = 2$ for each heteroatom.

To illustrate the calculation of the topological charge indices, let us consider *n*-butane. Its hydrogen-depleted graph is: •−•−•−•. If we number each vertex of this graph in the following way, 1−2−3−4, we can write the **A**, **Q**, and **M** matrices of eq 28, which are used to derive the following G values: $G_1 = |c_{12}| + |c_{23}| + |c_{34}| = {}^1/_4 + 0 + {}^1/_4 = 0.500$, $G_2 = |c_{13}| + |c_{24}| = {}^1/_9 + {}^1/_9 = 0.2222$, and $G_3 = |c_{14}| = 0$.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \qquad \boldsymbol{Q} = \begin{pmatrix} 0 & 1 & {}^1/_4 & {}^1/_9 \\ 1 & 0 & 1 & {}^1/_4 \\ {}^1/_4 & 1 & 0 & 1 \\ {}^1/_9 & {}^1/_4 & 1 & 0 \end{pmatrix}$$

$$\boldsymbol{M} = \boldsymbol{A}\cdot\boldsymbol{Q} = \begin{pmatrix} 1 & 0 & 1 & {}^1/_4 \\ {}^1/_4 & 2 & {}^1/_4 & {}^{10}/_9 \\ {}^{10}/_9 & {}^1/_4 & 2 & {}^1/_4 \\ {}^1/_4 & 1 & 0 & 1 \end{pmatrix} \quad (28)$$

3.1.1.4. Differences and Quotients of Connectivity Indices. The difference of connectivity indices, ${}^kD_t$, with $k = 0-4$ and $t = p, c, pc$, is defined in the following way,[96]

$${}^kD_t = {}^k\chi_t - {}^k\chi_t^{\,\nu} \quad (29)$$

The quotient of connectivity indices, ${}^kC_t$, with $k = 0-4$ and $t = p, c, pc$, is defined in the following way,[91]

$${}^kC_t = \frac{{}^k\chi_t}{{}^k\chi_t^{\,\nu}} \quad (30)$$

3.1.1.5. Electrotopological Indices. We have already seen the electrotopological state indices of eqs 16, which include electronic structural information for each atom of the graph as well as information about the topological environment of each atom.[57] Electrotopological state indices for each type of atom present in the molecule are obtained by summing the electrotopological states for the atoms of a given type. The symbol consists of "S" for "sum" followed by symbols for the bonds in the group (s = single, d = double, t = triple, a = aromatic) and symbols for the elements in the group. Thus, the hydroxyl group is SsOH, the ether oxygen is SssO, the keto oxygen is SdO, and the atom groups $-CH_3$, $=CH_2$, $\equiv CH$, $-CH_2-$, and $>C<$ are $SsCH_3$, $SdCH_2$, StCH, $SssCH_2$, and SssssC, respectively.

3.1.1.6. Shape Indices. The κ shape indices are the basis of a method of molecular structure quantitation in which attributes of molecular shape are encoded into three indices (κ values).[97] These κ values are derived from counts of one-bond, two-bond, and three-bond fragments, with each count being made relative to fragment counts in reference structures which possess a maximum and minimum value for that number of atoms. Furthermore, another shape index, namely, the shape index E, is defined as[91]

$$E = \frac{\sum_i n_i(d_i + 1)}{L} \quad (31)$$

In this formula, n_i is the number of vertices at a topological distance d_i from the "main path", where the latter is the shortest path joining the two most separated vertices. L is the diameter of the molecular graph. It is clear that the lower the E value, the more elongated the graph. Indeed, if the graph is considered as an ellipse, then E would correspond to the eccentricity. The discrete invariants, connectivity indices, TCIs, connectivity difference, and quotient descriptors can be calculated by using DESMOL11.[98] Electrotopological state and molecular shape indices can be calculated by using MOLCONN-Z.

Other indices have been used, such as, for example, versions of connectivity indices which include stereochemistry[100] or differences or quotients of connectivity or charge indices.[101]

3.1.2. Modeling

Modeling tools are mainly used to find new active compounds. To do this, two kinds of equations that are able to predict properties were obtained: the ones predicting quantitative properties (multilinear regression equations) and the others enabling recognition of the category to which the compound belongs (discriminant equations). A model is composed of several such equations of each type and the associated thresholds which define the intervals within which the active compounds generally fall for each equation. Thus, the model acts as a filter for potentially active new compounds by selecting only those which satisfy the equations within defined thresholds. Laboratory assays afford information on the efficiency of the model, which is useful in refining it.

3.1.3. Multilinear Regression Analysis (MLRA)

Several multilinear descriptive functions have been obtained by the linear correlation of physical properties with the aforementioned descriptors. The Furnival−Wilson algorithm[102] is used to obtain subsets of descriptors and equations with the least Mallows parameter, C_p,[103] as defined in eq 32.

$$C_p = \frac{\mathrm{RSS}}{s^2} + 2p - N \quad (32)$$

In eq 32, RSS is the residual sum of squares based on the selected independent variables, s^2 is the residual mean square based on the regression using all independent variables, p is one plus the number of subset independent variables, and N is the number of cases. This algorithm combines two methods of computing the residual sums of squares for all possible regressions to form a simple leap and bound technique for finding the best subsets without examining all possible ones. The result is a reduction by several orders of magnitude in the number of operations required to find the best subsets.

3.1.4. Validation of the Selected Equations

3.1.4.1. Stability. The stability of the selected mathematical models can be evaluated through a cross-validation by

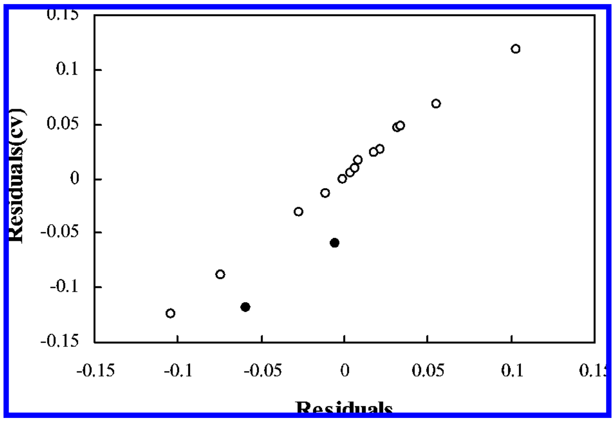


Figure 13. Validation of the equation obtained for LD_{50} of a set of antihistaminics. Residuals were obtained from the best regression versus residuals obtained by cross-validation (black points = outliers).

the leave-one-out method,[104] in which one compound of the set is extracted, the model is recalculated using as a training set the remaining $N - 1$ compounds, and the property is then predicted for the removed element. This process is repeated for all compounds of the set to obtain a prediction for each one. SEE(CV), the standard error of estimates for the cross-validation, is determined from the residuals for these predictions. The results can be plotted as residuals obtained from the best regression versus residuals obtained by cross-validation (see Figure 13).[105] This procedure also aids in the detection of outlying points. A more robust stability validation method is the leave-*n*-out method.[106]

The predictive ability of the equations obtained can be better assessed by an external test. A random subset of molecules is initially chosen (the "external test set"), and the modeling study is carried out with the remaining molecules (the "training set"). The predictive performance of the model is assessed by the results obtained when it is applied to the external test set.

3.1.4.2. Randomness. To detect the possible existence of fortuitous correlations, a randomization test can be done by randomly permuting the values of the property of each compound and then forming linear correlations with the aforementioned descriptors.[107] This process is repeated as many times as there are compounds in the set. The usual way to represent the results of a randomization test is to plot correlation coefficients versus predicted ones, r^2 and r^2_{cv}, respectively (Figure 14). r^2_{cv} is the prediction coefficient, defined as $r^2_{cv} = 1 - \text{PRESS/SD}$, with PRESS being the predictive residual sum of squares and SD being the squared deviation of the observed value from the mean observed value. It must be noted that, in contrast to r^2, the coefficient of prediction can take values less than zero.[38,108]

3.1.5. Linear Discriminant Analysis (LDA)

The objective of the linear discriminant analysis, LDA, is to find a linear combination of variables that allows discrimination between two or more categories or objects. Generally, two sets of compounds are considered in the analysis: first, a set of compounds with proven pharmacological activity and, second, a set of compounds known to be inactive. The selection of the descriptors is based on the Fisher−Snedecor parameter, and the classification criterion is the shortest *Mahalanobis* distance (i.e., the distance of each case from the mean of all cases used in the regression equation). Variables used in computing the linear classification functions are chosen stepwise. At each step, either the variable that adds the most to the separation of the groups is entered into the discriminant function or the variable that adds the least to the separation of the groups is removed from the discriminant function. The quality of the discriminant function is evaluated by Wilks' λ, which is a multivariate analysis of the variance parameter that tests the equality of group means for the variable(s) in the discriminant function. Minimization of Wilks' parameter allows selecting the predictors to be entered or deleted in the discriminant function.[16,107,108]

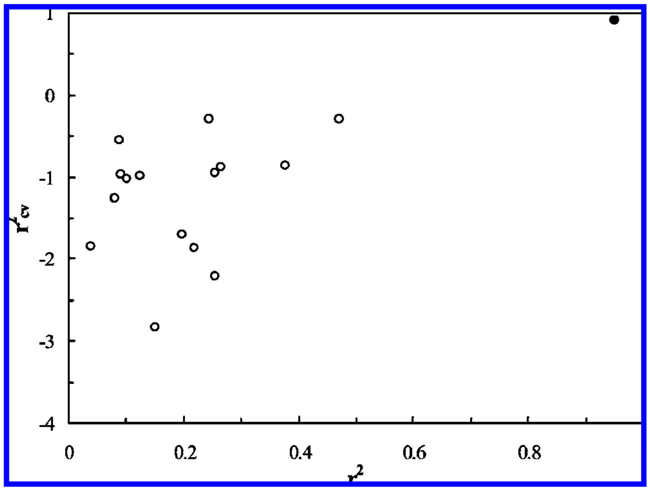


Figure 14. Validation of the mathematical model obtained for the property LD_{50}. The prediction coefficient, r^2_{cv}, versus correlation coefficient, r^2, was obtained by a randomization study.

The discriminant ability of the selected function is stated by the following:

(a) The *Classification matrix*, in which each case is classified into a group according to the classification function. The number of cases classified into each group and the percent of correct classifications are shown.

(b) The *Jack-knifed classification matrix,* in which each case is classified into a group according to the classification functions computed from all the data except the case being classified.[109]

(c) *Cross validation with random subsamples and classifying new cases.* Here, the cases in each group are randomly subdivided into two separate sets, the first of which is then used to estimate the classification function, and the second of which is classified according to the function. By observing the proportion of correct classifications produced for the second set, one obtains an empirical measure for the success of the discrimination.

(d) *Use of an External set test*, which entails the use of an external compound set to check the validity of the selected discriminant functions.

3.1.6. Pharmacological-Activity Distribution Diagrams (PDDs)

The pharmacological distribution diagram (PDD) is a useful tool for the selection of equations for the molecular design step.[110] PDDs are histogram-like plots in which the compounds, preferably from a test set, are grouped into intervals of the predicted value of the property *P*. The number of compounds in each interval of *P* is determined for each group. The expectancy *E* of finding a molecule with a desired value of *P* is obtained. For each arbitrary interval of whatever

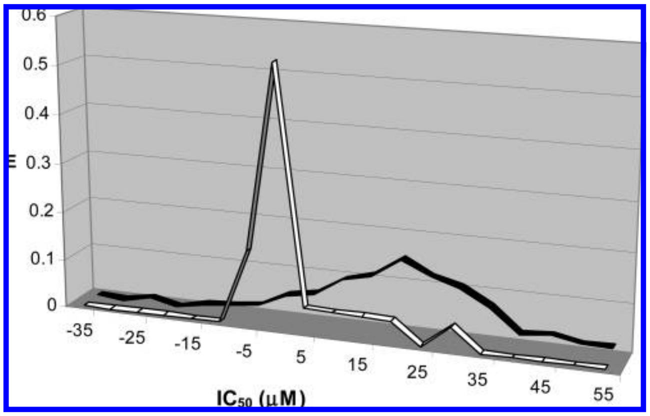


Figure 15. PDD for IC_{50} in HSV-1: $IC_{50} = -17.36({}^4\chi_p) + 41.39({}^4\chi_{pc}{}^v) + 21.71$. Abscissas: $IC_{50}/\mu M$. Ordinates: Expectancy of activity in white; expectancy of nonactivity in black. Interval: between −10 and 20.

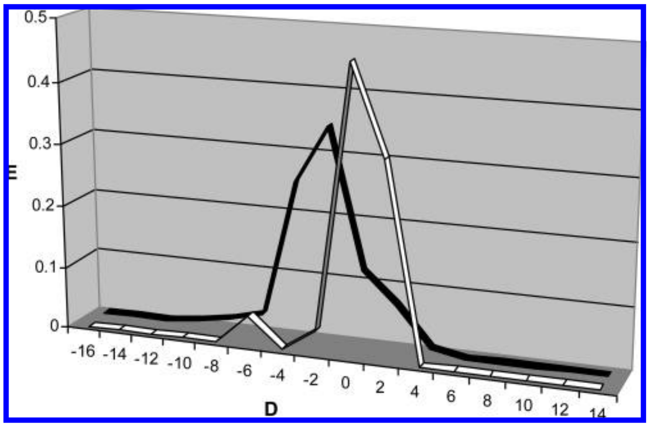


Figure 16. PDD for discriminant function of antiviral activity: $D = -1.17({}^0\chi^v) + 2.11({}^3\chi_p) + 2.79$. Abscissas: Classification function obtained by linear discriminant analysis. Ordinates: Expectancy of activity in white; expectancy of nonactivity in black. Interval: between −1 and 5.

function, the expectancy of the activity is defined as $E_a = a/(i + 1)$, where a is the ratio of the number of active compounds in this interval and the total number of active compounds, and i represents the ratio of inactive compounds in the same way. The expectancy of inactivity is similarly defined as $E_i = i/(a + 1)$. For a given equation, it is straightforward to see the zones in which the overlap of E_a and E_i is smallest and thereby decide if the equation being studied should be selected or not as useful in molecular design. This also permits us to determine the intervals of the property where the probability of finding new active compounds is greatest relative to the choice of a false active. Figures 15 and 16 show the PDDs obtained with a connectivity function and a discriminating function, respectively.[111]

3.2. Molecular Selection and Design

Several schemes of molecular design used in this approach will be discussed and illustrated with selected examples.

3.2.1. Database Search with External Validation

A mathematical model consisting of one or more equations with their corresponding thresholds is used to filter a structural database, and the selected structures are checked for the activity of interest in the database bibliography.

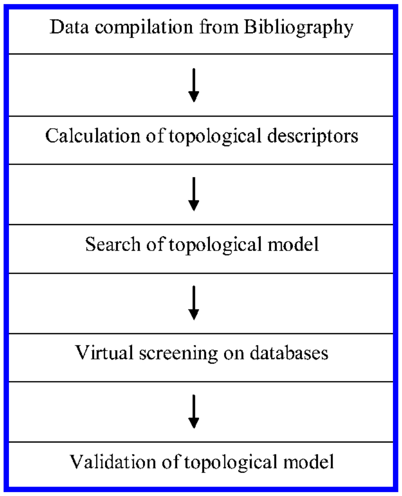

Data compilation from Bibliography
Calculation of topological descriptors
Search of topological model
Virtual screening on databases
Validation of topological model

Figure 17. Scheme of molecular design through database search with external validation.

Compounds found to be active validate the model, while those which are not listed as having been tested for the activity of interest are proposed for assay as new potentially active compounds. A possible scheme is shown in Figure 17.

Example: Anticonvulsant Activity, AC.[112] The model obtained from linear discriminant analysis was

$$\text{tam} = -28.88 - 1.94\,{}^4\chi_{pc}{}^{v} - 0.21G_1{}^{v} + 4.64G_5 + 20.11J_3{}^{v} - 45.87J_4 - 3.42\,{}^0D + 40.65({}^0C) - 10.47({}^3C_p) + 2.79({}^4D_p) + 1.32(PR_0)$$

$$F = 10,\quad \lambda = 0.54,\quad N = 128,\quad \text{Selection criterion: active if tam} > 0$$

This model has been used to predict the anticonvulsant activity (AC) of compounds not considered in its derivation. The 10330 compounds included in the Merck index were screened for this purpose. Using this model, 108 different compounds were selected as potential anticonvulsant drugs. A literature search for data on the pharmacological profile of these 108 compounds was done. It was found that in 41 cases the AC of the structures had been previously reported[112] (Table 15). This result shows the accuracy of the model, which found new potential leads from structures known to show a different pharmacological profile.

3.2.2. Molecular Selection by Virtual Screening on Databases

A mathematical model consisting of one or more equations with their corresponding thresholds is used to filter a structural database, and the structures selected are tested *in vitro* for the activity of interest. Compounds selected as actives but not showing activity *in vitro*, i.e., false positives, as well as the activities found for true positives are used to refine the model. A possible scheme is shown in Figure 18.

Example: Bronchodilator Activity.[113a] The mathematical model used to discover new compounds with bronchodilator activity is comprised of two discriminant functions: the first one, namely DF_1, was constructed using a large set of bronchodilator drugs consisting of more than 300 compounds including xanthines, β-adrenergic agonists, anti-cholinergics,

Table 15. Results of Prediction of Anticonvulsant Activity Obtained When Applying the *tam* Proposed to *The Merck Index* Base

compd	therapeutic category[a]	prob[b]
aconitine	neuralgia	1.000
adinazolan	antidepressant	0.931
α-methylenebutyrolactone		0.943
amido-G-acid		0.990
amlodipine	antianginal, antihypertensive	0.956
cannabidiol		0.978
cannabinol		0.907
caprolactam		0.982
cycloheptanone		0.976
cycloleucine		0.995
cyclopentanone		0.991
cychlothiazide	diuretic, antihypertensive	0.960
diethadione	analeptic	0.978
ectylurea	sedative, hypnotic	0.933
felodipine	antihypertensive, antianginal	0.939
guvacine		0.968
linoleic acid	nutrient	0.920
L-pyroglutamic acid		0.977
nifedipine	antianginal, antihypertensive	0.961
flunarizine	vasodilator	0.815
diltiazem	antianginal, antiarrhythmic	0.540
nicardipine	antianginal, antihypertensive	0.610
nisoldipine	antianginal, antihypertensive	0.974
nitrendipine	antihypertensive	0.600
nimodipine	vasodilator cerebral	0.584
verapamil	antianginal, antihypertensive	0.741
prenylamine	vasodilator coronary	0.862
nipecotic acid		0.984
phencyclidine	analgesic, anesthetic	0.911
phthalimide		0.917
pipecolic acid		0.961
proline		0.932
riluzole	neuroprotective	0.916
biperiden	anticholinergic, antiparkinsonian	0.719
scopolamine	anticholinergic	0.760
trihexyphenidyl	anticholinergic, antiparkinsonian	0.871
benactyzine	antidepressant, anticholinergic	0.643
benztropine	anticholinergic	0.685
sulfanilamide	antibacterial	0.921
caramiphen	anticholinergic, antitussive	0.690
carbetapentane	antitussive	0.571

[a] From *The Merck Index*, 12th ed. [b] Probability of anticonvulsant activity by the *tam* function.

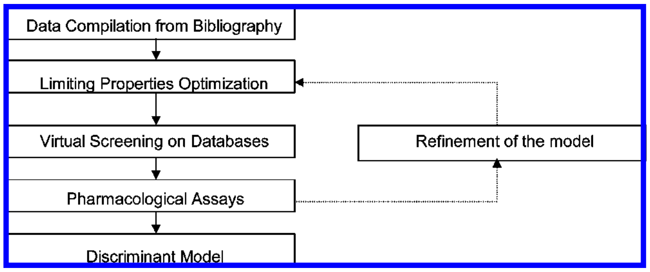

Data Compilation from Bibliography
Limiting Properties Optimization
Virtual Screening on Databases
Refinement of the model
Pharmacological Assays
Discriminant Model

Figure 18. Scheme of molecular design through molecular selection by virtual screening on databases.

and leukotriene antagonists as well as additional structurally heterogeneous drugs showing some extent of bronchodillator activity. A second discriminant function, DF_2, was also introduced solely to improve the DF_1 discriminant efficiency. This DF_2 function was obtained from a much smaller set of about 70 bronchodilator drugs which however did include as many representative compounds as possible in order to consider drugs belonging to every family of bronchodilators.

The discriminant functions chosen were as follows:[113b]

$$DF_1 = 3.07(^1\chi^v) - 3.58G_1 + 15.32J_2 + 55.50J_4 - 1.68PR_1 + 0.879PR_2 - 11.71$$

$$F = 287, \quad \lambda = 0.271, \quad N = 739$$

$$DF_2 = 17.40(^3D_p) - 12.27(^4D_p) - 6.61: \quad F = 129, \quad \lambda = 0.315, \quad N = 192$$

Based on this model, a compound is classified as inactive as a potential bronchodilator unless either $-1 < DF_1 < 10$ or $0 < DF_2 < 17$. Most of the active compounds do fall within these ranges. Application of the DF_2 and DF_1 functions to a collection of structures related variously to coumarines, flavonoids, and anthocyanosides drawn from our working databases resulted in the selection of 20 theoretically active compounds. Table 16 shows the structure of each compound and the values of DF_1 and DF_2 as well as the classification for each candidate for the different structures I−IV. From the mean relaxation values, E_{max}, the concentration−response curve was made (E_{max}(%) vs $-\log C$). The effective concentration 50% (EC_{50}) was calculated by interpolation, and it was expressed as pD_2 ($-\log EC_{50}$). The result obtained for every compound is illustrated in Table 17.

3.2.3. Virtual Combinatorial Syntheses and Computational Screening

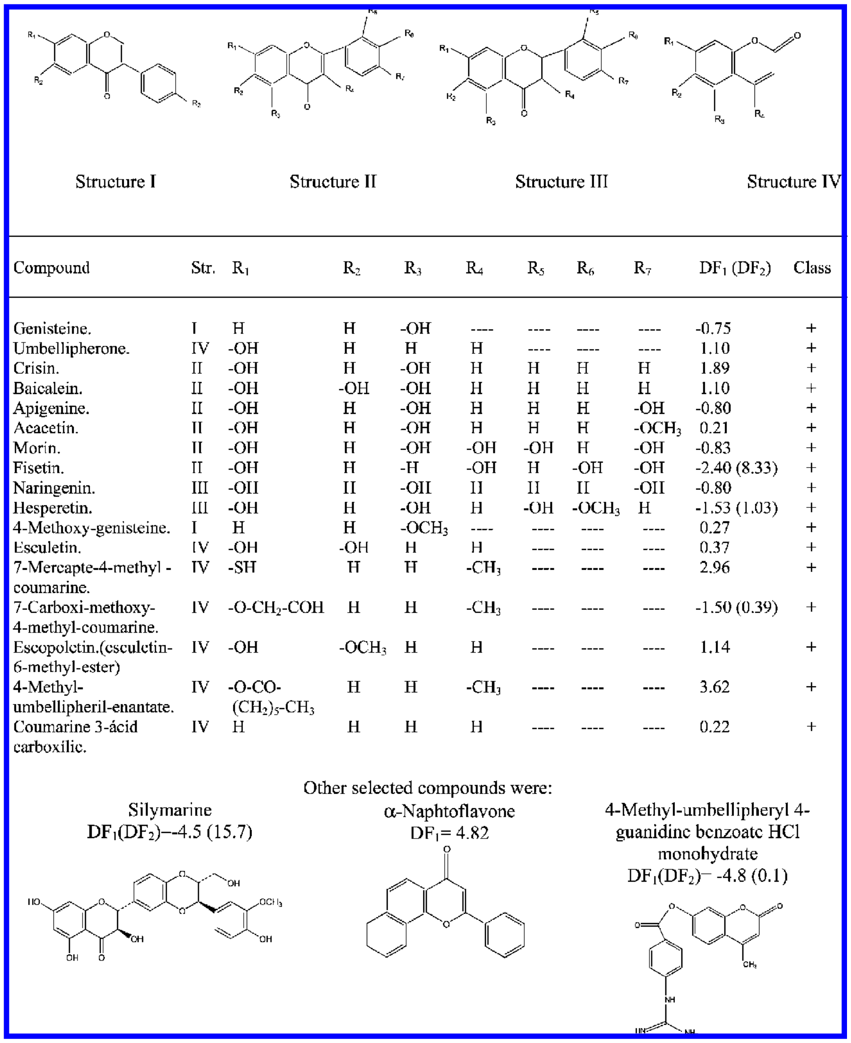
A mathematical model consisting of one or more equations with their corresponding thresholds is applied to a virtual combinatorial library made up of molecular structures resulting from a given synthetic scheme, and the structures selected from the virtual library by the model are then actually synthesized and tested. A possible scheme is displayed in Figure 19.

Example: Anti-herpes Activity.[111] The activity of new anti-herpes simplex virus type 1, anti-HSV-1, designed by virtual combinatorial chemical synthesis and selected by a computational screening, is determined. A virtual library of phenolesters and anilides was formed from two databases of building blocks, one with carbonyl fragments, and another containing both substituted phenoxy and phenylamino fragments. The virtually assembled compounds library was computationally screened, and those compounds that our mathematical model selected as active were finally synthesized and tested. The compounds shown in Figures 20 and 21 typify the compositions of the anhydride and the nucleophile databases, respectively. These compounds were chosen from among the commercially available ones, with care being taken to avoid substituents which would lead to side products in the real synthesis. The final compounds selected, shown in Figure 22, were synthesized and tested, and the results of the pharmacological assays are collected in Table 18.

3.2.4. Molecular Design of New Structures

In 1985 and 1988 the Valencia group completed two doctoral theses dealing with the prediction of drug properties and drug design using topological indices.[114] The results were published in a follow-up paper.[115] The main idea therein was the use of topological indices in an inverse way as compared to the usual: i.e. obtaining molecular structures from topological indices fulfilling predetermined properties.[91] This endeavor was supported by the fact that topological indices

Table 16. Structures of the 20 (17 + the 3 Figures at the Bottom) Compounds Selected by Molecular Topology Together with the Values of DF_1 and DF_2 When Relevant



Structure I Structure II Structure III Structure IV

Compound	Str.	R_1	R_2	R_3	R_4	R_5	R_6	R_7	DF_1 (DF_2)	Class
Genisteine.	I	H	H	-OH	----	----	----	----	-0.75	+
Umbellipherone.	IV	-OH	H	H	H	----	----	----	1.10	+
Crisin.	II	-OH	H	-OH	H	H	H	H	1.89	+
Baicalein.	II	-OH	-OH	-OH	H	H	H	H	1.10	+
Apigenine.	II	-OH	H	-OH	H	H	H	-OH	-0.80	+
Acacetin.	II	-OH	H	-OH	H	H	H	$-OCH_3$	0.21	+
Morin.	II	-OH	H	-OH	-OH	-OH	H	-OH	-0.83	+
Fisetin.	II	-OH	H	-H	-OH	H	-OH	-OH	-2.40 (8.33)	+
Naringenin.	III	-OII	II	-OII	II	II	II	-OII	-0.80	+
Hesperetin.	III	-OH	H	-OH	H	-OH	$-OCH_3$	H	-1.53 (1.03)	+
4-Methoxy-genisteine.	I	H	H	$-OCH_3$	----	----	----	----	0.27	+
Esculetin.	IV	-OH	-OH	H	H	----	----	----	0.37	+
7-Mercapte-4-methyl - coumarine.	IV	-SH	H	H	$-CH_3$	----	----	----	2.96	+
7-Carboxi-methoxy-4-methyl-coumarine.	IV	$-O-CH_2-COH$	H	H	$-CH_3$	----	----	----	-1.50 (0.39)	+
Escopolctin.(csculctin-6-methyl-ester)	IV	-OH	$-OCH_3$	H	H	----	----	----	1.14	+
4-Methyl-umbellipheril-enantate.	IV	-O-CO-$(CH_2)_5-CH_3$	H	H	$-CH_3$	----	----	----	3.62	+
Coumarine 3-ácid carboxilic.	IV	H	H	H	H	----	----	----	0.22	+

Other selected compounds were:

Silymarine
$DF_1(DF_2)$--4.5 (15.7)

α-Naphtoflavone
DF_1= 4.82

4-Methyl-umbellipheryl 4-guanidinc bcnzoatc HCl monohydrate
$DF_1(DF_2)$~ -4.8 (0.1)

Table 17. Values Obtained for Percentage of Relaxation and pD_2 ($pD_2 = -\log EC_{50}$) for Selected Compounds

compd	relaxation (%) (0.1 mM)	pD_2 ($-\log EC_{50}$)	no. of tests
theophylline (reference drug)	77.0 ± 0.0	4.69	6
3-coumarine carboxílic acid	0.0 ± 0.0		4
genisteine	73.7 ± 6.1	4.60	8
naringenin	70.5 ± 10.1	4.60	8
4-methoxy-genisteine	57.5 ± 4.9	4.65	9
umbellipherone	75.4 ± 6.5	4.50	8
esculetin	74.7 ± 6.8	5.35	9
fisetin	88.9 ± 2.0	4.60	14
hesperetin	87.4 ± 4.0	4.75	8
chrysin	60.8 ± 5.6	4.70	13
baicalein	65.4 ± 4.6	4.50	10
apigenine	58.2 ± 2.4	4.60	7
7-carboximethoxy-4-methyl-coumarine	12.8 ± 3.2	6.90	5
sylimarine	48.2 ± 3.5	4.50	8
morin	58.2 ± 6.6	4.50	8
acacetin	−38.4 ± 4.9		4
4-methyl-umbelliferyl-enantate	46.9 ± 5.6	4.65	17
escopoletin	61.7 ± 2.1	4.80	8
α-naphtoflavone	41.6 ± 4.7	4.95	10
7-mercapte-4-methyl-coumarine	55.1 ± 4.8	4.55	15
4-methyl-umbelliferyl-4-guanidine benzoate HCl monohydrate	80.8 ± 1.5	4.45	9

were not determined as simple structure-related descriptors but rather as an *algebraic description of the structure itself.* The method allowed for molecular construction from scratch or, alternatively, used a given base structure as scaffolding for a set of substituents drawn from hydrocarbon substructures and other functional groups.

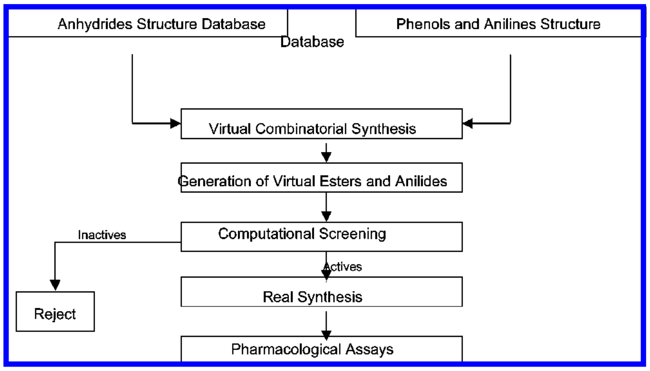

Anhydrides Structure Database
Phenols and Anilines Structure Database
Virtual Combinatorial Synthesis
Generation of Virtual Esters and Anilides
Computational Screening
Inactives
Actives
Reject
Real Synthesis
Pharmacological Assays

Figure 19. Scheme of molecular design through virtual combinatorial syntheses and computational screening.

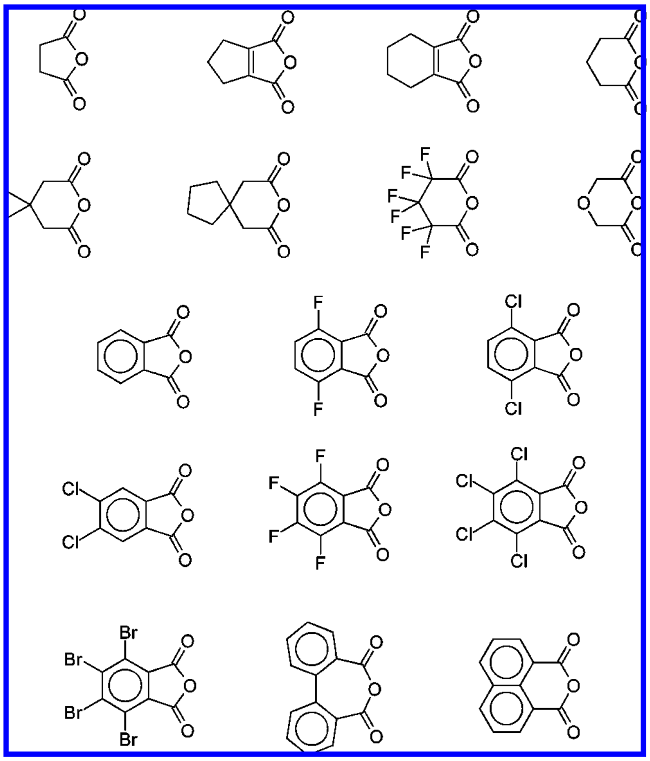


Figure 20. Symmetric cyclic anhydrides represented in the carbonyl fragments database.

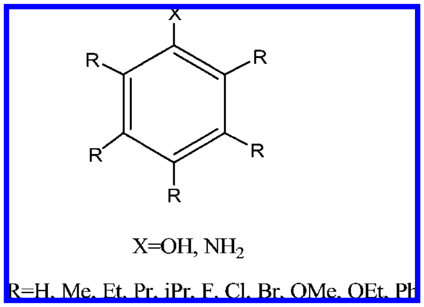


Figure 21. Phenols and anilines represented in the phenoxy and phenylamino fragments database.

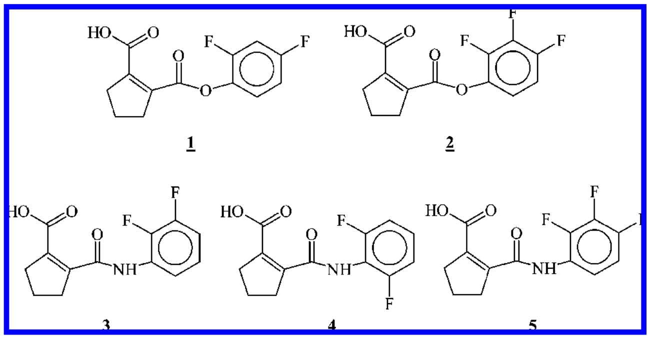


Figure 22. Compounds designed by virtual combinatorial chemical syntheses and selected by computational screening.

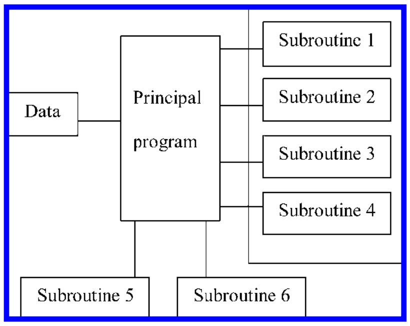


Figure 23. Scheme of molecular design of new structures (subroutines 1, 2, 3, and 4: formation of carbon−carbon bonds; subroutine 5: inclusion of functional groups; subroutine 6: molecular selection).

The substructural fragments were noncyclic and contained bond orders from one to three. The fragments and functional groups were computationally attached to the base structure either at single sites or as connections between previously defined attachment sites. These could be connected by each of their available atoms, and the formation of multiple bonds and cyclic structures was allowed in these steps. The computer program determined whether each new molecule designed was potentially active according to a predefined QSAR linear model based on Randić−Kier−Hall type indices.

The program also allowed the user to select the number of structures to be generated by implementing an adjustable screening level in the construction process which varied from a relentless exhaustive search of each isomer to wide leaps within the space of possible graphs. This procedure generated an adjustable molecular diversity. The usual strategy consisted of initial runs with wide intervals for the acceptable QSAR equation values and high diversity. In successive runs designed around the newly selected structures, the intervals were narrowed and the diversity was decreased. Finally, the compounds expected to be most active were synthesized and tested, and the test results were in turn used to refine the QSAR model. Figure 23 shows a scheme of the algorithm for molecular design of new structures.

Example: Non-narcotic Analgesics.[116] New analgesic drugs have been designed using molecular topology with linear discriminant analysis and connectivity functions using different topological descriptors.[117] The compounds 2-(1-propenyl)phenol, 2′,4′-dimethylacetophenone, *p*-chlorobenzohydrazide, l-(*p*-chlorophenyl)propanol, and 4-benzoyl-3-methyl-l-phenyl-2-pyrazolin-5-one were particularly notable, showing as they did analgesic values larger than 75% versus ASA (acetylsalicylic acid), the reference drug. The usefulness of this design method has been demonstrated in the search for new chemical structures having analgesic effects, some of which could become “lead drugs”.

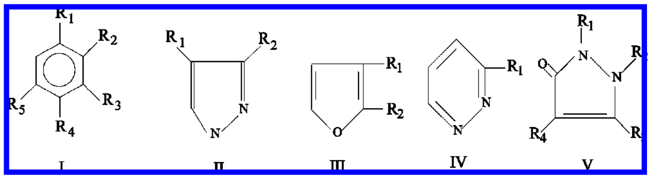
The different base structures used in this design step are collected in Table 19. The results of analgesia for each of the designed and selected compounds are shown in Table 20.

Table 18. Compounds Designed by Virtual Combinatorial Chemical Syntheses and Selected by Computational Screening, with Their Experimental IC_{50} on HSV-1, and Cytotoxicity Results

no.	compd	$IC_{50}/\mu M$	cytotoxicity[a]
1	2-(2,4-difluorophenoxycarbonyl)-1-cyclopentene-1-carboxylic acid	1.4	low
2	2-(2,3,4-trifluorophenoxycarbonyl)-1-cyclopentene-1-carboxylic acid	1.8	medium
3	2-(2,3-difluorophenylcarbamoyl)-1-cyclopentene-1-carboxylic acid	0.9	no
4	2-(2,6-difluorophenylcarbamoyl)-1-cyclopentene-1-carboxylic acid	0.9	no
5	2-(2,3,4-trifluorophenylcarbamoyl)-1-cyclopentene-1-carboxylic acid	0.9	no

[a] Effect on cell growth of noninfected cellular monolayers, at the corresponding IC_{50}.

Table 19. Base Structures Used in the Design Stage and Chemical Structures of the Compounds Selected as Theoretical New Analgesics[116]



compd	str	R_1	R_2	R_3	R_4	R_5
2-(1-propenyl)phenol	I	OH	$CHCHCH_3$	H	H	H
2′,4′-dimethylacetophenone	I	$COCH_3$	CH_3	H	CH_3	H
p-methylpropiophenone	I	COC_2H_5	H	H	CH_3	H
1-(*p*-chlorophenyl)propanol	I	Cl	H	H	$CHOHC_2H_5$	H
2-(1-hydroxy-3-butenyl)phenol	I	OH	$CHOHCH_2CHCH_2$	H	H	H
3-chlorosalicylic acid	I	COOH	OH	Cl	H	H
4-hydroxyantipyrine	V	C_6H_5	CH_3	CH_3	OH	
p-chloropropiophenone	I	Cl	H	H	COC_2H_5	H
5-chlorosalicylic acid	I	COOH	OH	H	H	Cl
3′,4′-dimethylacetophenone	I	$COCH_3$	H	CH_3	CH_3	H
1-(*p*-chlorophenyl)propylamine	I	Cl	H	H	$CHNH_2C_2H_5$	H
3-amino-4-carbethoxypyrazole	II	$COOC_2H_5$	NH_2			
3-methylpyridazine	IV	CH_3				
p-chlorobenzohydrazide	I	$CONHNH_2$	H	H	Cl	H
4,5-dichlorophthalic acid	I	Cl	Cl	H	COOH	COOH
2,4-dichlorobenzyl alcohol	I	Cl	H	Cl	CH_2OH	H
vanillic acid	I	COOH	H	OCH_3	OH	H
3-aminopyrazole	II	H	NH_2			
methyl 2-methyl-3-furancarboxylate	III	$COOCH_3$	CH_3			

Table 20. Results Obtained in the Study of Analgesic Activity for a Group of Designed and Selected Chemical Structures[116]

compd	analgesia (%)	ED_{50} (mg/kg)	LD_{50} (mg/kg)	TI (LD_{50}/ED_{50})
acetyl salicylic acid (ASA)[a]	49 ± 1^c	100 ± 8	$500 + 20$	5
2-(1-propenyl)phenol	85 ± 1^c	34 ± 5	720 ± 10	21
2′,4′-dimethylacetophenone	80 ± 1^c	45 ± 5	700 ± 10	16
p-methylpropiophenone	56 ± 1^c	100 ± 3	590 ± 20	6
1-(*p*-chlorophenyl)-propanol	78 ± 1^c	83 ± 2	720 ± 30	9
2-(1-hydroxy-3-butenyl)phenol	41 ± 4^a	114 ± 5	780 ± 20	7
3-chlorosalicylic acid	31 ± 1^c	106 ± 1	900 ± 30	8
3-amino-4-carbethoxy pyrazole	64 ± 2^b	87 ± 5	500 ± 30	6
3-methylpyridazine	45 ± 2^c	121 ± 11	>2000	>22
4-hydroxyantipyrine	14 ± 1^a			
p-chloropropiophenone	11 ± 1			
vanillic acid	28 ± 6			
5-chlorosalicylic acid	6 ± 2			
3′,4′-dimethylacetophenone	6 ± 3			
1-(*p*-chlorophenyl)propylamine				
p-chlorobenzohydrazide	80 ± 2^c	72 ± 10	282 ± 8	4
4,5-dichlorophthalic acid	23 ± 3			
2,4-dichlorobenzyl alcohol				
3-aminopyrazole				
methyl 2-methyl-3-furancarboxylate				

[a] ASA was taken as a reference point in the analgesic experiments: $^ap < 0.050$. $^bp < 0.010$. $^cp < 0.001$.

3.3. Applications

The most interesting results obtained by the methodology described in the preceding paragraphs have been grouped into the three categories shown in Tables 21−23. Each table shows the name of the property or activity modeled, the best equations obtained, and finally a reference to the corresponding article. The categories represented in Tables 21−23 are (a) the prediction of physicochemical parameters, (b) the

Table 21. QSAR Topological Models To Predict Physicochemical Properties

property	predictive equation	ref
viscosity	$\log \eta_{\text{exp}} = 0.664{}^3\chi_p - 0.263{}^4\chi_{pc} - 0.039\text{SsCH}_3 -$ $0.063\text{SssCH}_2 + 0.083\text{SOH} + 0.093\text{SsNH}_2 - 0.598$ $F = 175, r = 0.951, s = 0.17, s(\text{cv}) = 0.19, N = 117$	118
chromatographic properties R_{F}	$hR_{F2} = 107.6 + 10.6{}^4\chi_p{}^\nu - 4.9G_5{}^\nu - 14.7{}^3D_p - 25.9\text{PR}_0$ $F = 24, r = 0.916, s = 8.8, N = 23$ $hR_{F4} = 89.0 - 22.9{}^4\chi_p{}^\nu + 5.5G_5{}^\nu - 28.2{}^4D_{pc} - 3.2\text{PR}_3$ $F = 23, r = 0.955, s = 9.4, N = 23$	119
surface tension σ	$\sigma = 12.63 + 3.12\textit{Tipo}\text{Atom} - 12.11{}^0\chi + 8.99{}^3\chi_p{}^\nu + 2.03\kappa_0 +$ $2.93\text{Sum}I + 10.34\text{SHbint} - 3.22\text{SaOa} - 0.55\text{SF} + 4.40\text{Numhbd}$ $F = 32, r = 0.902, r^2(\text{cv}) = 0.73, s = 3.83, N = 77$	120
thermal conductivity λ	$\lambda = 0.124 - 0.003\text{Sum}\Delta I + 0.010\text{SsdCH} - 0.025\text{SsssCH} + 0.019\text{SssdC} +$ $0.011\text{SaNHa} - 0.004\text{SCl} - 0.004\text{SBr} + 0.0098\text{Numhbd} + 0.021\text{Numhba}$ $F = 22, r = 0.871, r^2(\text{cv}) = 0.700, s = 0.01, N = 74$	120
refractive indices (linear polymers) n_{R}	$n_R = 1.471 - 0.029{}^1\chi + 0.030{}^0\chi^\nu - 0.006\kappa_3 + 0.002H_{\text{max}} - 0.011\text{SsCH}_3 +$ $0.009\text{SaCHa} + 0.012\text{SssdC} - 0.006\text{SssO} - 0.003\text{SF} - 0.0001W$ $F = 170, r = 0.980r^2(\text{cv}) = 0.948, s = 0.01, N = 79$	121
glass transition temperatures, T_{g} (linear polymers)	$T_{\text{g}} = 7.533 - 3.394{}^1\chi^\nu + 1.284{}^2\chi^\nu + 1.148{}^3\chi_p{}^\nu + 0.295\kappa_{\alpha 1} -$ $0.174\text{SssO} - 0.038\text{SF} + 0.007W - 4.899J_1 + 2.252J_2 - 0.259V_4$ $F = 62, r = 0.945, r^2(\text{cv}) = 0.842, s = 0.44, N = 84$	121
chemiluminescent behavior[a] (pharmaceuticals and pesticides)	$\text{DF} = -0.20 - 87.98{}^7\chi_{\text{CH}} - 276.12{}^7\chi_{\text{CH}}{}^\nu + 1.224\text{SsdCH} - 35.38J_3{}^\nu + 66.81J_4{}^\nu$ $F = 20, l = 0.427, N = 96$	122
solubility (organic compounds)	$-\log S = 0.803 + 0.398{}^1\kappa + 0.234\text{SaaaC} - 1.195J_1{}^\nu - 1.536{}^0D$ $F = 141, r = 0.927, s = 0.90, N = 97$	123

[a] Chemiluminescent behavior if DF > 0.

Table 22. QSAR Topological Models To Predict Pharmacological Properties[a]

property	predictive equation	ref
t_{iw} (antihistaminic)	$t_{iw} = 67.19 - 32.94\textit{IShannon} - 0.72\text{Sum}I + 1.73\text{Sum}\Delta I$ $F = 52, r = 0.975, s = 0.77, s(\text{cv}) = 0.99, N = 12$	124
$\text{MIC}_{\textit{Epid.flocc.}}$ (antifungal)	$\log \text{MIC} = 0.097\text{SOH} - 0.243\textit{Phia}$ $F = 183, r = 0.961, s = 0.304, s(\text{cv}) = 0.324, n = 32$	125
carcinogenicity[b] discriminant function DF	$\text{DF}_{\text{carcino}} = 10.19 + 696.2{}^9\chi_{\text{CH}}{}^\nu - 45.88J_5{}^\nu - 10.16\text{Shannon} +$ $0.48\text{SsdN} - 0.07\text{SsOH} + 0.63\text{SaaaC} - 0.44\text{SssNH} - 0.44\text{SsssN} +$ $1.30\text{SdS} + 0.33n\text{class}$ $F = 27, \lambda = 0.43, N = 164$	126
IC_{50} (anti-*toxoplasma*) quinolones	$\log(1/\text{MIC}) = -6.1 + 0.3G_2 - 0.6G_3 - 9.3J_4 + 18.1J_4{}^\nu + 0.3\text{PR}_1$ $F = 24, r = 0.933, r^2(\text{cv}) = 0.74, s = 0.24, N = 24, C_{\text{p}} = 6.0$	127
IC_{50} (cyclooxigenase) analgesics	$\log IC_{50} = 0.32G_1{}^\nu + 6.34J_1 - 0.68V_4 + 1.4E - 2.25$ $F = 18, r = 0.908, s = 0.49, N = 20$	118
UDU (unchanged drug in urine) *anti-herpes*	$\log(\text{UDU}) = -4.67{}^1\chi^\nu + 8.70{}^2\chi - 3.64{}^3\chi_p + 3.15{}^3\chi_p{}^\nu - 8.05{}^3\chi_c - 9.23$ $F = 41, r = 0.957, s = 12.4, N = 25, C_{\text{p}} = 3.03$	111

[a] More details can be found in the cited references. [b] Carcinogenicity activity if $\text{DF}_{\text{carcino}} > 0$.

prediction of pharmacological properties, and (c) mathematical models for the selection and design of new active compounds, respectively.

Table 24 displays the results of a search using virtual screening and molecular design for compounds with previously unrecognized biological/pharmacological activities. This search was carried out on a group of compounds drawn from catalogs and from drugs of different therapeutic utility, and the activity discovered for each one was previously unknown.

3.3.1. Prediction of Physicochemical Parameters

The QSAR models obtained for the prediction of physicochemical properties are summarized in Table 21. Additional details are available in the references cited.

3.3.2. Prediction of Pharmacological Properties

The QSAR topological models used to predict pharmacological properties are given Table 22. Additional details are available in the references cited.

3.3.3. Mathematical Models for the Selection and Design of New Active Compounds

Mathematical−topological models for the selection and design of new active compounds are presented in Table 23. More details are available in the references cited.

3.3.4. New Biological Activities Discovered through Virtual Screening and Molecular Design

The new biological activities discovered through virtual screening are illustrated in Table 24. Details on assays and protocols can be found in the references therein.

4. Chemical Kinetics and Chemical Graph Theory

4.1. Kinetic Graphs

4.1.1. Kinetic Mechanisms

A short introduction will be *given to the* use of graph−theoretical concepts in chemical kinetics. The interested resder should refer to the bibliography for a deeper view of

Table 23. Mathematical-topological Models for the Selection and Design of New Active Compounds. Throughout the Last Column are the References.

therapeutic group	model	ref
anticonvulsant	$\mathrm{DF}_{\mathrm{tam}} = -28.88 - 1.94\,{}^4\chi_{pc}^{\nu} - 0.21G_1^{\nu} + 4.64G_5 + 20.11J_3^{\nu} - 45.87J_4 - 3.42\,{}^0D + 40.65\,{}^0C - 10.47\,{}^3C_p + 2.79\,{}^4D_p + 1.32PR_0$ $F = 10$, $\lambda = 0.54$, $N = 128$ selection criterion: active if $\mathrm{DF}_{\mathrm{tam}} > 0$	112
bronchodilator	$\mathrm{DF}_1 = 3.07\,{}^1\chi^{\nu} - 3.58G_1 + 15.32J_2 + 55.50J_4 - 1.68PR_1 + 0.879PR_2 - 11.71$ $F = 287$, $\lambda = 0.271$, $N = 739$ $\mathrm{DF}_2 = 17.40\,{}^3D_p - 12.27\,{}^4D_p - 6.61$ $F = 129$, $\lambda = 0.315$, $N = 192$ selection criterion: active if $\mathrm{DF}_1 > -1$ and $\mathrm{DF}_1 < 10$ or $\mathrm{DF}_2 > 0$ and $\mathrm{DF}_2 < 17$	113
antihistaminic	$\mathrm{DF}_1 = 7.20({}^1\chi_c^{\nu}) + 0.25G_1^{\nu} - 47.96J_1 - 22.98J_3^{\nu} - 4.89({}^4D_{pc}) - 0.36L + 12.65$ $F = 35$, $\lambda = 0.347$, $N = 146$ $\mathrm{DF}_2 = 2.13\mathrm{SdssC} + 1.37\mathrm{SaaCH} - 0.68\mathrm{SdsN} + 0.90\mathrm{SsssN} - 0.10\mathrm{SsOH} - 0.18\mathrm{SdO} - 2.77$ $F = 44$, $\lambda = 0.298$, $\lambda =146$ $t_{iw} = 67.19 - 32.94I\mathrm{Shannon} - 0.72\mathrm{Sum}I + 1.73\mathrm{Sum}\Delta I$ $F = 52$, $r^2 = 0.975$, $s = 0.8$, $s(\mathrm{cv}) = 1.0$, $\lambda = 12$, $p < 0.00001$ selection criterion: active if $0 > t_{\mathrm{iw}} > 0$, $9 > \mathrm{DF}_1 > 0$, and $10.5 > \mathrm{DF}_2 > 1.5$	124
antivirals (anti-herpes)	$\mathrm{IC}_{50} = -17.36({}^4\chi_p) + 41.39({}^4\chi_{pc}^{\nu}) + 21.71$ $\mathrm{F} = 38$, $r = 0.914$, $s = 0.6$, $n = 18$, $\mathrm{Cp} = 6.0$ $\log(\mathrm{ID}_{50}) = -1.42({}^0\chi) + 4.81({}^0\chi^{\nu}) - 11.41({}^3\chi_p^{\nu}) + 1.32({}^3\chi_c^{\nu}) + 4.17({}^4\chi_{pc}) - 8.42$ $F = 24$, $r = 0.929$, $s = 3.3$, $N = 25$, $C_{\mathrm{p}} = 5.04$ $\log(\mathrm{UDU}) = -4.67({}^1\chi^{\nu}) + 8.70({}^2\chi) - 3.64({}^3\chi_p) + 3.15({}^3\chi_p^{\nu}) - 8.05({}^3\chi_c) - 9.23$ $F = 41$, $r = 0.957$, $s = 12.4$, $N = 25$, $C_{\mathrm{p}} = 3.03$ $\mathrm{DF} = -1.17({}^0\chi^{\nu}) + 2.11({}^3\chi_p) + 2.79$ $F = 23.4$, $\lambda = 0.28$, $N = 81$ selection criterion: active if IC_{50} between -10 and 20, log ID_{50} between -5 and 3, log UDU between -4 and 4, and DF between -1 and 5.	111
cytostatics	$\mathrm{DF}_{\mathrm{cytostatic}} = -0.29 - 64.21J_5 + 2.89({}^4D_p)$ $F = 75$, $\lambda = 0.64$, $N = 264$ selection criterion: active if $\mathrm{DF}_{\mathrm{cytostatic}} > 0$	128
mycobacterium avium-M (quinolones)	$\mathrm{DF} = -2.6 + 20.1({}^3\chi_{\mathrm{CH}}) - 12.9({}^4\chi_c) + 42.5({}^4\chi_c^{\nu}) + 25.6({}^6\chi_{\mathrm{CH}}) - 2.2G_3^{\nu} + 2.4G_4^{\nu}$ $F = 31$, $\lambda = 0.37$, $N = 114$ selection criterion: active if $\mathrm{DF} > 0$	129
antibacterials	$\mathrm{DF}_{\mathrm{antibact}} = -3.635 + 0.934({}^0D) + 5.993\mathrm{DP}_1^{\nu}$ $F = 98$, $\lambda = 0.56$, $N = 355$ selection criterion: active if $\mathrm{DF}_{\mathrm{antibact}} > 0$	130
antifungals	$\mathrm{DF}_{\mathrm{antifung}} = 3.52 + 0.78G_1^{\nu} - 5.85G_5 + 34.85J_2 - 39.54J_2^{\nu} + 34.42J_3^{\nu} - 12.29({}^3\chi_p/{}^3\chi_p^{\nu}) + 4.21({}^3\chi_c/{}^3\chi_c^{\nu}) - 1.45\mathrm{PR}_0$ $F = 17$, $\lambda = 0.32$, $N = 90$ selection criterion: active if $\mathrm{DF}_{\mathrm{antifung}} > 1.0$	131
antimalarials	$\mathrm{DF}_1 = 0.56 + 3.25({}^3\chi_c) - 27.88({}^4\chi_c^{\nu}) - 8.64J_2$ $F = 9.8$, $\lambda = 0.55$, $N = 59$ $\mathrm{DF}_2 = -2.92 + 6.45({}^3\chi_c) - 2.78({}^4\chi_{pc}) + 0.39G_1^{\nu} - 1.72G_3 - 2.24({}^3C_c) - 2.62\mathrm{PR}_1 + 1.66\mathrm{PR}_2 + 0.04S$ $F = 8.9$, $\lambda = 0.35$, $N = 60$ selection criterion: active if DF_1 and DF_2 are >1.0	132
toxoplasma gondii	$\mathrm{DF}_1 = -26.99 + 3.06G_4^{\nu} + 76.20J_3 + 1.09({}^4C_c) - 1.30\mathrm{PR}_2 + 0.07S$ $F = 43$, $\lambda = 0.18$, $N = 66$ $\mathrm{DF}_2 = -54.6 - 2.8({}^4\chi_p^{\nu}) - 1.2G_2 + 1.4G_2^{\nu} - 4.6G_4 + 254.7J_4 + 176.9J_4^{\nu} - 190.8J_5^{\nu} + 3.3R + 1.4L - 1.5\mathrm{PR}_1 - 1.6\mathrm{PR}_3$ $F = 30$, $\lambda = 0.16$, $N = 98$; selection criterion: active if DF_1 and DF_2 are >0	133
non-narcotic analgesics	$\mathrm{DF} = -1.32({}^0\chi) + 4.67({}^1\chi) + 1.96({}^1\chi^{\nu}) - 6.56({}^2\chi^{\nu}) - 4.25({}^3\chi_p) - 4.11({}^3\chi_c) + 2.68({}^3\chi_p^{\nu}) + 13.31({}^3\chi_c^{\nu}) + 1.28({}^4\chi_p) + 11.75({}^4\chi_c) + 1.22({}^4\chi_{pc}) - 0.04$ $F = 9.3$, $\lambda = 0.198$, $N = 82$ $\log \mathrm{IC}_{50} = 0.32G_1^{\nu} + 6.34J_1 - 0.68V_4 + 1.4E - 2.25$ $F = 18$, $r = 0.908$, $s = 0.49$, $N = 20$ selection criterion: active if $\mathrm{DF} < 0.686$ and $3.5 > \log \mathrm{IC}_{50} > 0$	116, 117

the subject, and especially to the book by Temkin, Zeigarnik, and Bonchev.[26] In 1953, Christiansen[138] proposed a classification of reactions using diagrams similar to graphs. In 1956, King and Altman[139] further deepened the subject in an investigation of the derivation of the rate laws of steady-state reactions. In 1965, Temkin published a paper about the application of graphs to the analysis of steady-state reactions.[140] He proposed the concept of kinetic graphs, which reflect the structure of a mechanism in the space of intermediates. Work on the subject was developed further soon thereafter by Balaban,[141,142] who was followed by many others (see references in ref 26), especially in the 1970s. Graph theoretical studies of multiroute reactions with linear mechanisms make it possible to build a classification system based on the topological structure of reaction mechanisms and thus enumerate and code all classes of mechanisms involving any number of reaction routes. It will be assumed here that all reactions that constitute a reaction mechanism are elementary reactions, i.e., reactions which involve only one or two molecules. The term “reaction network” is

Table 24. New Biological Activities Discovered through Virtual Screening[a]

activity found	selected drugs	ref
cytostatic	6-azuridine, quinine	128
antibacterial	1-chloro-2,4-dinitrobenzene, 3-chloro-5-nitroindazole, 1-phenyl-3-methyl-2-pyrazolin-5-one, neohesperidin, amaranth, mordant brown 24, hesperidin, morine, niflumic acid, silymarine, fraxine	134, 130
antifungal	neotetrazolium chloride, benzotropine mesilate, 3-(2-bromethyl)-indole, 1-chloro-2,4-dinitrobenzene	131
hypoglycaemic	3-hydroxybutyl acetate, 4-(3-methyl-5-oxo-2-pyrazolin-1-yl) benzoic acid, 1-(mesitylene-2-sulfonyl) 1H-1,2,3-triazole	135
antivirals (anti-herpes)	3,5-dimethyl-4-nitroisoxazole, nitrofurantoin, 1-(pyrrolidinocarbonylmethyl)piperazine, nebularine, cordycepin, adipic acid, thymidine, α-thymidine, inosine, 2,4-diamino-6-(hydroxymethyl)pteridine, 7-(carboxymethoxy)-4-methylcoumarin, 5-methylcytidine	111
antineoplastic	carminic acid, tetracycline, piromidic acid, doxycycline	136
antimalarial	monensin, nigericin, vinblastine, vincristine, vindesine, ethylhydrocupreine, quinacrine, salinomycin	132
antitoxoplasma	cefamandole nafate prazosin andrographolide dibenzothiophene sulfone 2-acetamido-4-methyl-5-thiazolesulfonyl chloride	133
antihystaminic	benzydamine 4-(1-butylpentyl)pyridine *N*-(3-bromopropyl)phthalimide *N*-(3-chloropropyl)phthalimide *N*-(3-chloropropyl)piperidine hydrochloride 5-bromoindole	137
bronchodilator	griseofulvin, anthrarobin, 9,10-dihydro-2-methyl-4H-benzo 5,6-cyclohept[1,2-*d*]oxazol-4-ol, 2-aminothiazole, maltol, esculetin, fisetin, hesperetin, 4-methyl-umbellipheryl-4-guanidine benzoate	113
analgesics	2-(1-propenyl)phenol, 2′,4′-dimethylacetophenone, *p*-chlorobenzohydrazide, 1-(*p*-chlorophenyl)propanol, 4-benzoyl-3-methyl-1-phenyl-2-pyrazolin-5-one	116, 117

[a] For details, see the references in the last column.

normally used as a synonym for “reaction graph”, a construct which represents the different transformations of species participating in the elementary reaction. A reaction mechanism is said to be linear if it is expressed by a sequence of steps each of which contains at most one intermediate on one or both sides of the step. A nonlinear mechanism contains at least one elementary step in which the number of intermediates on one or both sides of the step is greater than one. A reaction graph is linear if there is no need to introduce bipartition (or multipartition) of edges and vertices. Bipartite reaction graphs are not linear because the set of vertices is divided into two proper subsets, and the vertices of different subsets are not equivalent. Vertices and edges of reaction graphs can have different labels that denote the ordinal numbers assigned to the corresponding reaction steps. In this section on reaction graphs, only linear reaction graphs will be treated in detail. Linear mechanisms can be described by both linear and nonlinear reaction graphs. However, nonlinear reaction graphs of linear mechanisms can always be reduced to linear reaction graphs without loss of generality. The vertices of kinetic graphs, as first proposed by Temkin, denote intermediates, while the edges (arcs) denote elementary steps. Intermediates are species that are produced in some reaction steps and consumed in other steps and do not appear in the overall reaction. Species that are not intermediates are called terminal. The probability (weight) ω_i of the ith step is assigned to the respective edge of a kinetic graph. Some species which are not involved in the overall stochiometry of a reaction are instead involved in reversible activation of a catalyst precursor, deactivation of a catalyst, or binding of intermediate species by reagents or products. These species are practically involved in mass balance of the intermediate species. Species of this sort are depicted with pendant vertices, i.e., with terminal vertices of degree one. As an example, let us take the three-step mechanism of the following overall catalytic reaction, with three intermediate species, I_1, I_2, and I_3: $A + B \rightleftarrows C + D$,

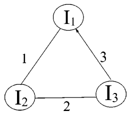


Figure 24. Kinetic graph of the three-step reaction mechanism. I_1, I_2, and I_3 are intermediates, which are normally represented by black vertices in the reaction graph.

$$A + I_1 \rightleftarrows I_2 + C \quad (1)$$

$$I_2 + B \rightleftarrows I_3 \quad (2)$$

$$I_3 \rightarrow I_1 + D \quad (3)$$

This mechanism is shown as a kinetic graph in Figure 24, in which the undirected edges 1 and 2 represent reversible reaction steps and the directed edge 3 represents an irreversible step. Let us now suppose that two new steps are added to this mechanism, i.e.,

$$I_1 + B \rightleftarrows I_1B \quad (4)$$

$$I_2 + D \rightleftarrows I_2D \quad (5)$$

Figure 25 shows the corresponding kinetic graph, in which two pendant vertices are now evident. The overall reaction, $A + B \rightleftarrows C + D$, is obtained by summing steps 1−3, upon which all I_i vanish. Figure 26 gives an example of a linear reaction network with two independent routes, five

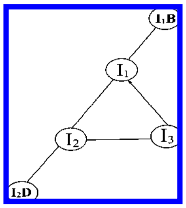


Figure 25. Kinetic graph with two pendant vertices used to describe the influence of two inactive intermediates, I_1B and I_2D.

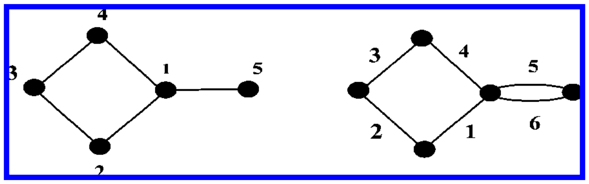


Figure 26. Kinetic graph (KG) of methane conversion by water vapor on Ni, whose mechanism is described in steps 1−6. Vertices 1−5 correspond to I, ICH_2, ICHOH, ICO, and IO, respectively. Edges 1−6 stand for the respective reaction steps 1−6.

intermediates, and six steps (1−6). It is the mechanism of methane conversion by water vapor on nickel,

$$CH_4 + I \rightleftarrows ICH_2 + H_2 \qquad (1)$$

$$ICH_2 + H_2O \rightleftarrows ICHOH + H_2 \qquad (2)$$

$$ICHOH \rightleftarrows ICO + H_2 \qquad (3)$$

$$ICO \rightleftarrows I + CO_2 \qquad (4)$$

$$I + H_2O \rightleftarrows IO + H_2 \qquad (5)$$

$$IO + CO \rightleftarrows I + CO_2 \qquad (6)$$

$$CH_4 + H_2O = CO + 3H_2$$

$$CO + H_2O = CO_2 + H_2$$

The intermediates from all of the steps are I (a divalent reaction site on the Ni surface), ICH_2, ICHOH, ICO, and IO (a chemisorbed species). The reaction route represented by the first of the overall equations incorporates steps 1−4, and the reaction route represented by the second overall equation incorporates steps 5 and 6. The KG (kinetic graph) cycles are given for these equations in Figure 26.

Normally, the task of classifying reaction graphs is facilitated by using kinetic face graphs (KFGs) of cycle graphs (CGs). Kinetic face graphs are defined on the basis of cycle adjacency in the initial kinetic graph. If KG is a kinetic graph with cyclomatic number μ, as defined in the treatment of dual indices, then μ independent cycles of KG induce a cycle graph CG which describes the adjacency of μ cycles of the kinetic graph KG upon which it is based [*B*(KG): the basis kinetic graph], with two vertices of CG being adjacent if the corresponding cycles of KG are adjacent. The loops of a KG are treated as cycles, and an edge is said to be a diagonal of a cycle when it joins two nonadjacent vertices of a cycle. Summing up, each vertex in a KFG (or CG) represents a cycle in the starting KG with a KFG edge representing the adjacency of a pair of KG cycles. When the edges of a cycle have a KG bridge between cycles in common, they are classified as an **A**-class cycle linkage. When the edges of a cycle have a vertex in common, they are classified as a **B**-class cycle linkage (Figure 27). And, finally, when the edges of a cycle have a KG edge in common, they are classified as a **C**-class cycle linkage. **A**, **B**, and **C** are the basic classes of linear reaction networks. Kinetic face graphs make possible the canonical numbering of KG vertices, edges, and cycles, which is what is needed for the coding and enumeration of linear networks (i.e., a reaction graph).

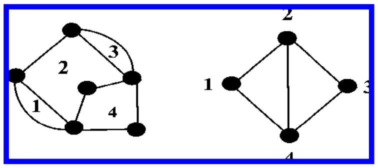


Figure 27. Kinetic graph (KG, left) and the corresponding kinetic face graph or cycle graph (KFG or CG). The CG edges [1,2], [2,3], and [2,4] are *C*-class cycle linkages, while [1,4] and [3,4] CG edges are *B*-class cycle linkages.

4.1.2. Categories and Subcategories

The principal factors determining the network category and subcategory of a KG are the number of cycles and vertices in it. The cycles in a graph are the basic determinants of graph complexity. Linear reaction networks are thus classified according to the number of independent routes, i.e., as *single-route* networks, *two-route* networks, etc. The number of graph vertices and graph edges are relevant to the number of intermediates and the number of elementary reactions (steps), respectively. The well-known Euler relation for a graph, $c = e - v + 1$, which ties together the number of cycles (c), the number of vertices (v), and the number of edges (e), lets us infer that in a KG the number of linearly independent routes is equal to c (also called the cyclomatic number), the number of steps is equal to e, and the number of intermediates is equal to v. This again means that either the number of steps (edges) or the number of intermediates (vertices) is independent once the routes are fixed. It is advantageous to use the KG vertices, since in multiroute networks there are fewer KG vertices than edges, i.e., fewer intermediates than steps, and for this reason the subcategories of *one-intermediate* network, *two-intermediate* network, etc. are useful.

It should be noticed that the original Euler relation is usually written as $v + f = e + 2$, where f is the number of faces. The given modified relation ($c = e - v + 1$) is correct for planar graphs (e.g., pyrene) but not for 3*D* structures such as adamantane, unless they are drawn as Schlegel diagrams[3,4,6]). In fact, the reduced formula does not account for the "outside" of a graph *as a face.* Thus, in adamantane from the Schlegel projections, one obtains $c = 12 - 10 + 1 = 3$, while from the Euler relation one obtains $f = 4$, since one of the four cycles is a linear combination of the three others.

4.1.3. Types

The way in which the KG cycles are connected determines the types and classes of linear networks. The previously described concept of a kinetic face graph, KFG, is central here. The larger the number of KFG edges, the more interconnected are the reaction routes, so that a more complex type of reaction network is produced in the hierarchical classification. Assigning serial numbers to KFGs according to the number of KFG edges to assign the type of a linear network is a satisfactory solution for linear networks with up to four routes, but starting with five-route networks,

Table 25. Enumeration of the Types (L) of Linear Reaction Networks with 1−5 Routes (KFG Vertices)

L M	0	1	2	3	4	5	6	7	8	9	10	Total
1	1											1
2		1										1
3			1	1								2
4				2	1	1	1					5
5					3	3	3	3	2	1	1	16

Table 26. Types L = 0−6 of Linear Reaction Networks Having M = 1−4 Routes

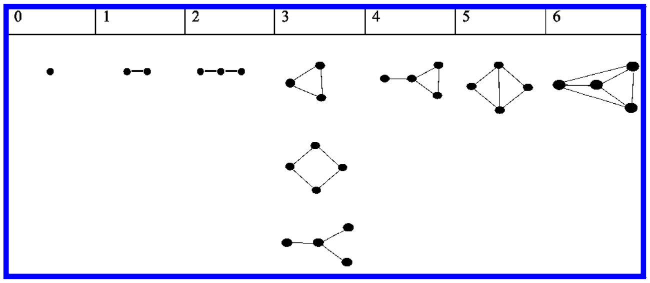
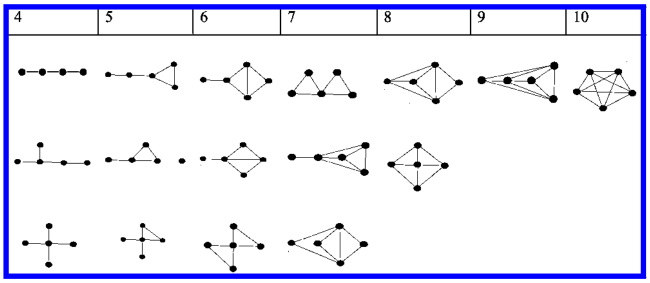


Table 27. Types L = 4−10 of Linear Reaction Networks Having M = 5 Routes



several groups of KFGs with the same number of edges exist, which would require additional classification criteria. However, to avoid introducing other serial numbers, only the number of edges will be used as a classification criterion. The number L of types of linear reaction networks with more strongly interrelated reaction routes increases when the number of edges is increased. Tables 25−27 illustrate the steadily increasing number and complexity of linear reaction networks having 1 to 5 routes with the number of KFG edges being from 0 to 10.

4.1.4. Classes

Classes of linear reaction networks within each type are defined depending on the manner in which pairs of reaction routes are interrelated. There are only three classes of two-route mechanisms: the already seen **A**-, **B**-, and **C**-classes, which correspond to a pair of KG cycles connected via a bridge, a common vertex, or a common edge, respectively (Figure 28). A bridge is the weakest kind of connection between two KG cycles. The cycle interrelation is stronger in class **B**, where the two routes share a common intermediate. It is particularly strong in class **C**, where the two routes share one or more steps. When dealing with multiroute networks, a fourth two-route class **Z** is formally introduced to mark the absence of linkages **A**, **B**, or **C** between the pair of corresponding KG cycles. The multiroute classes are combinations of the two-route classes, and they describe exhaustively all pairs of KG cycles. It is useful to introduce a larger classification unit termed a *generalized class* and denoted with $\mathbf{A}^a\mathbf{B}^b\mathbf{C}^c$, where a, b, and c are the total number of cycle adjacencies of types **A**, **B**, and **C**, respectively. Some examples of generalized classes are $\mathbf{A}^2$, $\mathbf{AB}^2$, $\mathbf{BC}^2$, $\mathbf{C}^3$, etc. The sum of the subscripts equals the total number of KFG edges ($a + b + c = L$), which determines the network type. The number of **Z** type indirect cycle linkages (here omitted) can be retrieved from $a + b + c + z = M(M - 1)/2$, where M is the number of routes. This relationship states that the sum of all four superscripts equals the total number of edges in the complete KFG with the same number of routes M.

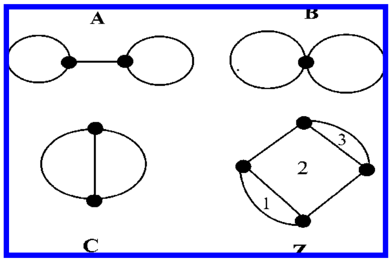


Figure 28. Four basic classes of linear two-route reaction networks. Class **Z** refers to the nonadjacent pair of cycles 1 and 3. Substituting any loop for a cycle of arbitrary size preserves the class.

4.2. Kinetics Graphs and Rate Laws

The general rule allowing one to use graphs for solving problems associated with a linear rate law, $y = ax$, is given in eq 33, wherein x_i and x_j relate to concentrations, while D_i and D_j are the determinants of the respective graph vertices.

$$x_i/x_j = D_i/D_j \tag{33}$$

Consider the KG graph of the following mechanism, which is depicted in Figure 29

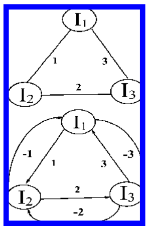


Figure 29. Kinetic graph of mechanisms 1−3 in standard (top) and expanded (bottom) forms.

$$\mathrm{A} + \mathrm{I}_1 \rightleftarrows \mathrm{I}_2 \quad (k_{+1}, k_{-1}) \tag{1}$$

$$\mathrm{I}_2 + \mathrm{A} \rightleftarrows \mathrm{I}_3 + \mathrm{B} \quad (k_{+2}, k_{-2}) \tag{2}$$

$$\mathrm{I}_3 \rightleftarrows \mathrm{I}_1 + \mathrm{B} \quad (k_{+3}, k_{-3}) \tag{3}$$

The weight ω_i of the ith arc is the ratio of the rate w_i of the ith step and the concentration of the respective intermediate species. Thus, for the KG of Figure 29, we have

$$\omega_{+1} = w_{+1}/[\mathrm{I}_1] = k_{+1}[\mathrm{A}][\mathrm{I}_1]/[\mathrm{I}_1] = k_{+1}[\mathrm{A}] \tag{34}$$

$$\omega_{-1} = w_{-1}/[\mathrm{I}_2] = k_{-1}[\mathrm{I}_2]/[\mathrm{I}_2] = k_{-1} \tag{35}$$

$$\omega_{+2} = w_{+2}/[\mathrm{I}_2] = k_{+2}[\mathrm{A}][\mathrm{I}_2]/[\mathrm{I}_2] = k_{+2}[\mathrm{A}] \tag{36}$$

$$\omega_{-2} = w_{-2}/[\mathrm{I}_3] = k_{-2}[\mathrm{B}][\mathrm{I}_3]/[\mathrm{I}_3] = k_{-2}[\mathrm{B}] \tag{37}$$

$$\omega_{+3} = w_{+3}/[\mathrm{I}_3] = k_{+3}[\mathrm{I}_3]/[\mathrm{I}_3] = k_{+3} \qquad (38)$$

$$\omega_{-3} = w_{-3}/[\mathrm{I}_1] = k_{+3}[\mathrm{B}][\mathrm{I}_1]/[\mathrm{I}_1] = k_{-3}[\mathrm{B}] \qquad (39)$$

The following steps may be used to derive the value of determinant D_i.

Step 1: Write the sum of weights of all arcs that leave each vertex, e.g.,

$$I_1{:}\quad \omega_{+1} + \omega_{-3},\quad \text{the first sum}$$

$$I_2{:}\quad \omega_{-1} + \omega_{+2},\quad \text{the second sum}$$

$$I_3{:}\quad \omega_{-2} + \omega_{+3},\quad \text{the third sum}$$

Step 2: Derive D_i* by taking the product of sums of step 1 so that the *i*th sum is excluded from the *i*th product, while others are preserved:

$$D_1{}^* = (\omega_{-1} + \omega_{+2})(\omega_{-2} + \omega_{+3});$$
$$D_2{}^* = (\omega_{-2} + \omega_{+3})(\omega_{+1} + \omega_{-3});$$
$$D_3{}^* = (\omega_{-1} + \omega_{+2})(\omega_{+1} + \omega_{-3}) \qquad (40)$$

Step 3: From the products derived in step 2, delete the terms that consist of weights of a circuit, remembering that in a trivial case these weights are the weights of the forward and reverse elementary steps. Thus, in the case under consideration, the terms $\omega_{+1}\omega_{-1}$, $\omega_{+2}\omega_{-2}$, and $\omega_{+3}\omega_{-3}$ should be deleted. Following these steps, we arrive at

$$D_1 = \omega_{-1}\omega_{-2} + \omega_{-1}\omega_{+3} + \omega_{+2}\omega_{+3};$$
$$D_2 = \omega_{+1}\omega_{-2} + \omega_{-3}\omega_{-2} + \omega_{+1}\omega_{+3};$$
$$D_3 = \omega_{-3}\omega_{-1} + \omega_{+1}\omega_{+2} + \omega_{-3}\omega_{+2} \qquad (41)$$

Formula 33 also allows deriving the concentration of any intermediate species through the concentration of a catalyst I_i, the concentration of a 0-species (which is equal to unity), or the total concentration of a catalyst. In the case of heterogeneous catalysis, the total catalyst concentration equals the total concentration of active sites on the catalyst surface, which is also unity. The related expressions are given in eqs 42 and 43.

$$[\mathrm{X}_i] = [\mathrm{I}_i]D_i/D_{\mathrm{I}1} \qquad (42)$$

$$[\mathrm{X}_i] = [\mathrm{X}]_\Sigma D_i/\Sigma D_i \qquad (43)$$

Here $[\mathrm{X}]_\Sigma$ is the overall concentration of all catalysts. If concentrations are the proportion of catalytst surface occupied by X_i, then $[\mathrm{X}]_\Sigma = 1$. For a noncatalytic reaction, we can derive eq 44, in which D_0 is the determinant of a zero species.

$$[\mathrm{X}_\mathrm{i}] = D_\mathrm{i}/D_0 \qquad (44)$$

For a kinetic graph with pendant vertices, such as in Figure 25, where there are two pendant vertices, the edges incident with these vertices depict equilibrium steps within steady-state or pseudo-steady-state processes. In a graph containing pendant vertexes, one can specify as in eq 45 the function F_i associated with each vertex *i*, which characterizes the extent to which a catalyst is bound by ligands or species.

$$F_i = 1 + \Sigma_\Sigma(\omega_\mathrm{s}/\omega_{-\mathrm{s}}) \qquad (45)$$

In eq 45, the sum is taken over all pendant vertexes *s* adjacent to vertex *i*. Thus, eq 43 can be modified to yield eq 46.

$$[\mathrm{X}_i] = [\mathrm{X}]_\Sigma D_i/\Sigma_i F_i D_i \qquad (46)$$

Thus, in the KG of Figure 25, functions *F* of vertices I_1 and I_2, which are adjacent to these two pendant vertices, are $F_1 = 1 + K_1[\mathrm{B}]$ and $F_2 = 1 + K_2[\mathrm{D}]$, respectively. Then, if $[\mathrm{X}]_\Sigma = 1$, we have

$$[\mathrm{I}_3] = D_{\mathrm{Z}3}/(D_{\mathrm{I}1}F_1 + D_{\mathrm{I}2}F_2 + D_{\mathrm{I}3}) \qquad (47)$$

$$[\mathrm{I}_1] = D_{\mathrm{Z}1}/(D_{\mathrm{I}1}F_1 + D_{\mathrm{I}2}F_2 + D_{\mathrm{I}3}) \qquad (48)$$

The rate of reaction which corresponds to the mechanism of Figure 25 can be written in two ways, as shown in eqs 49 and 50.

$$r = w_{+3} - w_{-3} = \omega_{+3}[\mathrm{I}_3] - \omega_{-3}[\mathrm{I}_1] \qquad (49)$$

$$r = (\omega_{+1}\omega_{+2}\omega_{+3} - \omega_{-1}\omega_{-2}\omega_{-3})/(D_{\mathrm{I}1}F_1 + D_{\mathrm{I}2}F_2 + D_{\mathrm{I}3}) \qquad (50)$$

If $[\mathrm{X}_i]$ in eq 46 is a concentration of a free catalyst (I_i), the quantity F_{st} in eq 51 can be viewed as a reciprocal value of the free catalyst steady-state concentration normalized to the total catalyst concentration, i.e.,

$$F_{\mathrm{st}} = [\mathrm{X}]_\Sigma/[\mathrm{X}_j] = \Sigma_i F_i D_i/D_j \qquad (51)$$

D_j in eq 51 is the determinant of the *j*th vertex. Obviously, if the process approaches equilibrium, F_{st} transforms to the reciprocal normalized concentration of a catalyst at the equilibrium point, as shown in eq 52.

$$F_{\mathrm{eq}} = [\mathrm{X}]_\Sigma/[\langle \mathrm{X}_j\rangle]_{\mathrm{eq}} \qquad (52)$$

Now we can derive the expressions given in eq 53,

$$[\langle \mathrm{X}_j\rangle]_{\mathrm{eq}}/[\mathrm{X}_j] = F_{\mathrm{st}}/F_{\mathrm{eq}} = \Sigma_i F_i D_i/D_j F_{\mathrm{eq}} \qquad (53)$$

and if the process approaches equilibrium so that $[\langle \mathrm{X}_j\rangle]_{\mathrm{eq}}/[\mathrm{X}_j] \rightarrow 1$, we obtain eq 54,

$$\Sigma_i F_i \langle D_i\rangle = \langle D_j\rangle F_{\mathrm{eq}} \qquad (54)$$

in which $\langle D_i\rangle$ and $\langle D_j\rangle$ are determinants of the *i*th and *j*th vertices that involve equilibrium or pseudoequilibrium concentrations of reagents and products. Equation 54 relates the kinetics and thermodynamics of complexation reactions and shows how coefficients of a rate law are related to the observed equilibrium constants as well as which of the complexes of rate constant should be associated with the equilibrium constants.

5. Application of Chemical Graph Theory to Biomacromolecules

5.1. Descriptors for Biomacromolecules

5.1.1. Polypeptides

Randić has introduced the molecular and shape profiles to quantitatively describe the molecular structure.[143,144] This approach is interesting because it allows the characterization of the 3D structure of small as well as large molecules. Molecular profiles were applied to protein 3D-sequences.[145]

Thus, Randić and Krilov found that the molecular profile approach is able to characterize the 3D structures of schematic representations of 27 amino acid peptides composed of polar and hydrophobic amino acids occupying all sites of a 3 × 3 × 3 cube, as proposed by Li et al.[146] More recently, these researchers have considered the D/D matrix, in which the elements are the quotients between the respective geometrical (or Euclidean) and topological (number of edges) atomic distances.[147] Matrices $^k(D/D)$ are then obtained by raising the elements of D/D to a power k. The folding profile of a structure is the sequence of indices $^k\Phi$, defined as the leading eigenvalues of $^k(D/D)$ divided by the number of vertices of the chain considered. For a given k, the $^k(D/D)$ elements of a more folded structure will be smaller than these elements calculated for a less folded one, and this is directly related to the matrix leading eigenvalue. The folding profiles have been successfully applied to the characterization of the different secondary structures found in proteins.

Randić et al.[148] have also devised a different approach based on a graphical representation of DNA triplets to obtaining descriptors useful for the description of protein sequences.

Estrada has proposed and developed descriptors useful in accounting for the 3D molecular structure in the description of features specific to polypeptides as well as in measuring the similarities between pairs of proteins.[149] In this approach, the 3D-structure index was defined as $I = \mathrm{tr}(\mathrm{e}^{\mathbf{B}})$, where **B** is the adjacency matrix of the dihedral angles but with the diagonal elements being the cosine of the corresponding angle.[149] This index accounted for the degree of folding in small molecules as well as in peptide models.[149]

The torsion degree in a backbone chain, which is related to the protein folding degree, can be characterized by the I_3 index defined in eq 55,[150,151]

$$I_3 = \frac{1}{t}\sum_{i=1}^{t} \exp(\lambda_i) = \langle \mathrm{e}^{\lambda} \rangle = \frac{\mathrm{tr}(\mathrm{e}^{\mathbf{B}})}{t} = \frac{I}{t} \qquad (55)$$

in which t is the number of dihedral angles in the backbone and λ_i are the eigenvalues of the matrix **B**. If a peptide has N amino acids, $t = 3N - 2$, since each amino acid has three dihedral angles and the extremes do not contribute. This definition of I_3 as a folding degree index is justified, since it accounts for the intuitive folding order for a given backbone chain having the same dihedral angles in different bond positions.[151,152] The computation of I_3 from its definition is cumbersome for the usual proteins, but it can be calculated with the required precision from the spectral moments μ_k of **B**. The k-th spectral moment of an $n \times n$ matrix **B** is the sum of the main diagonal elements (i.e., the trace) of the matrix **B** raised to the k-th power. This magnitude is related to the eigenvalues λ *of* **B** by eq 56.

$$\mu_k = \mathrm{tr}(\mathbf{B}^k) = \sum_{i=1}^{n} \lambda_i^{\,k} \qquad (56)$$

By using the definition of I_3, the MacLaurin series, and the relationship of spectral moments to eigenvalues, we find that

$$I_3 = \frac{1}{t}\sum_{i=1}^{t} \exp(\lambda_i) = \frac{1}{t}\sum_{k=0}^{\infty}\sum_{k=1}^{t}\frac{\lambda_i^{\,k}}{k!} = \frac{1}{t}\sum_{k=0}^{\infty}\frac{\mu_k}{k!} \qquad (57)$$

The contribution of spectral moments of orders higher than 10 is negligible, and the truncation at a point after this value is feasible without any loss of precision in the calculation of I_3.

The folding degree, measured as I_3, does not correlate with the number of amino acid residues, nor with several published measures of the protein packing, such as Pt (Liang and Dill’s total protein packing), OSP (Pattabiraman, Ward, and Fleming’s occluded surface packing), and RG (the radius of gyration of the protein).[151] These last results suggest that the concepts of folding and compactness are independent. Another interesting result regarding I-based peptide similarity is that the most folded peptides are not similar to each other while the less folded ones are. This is apparently because the more folded proteins have more specific functions than the less folded ones. This index has been successfully applied to the analysis of 3D protein similarity in lysozymes;[153] the numerical characterization of the protein secondary structure as demonstrated by a representative set of proteins;[151] and the influence of factors external to the amino acids sequence such as the effect of temperature on ribonuclease A.[151] The correlation of the reduction potential of azurins and pseudoazurins with the degree of folding was accomplished by a local version of the I_3 index.[152]

Proteomic maps have also been characterized through useful graph representations and invariants. Several models have been published based on $^k(D/D)$ matrices and folding profiles of selected spots,[154] the partial ordering of map spots and the distance−adjacency matrix,[155,156] the adjacency graphs of spots at a distance smaller than a critical value and different classes of associated matrices,[157] a similarity index and self-organizing map,[158] the nearest neighborhood of spots, [159−161] a canonical labeling of the vertices of a Hasse diagram embedded in the spot adjacency matrix,[162] and complex weights joined by relative abundance and matrix invariants.[163] For a review on this topic, see ref 164. Proteomic biodescriptors have been useful in the characterization of toxicity and, eventually, when combined with molecular descriptors, in its quantitative prediction.[165,166]

5.1.2. Polynucleotides

Nucleic acid sequences can be condensed into representations and numerical invariants to explore similarities between sequences, coding, ordering, and structure−activity correlation. Randić et al. have recently been exploring this interesting field. First, a condensed representation of sequences was developed to facilitate the comparison of similar sequences from different sources.[167] Then, a method to quantify the similarity of sequences was published.[168] This method relied on 2D walks defined by the sequence of bases and the corresponding eigenvalues of D/D matrices between bases. More approaches to nucleic acid sequence coding have been outlined:

- DNA profiles consisting of average matrix elements,[169] or the leading eigenvalues[170] of the 4 × 4 matrix that contains the average distance between base pairs, and other matrices whose elements are the higher powers of these average distances.
- Eigenvalues of matrices extracted from 4 × 4 × 4 tensors that contain the frequencies for each triplet of consecutive bases to be found in the sequence.[171]
- Directed walks on the plane generated by vectors that are functions of the type of base.[172]
- Process-control-like graphs that represent the sequence order in abscissas and base in ordinates, and diverse matrices

associated with the distances between vertices such as the Euclidean distances, the *D*/*D* matrix, or the (*L*/*L*) matrix, where the quotient between 3D Euclidean distances and the sum of geometrical lengths (by power *k*) of edges between two vetices are considered.[173] Invariants from these matrices and also from the matrix $^k(L/L)$, whose elements are elevated to power *k*, are discussed in ref 174.

- Dots on a worm curve representing an arbitrary two-digit binary code that encodes the different bases, while the nearest neighbor distance invariants encode the dots.[174,175]
- A cube with 3 × 3 × 3 edges whose vertices represent each codon in which a directed walk follows the sequence of codons, while the elements of the Euclidean distance matrix to power *k* represent the codon pairs.[176]
- Similarities based on bandwidth averages of Euclidean distance matrices obtained from a 4D representation in which, for every base in the sequence, each coordinate equals the number of A, T, G, and C found in the sequence until the base considered.[177]
- *N*-dimensional Jeffrey-like walks and Ward method for hierarchical clustering of DNAs from different species.[178]
- Four-color maps constructed from spiral sequence arrangements that give ten kinds of topological distance matrices between each pair of color regions and the average matrix elements associated with them.[179]

6. Meaning of Basic Molecular Connectivity Indices

In the following, four proposed interpretations of MC indices will be given. Recently, interpretations of other relevant topological indices, such as the Wiener index, the Hosoya index, the Balaban index, and the Harary index, have been proposed. The reader interested in these interpretations should consult ref 180.

6.1. The Molecular Connectivity Index—A Quantum Interpretation

It is beyond doubt that connectivity indices yield excellent results in the prediction of physical, chemical, and biological properties; however, no definitive statements can be made as to the physical meaning of such indices in particular and, more generally, even for other graph theoretical descriptors. Galvez has published papers providing some insight along these lines. In an initial paper,[181] he demonstrated that the first-order connectivity index may be related to both electronic and vibrational energies of alkenes and conjugated hydrocarbons according to the following formula (eq 58):

$$E_\pi = N_\pi \alpha + 4\beta^1\chi \qquad (58)$$

Here α and *β* stand for the Coulomb and resonance integrals, respectively, as defined in the original Hückel MO theory and N_π is the number of *π* electrons of the conjugated hydrocarbon. This equation allows for a very good fitting to, for example, the Hückel results for resonance energies. Moreover, the equation can be refined into a more accurate version if we include the influence of the overlap integral, *S*. On the other hand, it is easily seen that vibrational energy may also be expressed as a function of the topological valence *δ*. Thus, the values ranging from 1634 to 1675 cm^{-1} of the vibrational frequencies for all substituted ethylene derivatives follow surprisingly well the relationship in eq 59

$$\nu\ (\mathrm{cm}^{-1}) = 1780.6 - 147.2[(1/\delta_1) + (1/\delta_2)]^{1/2} \qquad (59)$$

where the subscripts 1 and 2 refer to the ethylene carbons.

In another paper, Galvez[182] has further developed an interpretation presented by Estrada and based on the concept of accessibility. The formalism is purely geometrical and is based on (i) the concentric spheres representing the covalent and Van der Waals volumes, (ii) the relation between molecular volume and surface area in alkanes, and (iii) connectivity indices. The demonstration is based in the loss of accessible volume per atom as two atoms bind each other. In this paper, the representation as concentric spheres of the covalent and Van der Waals volumes as well as the loss of accessible volume given by the intersection of the Van der Waals volumes when the two covalent spheres are in contact provide a measure of the loss of accessibility. According to these results, the molecular volume and surface area of alkanes can be expressed as a function of $^0\chi$ and $^1\chi$, i.e., the zero-order and first-order connectivity indices, respectively. The linear regression equations yield values of $r = 0.987$ and $r = 0.991$ for volume and surface area with $^0\chi$ and $^1\chi$, respectively. These results fit well with the known fact that with an increasing degree of branching, i.e., number of tertiary or quaternary carbons, the molecular volume increases while the molecular surface area decreases. The key role played by the molecular volume and surface area is well-known for experimental properties ranging from molecular polarizibility to boiling temperatures and, in general, for all intermolecular forces and interactions. The molecular accessibility, *A*, is an important concept which aids in deriving the given model for the connectivity indices. Alternative definitions of this concept can be given in terms of connectivity indices, as, for example, in eq 60.

$$A = 4(^1\chi^v) - {^0\chi^v} \qquad (60)$$

Despite the simplicity of this definition, it is possible to predict the increment of the standard free energy (and hence the spontaneity of the process) for positional isomerization between pairs of hydrocarbons and even more complex molecules. For instance, the equilibria between *n*-butane and isobutane, between 2-methylpentane and 3-methylpentane, between propylamine and isopropylamine, and between *o*-methylaniline and *m*-methylaniline may all be predicted. All these features reinforce the idea that at least some important physical and geometrical molecular parameters can be expressed as functions of topological indices and thus bypass the need for cumbersome calculations. Altogether, Galvez's results support the idea that molecular topology is not only an alternative but also an independent approach as compared to conventional methods based on quantum or classical mechanics.

6.2. The Molecular Connectivity Index—A Kinetic Interpretation

The *kinetic* interpretation of the first-order molecular connectivity index,$^1\chi$, was proposed by Kier and Hall.[57,183] The Kier–Hall interpretation turns around the concept of bimolecular encounter accessibility, A_{ij}. This interpretation is centered on the bond index contribution of $^1\chi$ (see eq 5), $C_{ij} = (\delta_i\delta_j)^{-0.5}$. This algorithm encodes the relative accessibility of the *ij* biatomic fragment of a molecule in encountering another fragment of a different molecule, i.e.,

$C_{ij} = A_{ij}$. The δ values themselves are thus viewed as a count of neighboring atoms bonded to an atom in the HS chemical graph, which corresponds to the number of σ electrons contributed by that atom to bonded atoms, while the $1/\delta$ values are viewed as the fraction of the total number of σ electrons contributed to each bond formed with that particular atom. In an alkane chemical graph (excluding the trivial case of ethane, where the biatomic fragment and the molecule coincide), the (δ_i, δ_j) values vary widely, ranging from (1, 2) for highly exposed fragments to (4, 4) for highly buried fragments. The corresponding A_{ij} values range from 0.707 to 0.250. Thus, exposed ij fragments with large A_{ij} values have a high accessibility while buried fragments have a low accessibility and low A_{ij} values. Let us now imagine that pairs of such fragments on two different molecules, M and N, undergo a bimolecular encounter. This would yield four different types of encounters, i.e., $(1, 2)_M$-$(1, 2)_N$, $(1, 2)_M$-$(4, 4)_N$, $(4, 4)_M$-$(1, 2)_N$, and $(4, 4)_M$-$(4, 4)_N$, and the single accessibilities of each fragment are 0.707 and 0.250. If the encounter probability between the two fragments is assumed to be the product $p_{ij,kl} = (A_{ij})_M(A_{kl})_N$, then for the four different types of encounters (rounding): $p_{12,12} = (p_{12})^2 = 0.500$, $p_{12,44} = p_{44,12} = 0.177$, and $p_{44,44} = (p_{44})^2 = 0.063$. The model can present some degeneracies; that is, encounters with (2, 2) fragments and with (1, 4) fragments are equiprobable. The total bimolecular encounter probability for the present (rather strange) case is $p_T = (p_{12})^2 + 2p_{12,44} + (p_{44})^2 = (p_{12} + p_{44})^2 = p_M p_N = (0.957)^2 = 0.916$. Here, but not always, $p_M = p_N = 0.957$. The interesting final result is that $p_{12} + p_{44} = {}^1\chi(M) = {}^1\chi(N)$. Thus, the total bimolecular encounter probability for any molecule which is encoded by a chemical graph is $p_T = p_M p_N = {}^1\chi(M)^1\chi(N) = \sum(A_{ij})_M \sum(A_{kl})_N$, and the single probabilities are $p_M = {}^1\chi(M) = \sum(A_{ij})_M$ and $p_N = {}^1\chi(N) = \sum(A_{kl})_N$. Thus, we see that the molecular connectivity index, a graph-theoretical index, encodes information about bimolecular interactions by partitioning them into the interactions among the single biatomic fragments of a molecule.

For hydrogen-suppressed chemical pseudographs, which encode atoms other than carbon atoms and for which the index${}^1\chi^v$ is used, things are more complex. Here, different types of bimolecular encounters are possible among the different fragments comprising the molecule, as in, for example, an apolar−apolar external encounter with high probability such as CH_2-CH_3 on CH_2-CH_3 or an apolar−polar external encounter such as C−OH on CH_2-CH_3 with low probability due to the large $\delta^\nu(ps)_{OH} = 5$, as expected. Now, the polar−polar external encounter (e.g., C−OH on C−OH) has within this model a very low probability, which is a seemingly unexpected result. Actually, a way out of this difficulty is to admit that the polar encounter cannot be treated in the same way as the apolar one. It could be suggested that polar fragments interact, giving rise to van der Waals bimolecular (or supramolecular) entities, and that the encounters are among these entities. The total bientity encounter probability between [M−N] and [O−P] could be defined as $p_T = p_{MN}p_{OP} = {}^1\chi^\nu(M-N)^1\chi^\nu(O-P)$, where $p_{MN} = {}^1\chi^\nu(M) + {}^1\chi^v(N)$ and $p_{OP} = {}^1\chi^\nu(O) + {}^1\chi^\nu(P)$. This approach as well as one centered on the concept of a variable connectivity index${}^1\chi^f$ should be pursued. The suggestion that connectivity indices should be discussed in terms of their partitioning into bond contributions to better differentiate between contributions from exposed and buried fragments has recently been further developed and extended to other type of indices.[180]

6.3. The Molecular Connectivity Index—A Geometric Interpretation

This interpretation advanced by Estrada[184] is centered on the concept of the molecular accessibility defined as Acc(i) = $\alpha(\delta_i)^{-0.5}$, where α is a proportionality constant. With this definition, the zeroth-order molecular connectivity index is strictly related to the accessibility parameter. We also have Acc(i) = $\beta(L - I)$, where L is the circle perimeter surrounding the atom, I is the arc shared with the circumference of the neighbor atom, and β is a proportionality constant which can be seen as the specific atomic accessibility and set equal to one, giving Acc(i) = $(L - I)$. Thus, the inverse square root of the vertex degree can be seen as a component of the *relative atomic accessibility perimeter* (RAP). The first-order molecular connectivity index, ${}^1\chi$, is made up of bond contributions of the type $C_{ij} = (\delta_i)^{-0.5}(\delta_j)^{-0.5}$, which, from what has been said, represents nothing else than the relative bond accessibility areas (RBAs). Now, since ${}^1\chi = \sum_{i\neq j} C_{ij}$, the ${}^1\chi$ can be considered as the relative molecular accessibility area (RMA) when molecules, for this and the previous index, are encoded with pure hydrogen-suppressed chemical graphs. These areas represent the total areas that are accessible from the environment surrounding the molecules, and the smaller the $(\delta_i\delta_j)$ is, the larger the area contributed by a single C_{ij}, and consequently the accessibility of the ij group, is. From this point of view, a cycle is made up of equiaccessible points, while a chain is more accessible at its external points while buried vertices are hardly accessible. In going to pseudographs, which can encode multiple bonds and nonbonding electrons, and when δ^ν replaces δ, things become more complicated, but even here, the concept of accessibility continues to be valid. For pseudographs, a physical meaning should be given to the circles belonging to the vertices, and this can be done by introducing the van der Waals radii, r_W. These radii decrease with an increase in the number of electrons in the valence shell. Now, it is possible to define the circumference, L, and the length of the overlap arc, I, as $L = 2\pi r_W$ and $I = \theta r_W$, where θ is the overlap angle between two adjacent circumferences. Figure 30, top, shows the case of two overlapping circles, centered at C and C'. Here, r_W is the van der Waals radii (distance CP or CP′), r_{cv} is the covalent radii (bond distance CO), and θ is the angle formed by PCP′. The definition for the atomic accessibility for groups with the same number of neighbors formally does not change, and in fact, Acc(i) = $(L - I) = r_W(2\pi - \theta)$. With multiple bonds, the atomic accessibility is accounted for by the change in the covalent radii, r_{cv}, of the atoms supporting multiple bonds, which is reflected in θ and thus in Acc(i).

In the bottom part of Figure 30, the upper values are the Acc(i) values as a function of the δ_i^ν. Here, a clear proportionality is evident between both parameters, and it is evident that *Acc*(i) depends on a negative power of δ_i^ν. Thus, for groups with the same number of neighbors, *Acc*(i) = $\alpha'(\delta^\nu{}_i)^{-p}$, where a p value of 0.5 can be suggested. For groups with a different number of neighbors, the perimeter of the atom is overlapped by the circumference of each neighboring atom. Here, the Acc(i) definition changes only a bit, being now Acc(i) = $(L - \delta I)$, where δ accounts for the number of neighbors bonded to the corresponding atom. In the bottom part of Figure 30, the bottom values are the Acc(i) values plotted as a function of the δ_i^ν for the case of two neighbors. Comparing the two cases, we see that Acc(i) is made up of two contributions: one is related to the

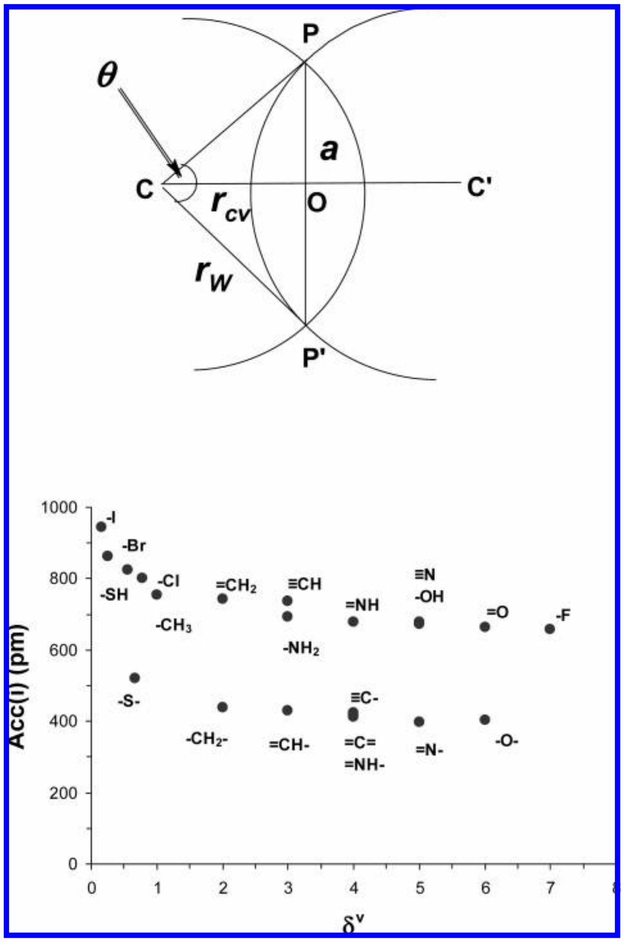


Figure 30. (Top) Scheme of the van der Waals circumferences of two neighboring atoms. (Bottom) Plot of the atomic accessibility perimeter, Acc(i) in pm, versus the valence degree, δ^ν, of different heteroatoms bonded to one (top values) or two neighbors (bottom values).

pseudograph aspect of the molecule and the other only to its graph aspect. Both contributions can be encoded with a single formula, i.e., Acc(i) = $128(\delta_i^\nu)^{-0.5} + 1149(\delta_i) - 511$.

Concerning higher-order indices, let us consider as an example the second-order path index, ${}^2\chi_p$, which is defined by ${}^2\chi_p = \sum C_{ijk}$, with $C_{ijk} = (\delta_i)^{-0.5}(\delta_j)^{-0.5}(\delta_k)^{-0.5}$. This formula can be considered as the multiplication of three accessibility parameters which encodes a volume; that is, this term is the volume that is accessible from the outside to three adjacent atoms in a path of length 2 in a molecule. Similarly, higher-order connectivity indices can be thought of as volumes in hyperspaces. A deeper study of the exponent p in Acc(i) = $\alpha'(\delta^\nu{}_i)^{-p}$ reveals that the optimal value for p is 0.27. If higher row atoms are considered instead, then the contribution of inner shell electrons should be included, as is done in the δ^ν definition of eq 3, even if, here, the older Kier−Hall[13] definition for δ^ν is used.

6.4. The Molecular Connectivity Index—A Variable Index Interpretation

The variable connectivity index[64−70] can be considered as the index of an unknown pseudograph, as is evident from its adjacency matrix, where x is on the diagonal. The "reality" of the unknown pseudograph is fixed by the value of x (some of these arguments are valid for y_H). Let us look at some validity domains for x, which can give rise to different mathematical objects (see eq 23). For $x \ll \delta$, ${}^1\chi^f = {}^1\chi$; that is, the unknown pseudograph goes over into the corresponding graph of the molecule. For $x \sim \delta$, ${}^1\chi^f = {}^1\chi^v$; that is, the unknown pseudograph becomes practically a pseudograph, and x represents multiple bonds or/and self-connections. In this context, we could even speak of partial pseudographs, as x could represent partial multiple bonds and/or partial multiple self-connections. With this last concept, we are moving away from graph theory and into the domain of a non-graph-theoretical index whose real entity will become clear in the next domain. For either $x \gg \delta$ or $x \ll 0$, we have${}^1\chi^f = 1/x$; that is, the graph characteristics completely disappear, and we have a pure "*ad hoc*" computational index. For $x < 0$ and $|x| \approx \delta$ and approaching δ either from the left or from the right, ${}^1\chi^f = 1/\epsilon \gg {}^1\chi$, with $0 < \epsilon \ll 1$; that is, the graph is approaching the graph of a set of nonconnected vertices with $\delta = 0$. Here, the molecular characteristics of the graph go by the board, and we face a practically "*atomic*" graph. Further, analyzing eq 24, if $x \gg \delta$ and $y_H < 0$, but $|y_H| \approx \delta$, then the whole graph shrinks into the graph of its part, i.e., that part of the graph which encodes the heteroatom only. In many of these cases, the graph interpretation of the molecule blurs even though the chemical interpretation of the computation is enhanced, since the main and the secondary contributions to the model can be better distinguished. Theoretically, the ${}^1\chi^f$ index, even though defined as a graph-theoretical index within the frame of graph theory, ends up on one side, rendering the purely graphical representation of the molecule superfluous, but gives rise on the other side to a new kind of "*ad hoc*" molecular computational index.

The complete graph algorithm (see eq 3) may also have a variable solution when q is subjected to an optimization procedure. Here also, there are some domains where the graphical interpretation blurs. For $q \gg p \cdot r$, we have $\delta^\nu \gg \delta^\nu(ps)$, and the graph characteristics of the molecule go by the board, as also happens for $0 < q \ll 1$, giving $\delta^\nu \ll \delta^\nu(ps)$. In both cases, together with the case $q < 0$, δ^ν becomes an "*ad hoc*" computational parameter, and it gives rise to "*ad hoc*" molecular computational indices.

7. Conclusions

During the present excursion through some quite recent developments in mathematical chemistry, and especially in chemical graph theory, we have encountered a wealth of new graph-theoretical concepts and their successful applications in modeling techniques which, as we have seen, extend even to chemical kinetics and some characteristics of biomacromolecules. The field of mathematical chemistry and of the corresponding graph-theoretical methods applied to chemistry is an ever increasing field; the reader ought to see Randić's recent review on phenomena which are usually treated by quantum chemistry methods.

There are phenomena, especially in physics, that can be interpreted by "absolute theories", i.e., by theories that can quantitatively predict a certain property of a system from its fundamental parameters without using "adjustable external" parameters. In these predictions, balancing computational efficiency with scientific accuracy is always a compromise, and in chemistry (and especially in physical chemistry), this is a rather difficult task to achieve without the help of "adjustable external" parameters, i.e., without the help of "nonabsolute theories". This is what happens, e.g., with quantum chemistry whose modeling ability is mainly based on adjustable external parameters, and even in the newest approach in this field, the density functional theory (DFT), there is a lot of (inspired) guesswork involved. Problems arising with quantum methods in predicting properties have brought into the field classical molecular mechanics (MM) methods, where molecules are seen as a group of tiny moving balls attached to each other by massless springs. Chemical graph theory was brought into the field in the second half of the 19th century to aid in solving a chemical problem, the enumeration of organic isomers, by also using an "absolute theory". It has since then developed into a highly structured field of scientific endeavor, sometimes remaining an absolute theory and at other times developing into a "semiempirical" theory. In both cases, chemical graph theory has been able to adhere to a more than satisfactory degree to Giordano's dictum regarding a physicist's task:[185] "*A physicist is a person who can calculate anything within an order of magnitude*."

8. Acknowledgments

The present review owes its final format to the many suggestions of the three reviewers and especially to the accurate editing of the language editor John Downing. Needless to say, errors, omissions, and oversights are the responsibility only of the authors. One of the authors (L.P.) would like to thank the other authors for their helpfulness and hospitality during his sabbatical at their department. The authors (J.G. and R.G.D.) acknowledge financial support from the Fondo de Investigación Sanitaria, Ministerio de Sanidad, Spain (project: SAF2005-PI052128). Financial support from the I3P program of the Spanish 'Consejo Superior de Investigaciones Cientificas (CSIC)' is acknowledged by J.V.d.J.-O.

9. Appendix

9.1. An Update

The present review is not a book and for this reason cannot cover every new aspect of chemical graph theory, which has consistently grown during these last years. This means that some "new trends" have been left out to avoid overburdening the review. In any case, some (but not all) aspects of this growth can be briefly mentioned here. References 186−190 develop applications of topological computational methods to be used for the development of new drugs, among them being antihistaminic, antimalarial, antitubercolotic, and cytostatic drugs. References 191−193 develop further interesting work on a mathematical characterization of proteomics and DNA. Reference 191 is just a reference book, and ref 192 is a paper about a dose/response curve for proteome which illustrates for the first time the detection of hormesis on the cellular level (up to now it has been known at a whole organism level). Reference 193 illustrates instead the use of a graphical representation for solving the DNA alignment problem. Concerning reaction graphs, three international works of Kvasnicka et al., who normally published in *Collect. Czech. Chem. Commun*., deserve to be cited, and they can be found as refs 194−196. Among Estrada's many achievements during these last years are not only further works on the graph-protein problem[197,198] but also works on the generalized topological indices (GTI),[199−202] which would by themselves deserve a review.

Last but not least, let us cite the stereoisomer enumeration problem by graph methods, which is largely treated in Fujita's last volume (Vol. 4) of the *Mathematical Chemistry Monographs* edited by *MATCH* and cited at the end of this Appendix. Actually, Vols. 1, 3, and 4, and the ones not yet published, offer a wide and deep perspective on chemical graph theory. Volume 2 is a nontechnical introduction to some general aspects of science and is intended for the reader with no more than high school algebra. In it the author tries to explain the nature of some basic concepts in mathematics and why, when, and how they are used in science and the humanities.

9.2. Graph-Theoretical Software for Model Purposes

9.2.1. APPROBE (& POLLY)

These two QSAR/QSPR related programs have been developed at the Natural Resources Research Institute, University of Minnesota, Duluth, MN, under the leadership of S. Basak. (http://wyle.nrri.umn.edu/Basak/). Both programs have been licensed to Glaxo, now part of Glaxo Smith Kline and Upjohn, now part of Pfizer. APPROBE is the acronym for *Atom Pair PROBE*. It is a C program to manipulate either atom pairs or topological torsions to determine structure similarity/dissimilarity. It accepts SMILES strings as molecular structure input. Using APPROBE, a descriptor file or library file of structures may be generated. This descriptor file may be either atom pairs or topological torsions. SMILES is the acronym for *Simplified Molecular Input Line Entry System*. SMILES is a linear notation for chemical structures. SMILES strings afford an easy way of entering certain types of structures in many programs. These strings are unique for each structure and suitable for database searching requirements. Correct strings should be capable of being interpreted by the modeling system and yield a 2D- or 3D-structure.[203,204] POLLY is a nickname for a tested and documented computer program which calculates 98 different types of topological indices for molecules containing up to 120 atoms composed of H, C, N, O, F, P, S, Cl, Br, or I atoms. Among these indices, we can find, e.g., the Wiener index, the molecular connectivity and valence connectivity indices, the bonding connectivity indices, and the information theoretic indices. It accepts a molecular description in the form of SMILES notation but is easily adapted to other user-oriented forms of input. The output consists of values of the 98 indices for each molecule including molecular connectivity indices, information theoretic indices, and indices of neighborhood symmetry.

9.2.2. CODESSA

This is the acronym for *Comprehensive Descriptors for Structural and Statistical Analysis*. It is a program designed by Katrizky, A. R.; Karelson, V.; Lobanov, A. R. University of Florida, Gainsville, FL, for use in the study of structure versus property or biological activity, and it calculates some 400 molecular descriptors which were designed before 1995. Half of these descriptors represent various topological indices.[205] This package has recently been updated and is marketed as CODESSA PRO (www.codessa-pro.com or katrizky@chem.ufl.edu), and it is furnished with an advanced variable selection procedure and with an even larger pool of theoretical descriptors. The descriptors are calculated solely from molecular structures: directly from the molecular formula in the case of constitutional and topological descriptors and utilizing the molecular 3D geometry for geometrical descriptors. CODESSA PRO is implemented with a semiempirical quantum-chemical program CMOPAC, which is based on MOPAC version 7 (Stewart, J.J.P. MOPAC Program Package, QCPE 1989, No. 455). Structures in CODESSA PRO are used to calculate constitutional, topological, geometrical, thermodynamic, quantum chemical, and electrostatic descriptors. MOPAC, named from an acronym for *Molecular Orbital Package*, is a general-purpose semiempirical quantum mechanics package for the study of chemical properties and reactions in gas, solution, and solid state.

9.2.3. CLUJTOPO

A new version of Clujtopo (2.0) is now ready and has a job function which enables one to run over all molecules within the work directory.[206] It includes a routine for finding molecular similarities, proteins included. This program is designed to calculate descriptors from topological matrices and/or polynomials. Several weighting schemes including group electronegativity, group mass, and partial charges are proposed. Topological indices derived from matrices such as adjacency, distance, detour, distance-path, detour-path, Cluj, their reciprocal matrices, walk-matrices, layer-matrices, and shell-matrices are obtained with this package. It starts from the Hyper-Chem figure of the molecule and derives from there the vast array of matrices. The package has been developed at the Faculty of Chemistry and Chemical Engineering of the University of Cluj, Romania, by Diudea, M. (diudea@chem.ubbcluj.ro), Ursu, O., and Levente, C. The name is a contraction of *Cluj* and *topology*.

9.2.4. DRAGON

Dragon is a regularly updated application[16] for the calculation of molecular descriptors. The new 2003 version DRAGON 4 calculates 1612 molecular descriptors divided into 20 logical blocks, among which are topological descriptors, constitutional descriptors, walk and path counts, connectivity indices, information indices, edge adjacency indices, topological charge indices, eigenvalue-based indices, Randić molecular profiles, geometrical descriptors, charge descriptors, etc. Principal components (PCs) can also be calculated on the molecular descriptor blocks, and any user-defined descriptor/response file can be added to the calculated descriptor. The program includes interactive graph menus as well as histogram graphs and univariate statistics. Line plots and 2D- and 3D-scatter-plots are also available, allowing a preliminary analysis of molecule distribution in the descriptor or PC space, as well as a preliminary correlation analysis when user-defined responses have already been loaded. The name Dragon is taken from the Ishtar gate of Babylon at the Pergamon Museum in Berlin. The main figure of this gate is a dragon. This software is distributed by Talete srl, a private company which provides software, consulting, seminars, and courses in QSAR and chemometrics. This program has been developed by the *Milano Chemometrics and QSAR research group*, headed by Todeschini, R., Department of Environmental Sciences, University of Milano-Bicocca (roberto.todeschini@unimib.it or www.disat.unimib.it/chm/Dragon4.htm). This research group, which edits the official *Bulletin* of the *International Academy of Mathematical Chemistry* (IAMC), also organizes a school on molecular descriptors and chemometrics, with an introduction to Dragon and other types of software. The *Milano Chemometrics and QSAR research group* regularly updates the Dragon book and the software as well as the published literature in mathematical chemistry (it seems that up to now it has collected more than 7000 references). This group has recently introduced a website (www.moleculardescriptors.eu/index.htm) dedicated to all those scientists working in the field of molecular descriptors. The head of the group, Roberto Todeschini, is the actual president of the IAMC.

9.2.5. MOLGEN-4.0

This is the latest version of a series of the MOLGEN (molecular Generator; MOLGEN 5.0 is on release) project that has now run for 17 years and which is widely used in industry and academia. The 1997 MOLGEN 3.5 version was awarded as the best scientific software for chemists. This program is devoted to the computation of all structural formulas (=connectivity isomers) that correspond to a given molecular formula and, optionally, satisfy additional conditions such as prescribed and forbidden substructures. It has several components: (i) a generator for chemical graphs, (ii) a generator for connectivity isomers, (iii) a graphical molecule editor and 2D-display, (iv) a display for 3D-placements using energy optimization, and (v) a generator for all configurational isomers. It is also available in an educational version. A version of MOLGEN has been developed for the simulation of combinatorial chemistry and the optimization of such experiments. A MOLGEN-CID (MOLGEN−*C*hemical *Id*entifier) program has recently been developed and is freely accessible for use via the Internet. It works on hydrogen-suppressed graphs and uses information on bond multiplicity. Thus, the output does not contain hydrogen atoms. The MOLGEN software family has been developed at the University of Bayreuth, Germany, by Kerber, A.; Laue, R.; and Ruckdeschel, A. (http://www.molgen.de/).

9.2.6. MOLCONN-Z

MOLCONN is the acronym for *Molecular Connectivity*. A modern version of Molconn-Z is marketed by eduSoft LC QSAR Software (www.edusoft-lc.com/toolkits or haney@hbond.com) and includes the largest extant set of Kier and Hall descriptors, originally developed by Kier and Hall 20 years ago and subsequently improved by them. It also includes E-State (i.e., atom- and group-type) descriptors, H-bond (i.e., donor and acceptor) descriptors, polarity descriptors, and 3D QSAR descriptor fields. The MOLCONN-Z software was designed to carry out the computation of a wide range of topological indices of molecular structure, such as the χ indices, the κ shape indices, the electrotopological state indices I_S and E_S, the hydrogen electrotopological state indices, topological equivalence indices, counts of graph paths, several information indices such as the Shannon and the Bonchev-Trinajstić information indices, and other indices. The authors, L.H. Hall and L.B. Kier (hall@enc.edu, kier@gems.vcu.edu) have recently authored a book[57] which is marketed with a CD for the computation of the I_S and E_S atom-based indices.

9.2.7. TOPS-MODE and MODesLab

These are products developed by Estrada and by Estrada et al., at the University of Santiago de Campostela, Spain.

Modeslab studio (2002, version 1.0 b) (Estrada66@yahoo.com; www.modeslab.com/, accessed March 2004)) is the marketing studio and provides all the necessary tools for performing QSAR studies, starting with the input of a large number of molecules and going on to the calculation of molecular descriptors, property prediction, and substructural analysis.[207−210] It also provides a way to define the properties of atoms, bonds, and fragments by an extension of the SMILES language, and to use them in molecular descriptor calculations. TOPS-MODE is an acronym for *Topological Substructural Molecular Design*. This method is based on the computation of the spectral moments of the bond adjacency matrix with the appropriate weights for each molecule in the data set. The bond adjacency matrix is a square symmetric matrix whose nondiagonal entries are one or zero if the corresponding bonds share an atom or not, respectively. The main diagonal entries are bond weights describing the hydrophobic/polarity, electronic, and steric features of compounds. The spectral moments are defined as the trace, i.e., the sum of the main diagonal entries, of the corresponding powers of the bond adjacency matrix. A table is generated in which the rows correspond to the compounds and the columns to the spectral moments. The program extracts QSPR/QSAR relationships by using linear or nonlinear multivariate statistical methods, and it tests the predictive QSPR models with cross-validation techniques. Further, it draws the hydrogen-depleted molecular graphs for each molecule of the data set. The bond weights are used in order to differentiate the molecular bonds, e.g., bond distances, bond dipoles, polarizabilities, and thus allow the derivation of meaningful information about the partition coefficient, polar surface area, polarizability, Gasteiger−Marsili atomic charges, van der Waals atomic radii, molar refraction, and Abraham molecular descriptors. Recently, there was an interesting debate at the Internet Electronic Conference of Molecular Design 2003 (www.biochempress..com/iecmd_2003.html, accessed January 2004) about the possibilities of TOPS-MODE and DRAGON 2.1 descriptors in QSAR.

9.2.8. The Variable Connectivity Index

This computer program was designed for efficient computation of the variable connectivity index mentioned in previous paragraphs. It was designed by Kezele, N. (nenad@joker.irb.hr); Klasinć, L.; von Knop, J.; Ivaniš, S.; and Nikolić, S. at the Rudjer Bošković Institute, Zagreb, Croatia.[70]

9.2.9. Toolkit

This is a program that accompanies the book *Handbook for Estimating Physicochemical Properties of Organic Compounds*.[29] The aim of this program is to provide the readers of the book with the opportunity to try some of the methods described in the book or to apply them to their own problems. In addition to this, Toolkit offers several features to help in determination of desired physicochemical properties. For example, Toolkit includes the Registry of Physicochemical Data, containing data for more than 24 000 compounds. Toolkit also includes powerful tools for generating three-dimensional models of compounds from SMILES codes imported from external editors or from your pictures drawn with external chemical editors such as CS ChemDraw, ISIS Draw, and ChemWindow.

9.3. Mathematical Chemistry Monographs

A mathematical chemistry monograph (MCM) is not a program but instead is a series of books on different topics in mathematical chemistry that are published in connection with the journal MATCH (*MATCH Communications in Mathematical and in Computer Chemistry*). Four numbers have already been published. MATCH is an international journal which contains original research papers on various applications of mathematics in chemistry, especially graph theory, but also including mathematical research inspired by chemical problems as well as the development of chemistry-related algorithms and computer software, and also the historical aspects thereof. MATCH was founded by Oskar E. Polansky in October 1975, and its actual editor is Ivan Gutman. The publisher of the MCM series is the University of Kragujevac and Faculty of Science Kragujevac. Those interested in the series should contact match@kg.ac.yu. The following four volumes have already been edited:

Vol. 1: Li, X.; Gutman, I. *Mathematical Aspects of Randić-Type Molecular Structure Descriptors*.

Vol. 2: Pogliani, L. *Numbers Zero, One, Two, and Three in Science and Humanities*.

Vol. 3: Janezić, D.; Milicević, A.; Nikolić, S.; Trinajstić, N. *Graph-Theoretical Matrices in Chemistry*.

Vol. 4: Fujita, S. *Diagrammatical Approach to Molecular Symmetry and Enumeration of Stereoisomers*.

10. Note Added in Proof

Recently, it has been published that a novel antineoplasic quinoline designed by molecular topology, MT477, is in preclinical trials (Jasinski, P.; Welsh, B.; Galvez, J.; Land, D.; Zwolak, P.; Ghandi, L.; Terai, K.; Dudeck, A. Z. *Invest. New Drugs* **2007**, Oct. 24).

11. References

(1) Pogliani, L. *Chem. Rev.* **2000**, *100*, 3827.
(2) Randić, M. *Chem. Rev.* **2003**, *103*, 3449.
(3) Harary, F. *Graph Theory*, 2nd printing; Addison-Wesley: Reading, MA, 1971. Buckley, F.; Harary, F. *Distance in Graphs*; Perseus Books: Jackson, TN, 1990. See also: Essam, J. W.; Fisher, M. E. *Rev. Mod. Phys.* **1970**, *42*, 272.
(4) Hartsfied, N.; Ringel, G. *Pearls in Graph Theory*; Academic Press: New York, 1990. Bondi, J. A.; Murty, U. S. R. *Graph Theory with Applictions*; North-Holland: New York, 1976 (on the web). Biggs, N. L.; Lloyd, E. K.; Wilson, R. J. *Graph Theory 1736−1936*; Oxford University Press: Oxford, 1986.
(5) Bonchev, D. G. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U.K., 1983.
(6) Balaban, A. T., Ed. *Chemical Applications of Graph Theory*; Academic Press: London, 1986.
(7) Sinanoglu, O. *J. Am. Chem. Soc.* **1975**, *97*, 2309; *J. Math. Phys.* **1981**, *22*, 1504; *Theor. Chim. Acta* **1984**, *65*, 267. See also: Turro, N. J. *Angew. Chem., Int. Ed. Engl.* **1986**, *25*, 882.
(8) Zefirov, N. S., Kuchanov, S. I., Eds. *Graph Theory Applications of Chemistry*; Nauka: Novosibirks, 1988 (in Russian).
(9) King, R. B., Rouvray, D. H., Eds. *Graph Theory and Topology in Chemistry*; Elsevier: Amsterdam, 1987.
(10) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(11) Devillers, J., Balaban, A. T., Eds. *Topological Indices and Related Descriptors in QSAR/QSPR*; Gordon and Breach: Amsterdam, 1999.
(12) Bonchev, D.; Rouvray, D. *Chemical Topology: Introduction and Fundamentals*; Taylor & Francis: London, 1999. *Chemical Topology: Applications and Techniques*; Taylor & Francis: London, 2000.
(13) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.
(14) Randić, M.; Trinajstić, N. *Croat. Chem. Acta* **1994**, *67*, 1.
(15) Diudea, M. V., Ed. *Studies by Molecular Descriptors*; Nova Science Pub.: New York, 2000.
(16) Todeschini, R.; Consonni, V. *The Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
(17) Pilar, F. L. *Elementary Quantum Chemistry*; Mcgraw-Hill: New York, 1968; Chapter 18.
(18) Estrada, E.; Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573.

(19) Pogliani, L. *J. Chem. Inf. Comput. Chem*. **2004**, *44*, 42.

(20) Harary, F.; Read, R. *Proc. Graphs and Combinatorics Conference*; George Washington University; Springer: New York, 1977. Cited in: Barrow, J. D. *The Book of Nothing*; Vintage Books: New York, 2000; p 155.

(21) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.

(22) Randić, M. Topological Indices. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. V. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; pp 3018.

(23) Rouvray, D. H. *Chemical Graph Theory*—*Introduction and Fundamentals*; Bonchev, D., Rouvray, D. H., Eds.; Gordon & Breach: New York, 1991.

(24) Rouvray, D. H. *Chem. Br.* **1977**, *13*, 52.

(25) Rouvray, D. H. *THEOCHEM* **1989**, *185*, 1.

(26) Temkin, O. N.; Zeigarnik, A. V.; Bonchev, D. *Chemical Reaction Network*; CRC Press: Boca Raton, FL, 1996. See also: Temkin, O. L.; Bonchev, D. *J. Chem. Educ*. **1992**, *69*, 544.

(27) Pogliani, L. *MATCH Commun. Math. Comput. Chem*. **2003**, *49*, 141.

(28) Nye, M. J. *A History of Modern Physics 1800*−*1954, Vol IV: The Question of the Atom*, 2nd ed.; Tomash Pub.: Los Angeles, San Francisco, 1986; p 167.

(29) Reinhard, M.; Drefahl, A. *Handbook for Estimating Physicochemical Properties of Organic Compounds*; Wiley: New York, 1999.

(30) Hansch, C.; Leo, A. *Exploring QSAR*—*Fundamentals and Applications in Chemistry and Biology*; ACS Professional Reference Book; American Chemical Society: Washington, DC, 1995.

(31) Pogliani, L.; de Julián-Ortiz, J. V. *Chem. Phys. Lett.* **2004**, *393*, 327.

(32) Besalu, E.; de Julián-Ortiz, J. V.; Iglesias, M.; Pogliani, L. *J. Math. Chem.* **2006**, *39*, 475.

(33) Besalu, E.; de Julián-Ortiz, J. V.; Pogliani, L. *J. Chem. Inf. Model.* **2007**, *47*, 751.

(34) Carbó-Dorca, R.; Robert, D.; Amat, Ll.; Girones, X.; Besalu, E. *Molecular Quantum Similarity in QSAR and Drug Design*; Springer: Berlin, 2000.

(35) Randić, M. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 311.

(36) Randić, M. *New J. Chem.* **1991**, *15*, 517.

(37) Randić, M. *J. Comput. Chem.* **1991**, *12*, 970.

(38) Randić, M. *THEOCHEM* **1991**, *233*, 45.

(39) Randić, M. *Int. J. Quant. Chem. Quant. Biol. Symp.* **1994**, *21*, 215.

(40) Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell*. **2002**, *20*, 269.

(41) Hawkins, D. M. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 1.

(42) Peterangelo, C.; Seybold, P. G. *Int. J. Quant. Chem.* **2004**, *96*, 1.

(43) Pecka, J.; Ponec, R. *J. Math. Chem.* **2000**, *27*, 13.

(44) Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579.

(45) Golbraikh, A.; Shen, M.; Xiao, Z.; Xioa, Y.; Lee, K.-H.; Tropsha, A. *J. Comput.-Aid. Mol. Des.* **2003**, *17*, 241.

(46) Spiegel, M. R. *Probability and Statistics*; McGraw-Hill: New York, 1975.

(47) Draper, N. R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, 1966.

(48) Anscombe, F. J. *Am. Stat.* **1973**, *27*, 17.

(49) Pogliani, L. *New J. Chem*. **2003**, *27*, 919.

(50) Pogliani, L. *J. Comput. Chem*. **2003**, *24*, 1097.

(51) Pogliani, L. *J. Comput. Meth. Sci. Eng.* **2004**, *4*, 737.

(52) Pogliani, L. In *Computational Aspects of Electric Polarizability Calculations: Atom, Molecules, and Clusters*; Maroulis, G., Ed.; IOS Press: Amsterdam, 2005; Part II, pp 737−751.

(53) Pogliani, L. *J. Pharm. Sci.* **2007**, *96*, 1856.

(54) Pogliani, L. *New J. Chem.* **2005**, *29*, 1082−1088.

(55) Randić, M. *Chemom. Intell. Lab. Syst*. **1990**, *10*, 213.

(56) Chang, R. *Chemistry*, 5th ed.; McGraw-Hill: New York, 1994.

(57) Kier, L. B.; Hall, L. H. *Molecular Structure Description*. *The Electrotopological State*; Academic Press: New York, 1999.

(58) Yang, C.; Zhong, C. *Chem. Inf. Comput. Sci.* **2003**, *43*, 1998.

(59) Pogliani, L. *J. Comput. Chem.* **2006**, *27*, 868.

(60) Pogliani, L. *J. Phys. Chem. A* **2000**, *104*, 9029.

(61) Pogliani, L. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 836.

(62) Pogliani, L. *THEOCHEM* **2002**, *581*, 87.

(63) Pogliani, L. In *Topology in Chemistry*; Rouvray, D. H., Bruce King, R., Eds.; Horwood: Chichester, U.K., 2002; p 208.

(64) Randić, M. *Chemom. Intell. Lab. Syst*. **1990**, *10*, 213.

(65) Randić, M.; Dobrowolski, J. *Int. J. Quantum. Chem*. **1998**, *70*, 1209.

(66) Randić, M. *New J. Chem.* **2000**, *24*, 165.

(67) Randić, M.; Pompe, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575 and 631.

(68) Randić, M.; Plavšić, D.; Lerš, N. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 657.

(69) Randić, M.; Basak, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614 and 650.

(70) Kezele, N.; Klasinc, L.; von Knop, J.; Ivaniš, S.; Nikolić, N. *Croat. Chem. Acta* **2002**, *75*, 651.

(71) Yang, C.; Zhong, C. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1998.

(72) Sharma, V.; Goswami, R.; Madan, A. K. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 273.

(73) Ren, B. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 139.

(74) Ren, B. *J. Comput.-Aided Mol. Des*. **2003**, *17*, 607.

(75) Schmid, R. *J. Chem. Educ.* **2003**, *80*, 931.

(76) Pogliani, L. *Int. J. Quantum. Chem.* **2005**, *102*, 38.

(77) Ma, B.; Lii, J.-H.; Allinger, N. L. *J. Comput. Chem*. **2000**, *21*, 813.

(78) Cargas, M. L.; Seybold, P. G.; Andersen, M. E. *Toxicol. Lett*. **1988**, *43*, 235.

(79) Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. A. *SAR QSAR Environ. Res*. **2002**, *13*, 649 (and references therein).

(80) Seybold, P. G. *SAR QSAR Environ. Res.* **1999**, *10*, 101.

(81) Seybold, P. G. *Adv. Quant. Struct. Prop. Rel*. **2002**, *3*, 109.

(82) Mihalić, Z.; Nikolić, S.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 28.

(83) Pogliani, L. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 801.

(84) Matta, C. F.; Gillespie, R. J. *J. Chem. Educ.* **2002**, *79*, 1141.

(85) Randić, M. *Croat. Chim. Acta* **1991**, *64*, 43−54.

(86) Randić, M. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 311.

(87) Randić, M. *THEOCHEM* **1991**, *233*, 45.

(88) Pogliani, L. *Internet Electron. J. Mol. Des.* **2006**, *5*, 364.

(89) Galvez, J.; Garcia-Domenech, R. *Farmaindustria* **1994**, 357.

(90) Gálvez, J.; García-Domenech, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520.

(91) Gálvez, J.; García-Domenech, R.; Julián-Ortiz, J. V. de; Soler, R. *J. Chem. Inf. Comp. Sci.* **1995**, *35*, 272.

(92) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.

(93) Wiener, H. *J. Chem. Phys*. **1947**, *15*, 766.

(94) Wiener, H. *J. Phys. Chem.* **1948**, *52*, 425.

(95) Wiener, H. *J. Phys. Chem.* **1948**, *52*, 1082.

(96) Kier, L. B.; Hall, L. M. *Pharm. Res.* **1989**, *6*, 497.

(97) Kier, L. B. *Quant. Struct.*−*Act. Relat.* **1985**, *4*, 109.

(98) DESMOL11 software; Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Facultad de Farmacia, Universitat de Valencia, Spain.

(99) MOLCONN-Z software, version 3.50; L. H. Hall, Eastern Nazarene College, Quincy, MA.

(100) Julián-Ortiz, J. V.; de Gregorio Alapont, C.; de Ríos-Santamarina, I.; García-Domenech, R.; Gálvez, J. *J. Mol. Graphics Modell.* **1998**, *16*, 14.

(101) Julián-Ortiz, J. V.; de Besalú, E.; García-Domenech, R. *Ind. J. Chem.* **2003**, *42A*, 1392.

(102) Furnival, G. M.; Wilson, R. W. *Technometrics* **1974**, *16*, 499.

(103) Hocking, R. R. *Technometrics* **1972**, *14*, 967.

(104) Allen, D. M. *Technometrics* **1974**, *16*, 125.

(105) Duart, M. J.; Antón-Fos, G. M.; de Julián-Ortiz, J. V.; Gozalbes, R.; Gálvez, J.; García-Domenech, R. *Int. J. Pharm.* **2002**, *246*, 111.

(106) Besalú, E. *J. Math. Chem.* **2001**, *29*, 191.

(107) Wold, S.; Eriksson, L. *Statistical validation of QSAR results*. In *Chemometric methods in molecular design*; Van de Waterbeemd, H., Ed.; VCH: New York, Vol. 2, 1995; pp 309−318.

(108) Dixon, W. J.; Brown, M. B.; Engelman, L.; Jennrich, R. I. *BMDP Statistical Software Manual*; University of California Press: San Francisco, 1990, Vol. 1.

(109) Lachenbruch, P. A.; Mickey, R. M. *Technometrics* **1968**, *10*, 1.

(110) Gálvez, J.; García-Domenech, R.; Gregorio-Alapont, C.; de Julián-Ortiz, J. V.; Popa, L. J. *J. Mol. Graphics* **1996**, *14*, 272.

(111) De Julian-Ortiz, J. V.; Galvez, J.; Muñoz-Collado, C.; Garcia-Domenech, R.; Jimeno-Cardona, C. *J. Med. Chem.* **1999**, *42*, 3308.

(112) Bruno-Blanch, L.; Gálvez, J.; García-Domenech, R. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2749.

(113) **113a**: Rios-Santamarina, I.; García-Domenech, R.; Cortijo, J.; Santamaria, P.; Morcillo, E. J.; Gálvez, J. *Internet Electron. J. Mol. Des.* **2002**, *1*, 70. **113b**: Ríos Santamarina, I.; García-Domenech, R.; Gálvez, J.; Santamaría, P.; Cortijo, J.; Morcillo, E. J. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 477.

(114) **114a**: Arviza, M. P. Predicción e interpretación de algunas propiedades fisicoquímicas y biológicas de un grupo de barbitúricos y sulfonamidas por el método de conectividad molecular; Universidad de Valencia: 1985. **114b**: Bernal, J. Desarrollo de un nuevo método de diseño molecular asistido por ordenador. Su aplicación a fármacos betabloqueantes y benzodiazepinas; Universidad de Valencia: 1988.

(115) Gálvez, J.; García-Domenech, R.; Bernal, J. M.; García-March, F. J. *An. Real Acad. Farm.* **1991**, *57*, 533.

(116) García-Domenech, R.; García-March, F. J.; Soler, R. M.; Gálvez, J.; Antón-Fos, G. M.; de Julián-Ortiz, J. V. *Quant. Struct.*−*Act. Relat.* **1996**, *15*, 201.

(117) Gálvez, J.; García-Domenech, R.; de Julián-Ortiz, J. V.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1198.

(118) García-Domenech, R.; Villanueva, A.; Gálvez, J.; Gozalbes, R. *J. Chim. Phys.* **1999**, *96*, 1172.

(119) Gozalbes, R.; de Julián-Ortiz, J. V.; Antón-Fos, G. M.; Gálvez, J.; García-Domenech, R. *Chromatographia* **2000**, *51*, 331.

(120) García-Domenech, R.; Muñoz-Espí, R.; Roda-Fenollosa, G.; Villanueva-Montesinos, A.; Gálvez, J. *Afinidad* **2003**, *60* (504), 161.

(121) García-Domenech, R.; de Julián-Ortiz, J. V. *J. Phys. Chem. B* **2002**, *106*, 1501.

(122) Lahuerta Zamora, L.; Fuster Mestre, Y.; Duart, M. J.; Antón Fos, G. M.; García-Domenech, R.; Gálvez Alvarez, J.; Martínez Calatayud, J. *Anal. Chem.* **2001**, *73*, 4301.

(123) Murcia, M.; García-Domenech, R.; Castillo, M. E.; Font, R.; Porcar, M.; Simón, V. E.; Gálvez, J. *Afinidad* **2000**, *LVII-489*, 337.

(124) Duart, M. J.; Garcia-Domenech, R.; Anton-Fos, G. M.; Galvez, J. *J. Comput.-Aided Mol. Des*. **2001**, *15*, 561.

(125) García-Domenech, R.; Catalá, A. I.; García-García, A.; Soriano, A.; Pérez-Mondejar, V.; Gálvez, J. *Ind. J. Chem.* **2002**, *41B*, 2376.

(126) García-Domenech, R.; de Julián-Ortiz, J. V.; Duart, M. J.; García-Torrecillas, J. M.; Antón-Fos, G. M.; Ríos-Santamarina, I.; de Gregorio-Alapont, C.; Gálvez, J. *SAR QSAR Environ. Res.* **2001**, *12*, 237.

(127) Gozalbes, R.; Brun-Pascaud, M.; García-Domenech, R.; Gálvez, J.; Girard, P. M.; Doucet, J. P.; Derouin, F. *Antimicrob. Agents Chemother*. **2000**, *44* (10), 2771.

(128) Gálvez, J.; García-Domenech, R.; Gómez-Lechón, M. J.; Castell, J. V. *THEOCHEM* **2000**, *504*, 241.

(129) Gozalbes, R.; Brun-Pascaud, M.; García-Domenech, R.; Gálvez, J.; Girard, P. M.; Doucet, J. P.; Derouin, F. *Antimicrob. Agents Chemother*. **2000**, *44* (10), 2764.

(130) de Gregorio-Alapont, C.; García-Domenech, R.; Gálvez, J.; Ros, M. J.; Wolski, S.; García, M. D. *Bioorg. Med. Chem. Lett*. **2000**, *10*, 2033.

(131) Pastor, L.; García-Domenech, R.; Gregorio Alapont, C.; de Gálvez, *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2577.

(132) Mahmoudi, N.; de Julian-Ortiz, J. V.; Ciceron, L.; Galvez, J.; Mazier, D.; Danis, D.; Derouin, F.; Garcia-Domenech, R. *J. Antimicrob. Chemother.* **2006**, *57*, 489.

(133) Gozalbes, R.; Gálvez, J.; García-Domenech, R.; Derouin, F. *SAR QSAR Environ. Res.* **1999**, *10*, 47.

(134) Gálvez, J.; García-Domenech, R.; Gregorio Alapont, C.; de Julián-Ortiz, J. V.; de Salabert-Salvador, M. T.; Soler-Roca, R. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: London, 1996; Vol. I, pp 267.

(135) Antón-Fos, G. M.; García-Domenech, R.; Pérez-Giménez, F.; Peris-Ribera, J. E.; García-March, F.; Salabert-Salvador, M. T. *Arzneim.-Forsch./Drug Res*. **1994**, *44*, 821.

(136) Gálvez, J.; Gómez-Lechón, M. J.; García-Domenech, R.; Castell, J. V. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 2301.

(137) Casabán-Ros, E.; Antón-Fos, G. M.; Gálvez, J.; Duart, M. J.; García-Domenech, R. *Quant. Struct.*−*Act. Relat.* **1999**, *18*, 35.

(138) Christiansen, J. A. In *Advances in Catalysis*; Frankenburg, W. G., Komarewsky, V. I., Rideal, E. K., Eds.; Academic Press: New York, 1953; Vol. 5, p 311.

(139) King, E. L.; Altman, C. A. *J. Phys. Chem.* **1956**, *60*, 1375.

(140) Temkin, M. I. *Dokl. Akad. Nauk SSSR* **1965**, *165*, 615.

(141) Balaban, A. T.; Fărcaşiu, D.; Bănică, R. *Rev. Roum. Chim.* **1966**, *11*, 1205. Balaban, A. T. *Rev. Roum. Chim.* **1967**, *12*, 875.

(142) Balaban, A. T. In *Graph Theoretical Approaches to Chemical Reactivity*; Bonchev, D., Mekenyan, O., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1994; pp 137−180.

(143) Randić, M.; Razinger, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 140.

(144) Randić, M. *J. Math. Chem.* **1996**, *19*, 375.

(145) Randić, M.; Krilov, G. *Chem. Phys. Lett.* **1997**, *272*, 115.

(146) Li, H.; Helling, R.; Tang, C.; Wingreen, N. *Science* **1996**, *273*, 666.

(147) Krilov, G.; Randić, M. *New J. Chem.* **2004**, *28*, 1608.

(148) Randić, M.; Zupan, J.; Balaban, A. T. *Chem. Phys. Lett.* **2004**, *397*, 247.

(149) Estrada, E. *Chem. Phys. Lett.* **2000**, *319*, 713.

(150) Estrada, E. *Bioinformatics* **2002**, *18*, 697.

(151) Estrada, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1238.

(152) Estrada, E. *Comput. Biol. Chem.* **2003**, *27*, 305. Estrada, E.; Uriarte, E. *Comput. Biol. Chem.* **2005**, *29*, 345.

(153) Estrada, E. *Proteins*: *Struct., Funct., Bioinf.* **2004**, *54*, 727.

(154) Randić, M.; Witzmann, F.; Vracko, M.; Basak, S. C. *Med. Chem. Res.* **2001**, *10*, 456.

(155) Randić, M. *Int. J. Quantum Chem.* **2002**, *90*, 848.

(156) Randić, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 983.

(157) Bajzer, Z.; Randić, M.; Plavšić, D.; Basak, S. C. *J. Mol. Graphics Modell.* **2003**, *22*, 1.

(158) Vračko, M.; Basak, S. C. *Chemometrics Intell. Lab. Syst.* **2004**, *70*, 33.

(159) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C. *J. Proteome Res.* **2004**, *3*, 778.

(160) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C. *Croat. Chem. Acta* **2004**, *77*, 345.

(161) Randić, M.; Witzmann, F. A.; Kodali, V.; Basak, S. C. *J. Chem. Inf. Model.* **2006**, *46*, 166.

(162) Randić, M.; Lerš, N.; Vukičević, D.; Plavšić, D.; Gute, B. D.; Basak, S. C. *J. Proteome Res.* **2005**, *4*, 347.

(163) Balasubramanian, K.; Khokhani, K.; Basak, S. C. *J. Proteome Res.* **2006**, *5*, 1133.

(164) Randić, M.; Basak, S. C. *Med. Chem. Res.* **2004**, *13*, 800.

(165) Hawkins, D. M.; Basak, S. C.; Kraker, J.; Geiss, K. T.; Witzmann, F. A. *J. Chem. Inf. Model.* **2006**, *46*, 9.

(166) Vračko, M.; Basak, S. C.; Witzmann, F. A. *J. Chem. Inf. Model.* **2006**, *46*, 130.

(167) Randić, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50.

(168) Randić, M.; Vračko, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599.

(169) Randić, M. *Chem. Phys. Lett.* **2000**, *317*, 29.

(170) Randić, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 561.

(171) Randić, M.; Guo, X.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619.

(172) Guo, X.; Randić, M.; Basak, S. C. *Chem. Phys. Lett.* **2001**, *350*, 106.

(173) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. *Chem. Phys. Lett.* **2003**, *368*, 1.

(174) Randić, M.; Vračko, M.; Zupan, J.; Novič, M. *Chem. Phys. Lett.* **2003**, *373*, 558.

(175) Randić, M. *Chem. Phys. Lett.* **2004**, *386*, 468.

(176) Balaban, A. T.; Plavsić, D.; Randić, M. *Chem. Phys. Lett.* **2003**, *379*, 147.

(177) Randić, M.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532.

(178) Zupan, J.; Randić, M. *J. Chem. Inf. Model.* **2005**, *45*, 309.

(179) Randić, M.; Lers, N.; Plavsić, D.; Basak, S. C.; Balaban, A. T. *Chem. Phys. Lett.* **2005**, *407*, 205.

(180) Randić, M.; Zupan, J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550.

(181) Galvez, J. *THEOCHEM* **1998**, *429*, 255.

(182) Galvez, J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1231.

(183) Kier, L. B.; Hall, L. H. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 792.

(184) Estrada, E. *J. Phys. Chem. A* **2002**, *106*, 9085.

(185) Giordano, N. J. *Computational Physics*; Prentice Hall: Upper Saddle River, New Jersey, 1977; 1st page.

(186) Duart, M. J.; Anton-Fos, G. M.; Aleman, P. A.; Gay-Roig, J. B.; Gonzales Rosende, M. E.; Galvez, J.; Garcia-Domenech, R. *J. Med. Chem.* **2005**, *48*, 1260.

(187) Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Muñoz, C.; Guna, R.; Borras, R. *J. Biomol. Screening* **2005**, *10*, 206.

(188) De Julian-Ortiz, J. V.; Garcia-Domenech, R.; Galvez, J.; Pogliani, L. *SAR QSAR Environ. Res.* **2005**, *15*, 2643.

(189) Llacer, M. T.; Galvez, J.; Garcia-Domenech, R.; Gomez-Lechon, M. J.; Mas Arcas, C.; de Julian-Ortiz, J. V. *Internet Electron. J. Mol. Des.* **2006**, *5*, 306.

(190) Duart, M. J.; Garcia-Domenech, R.; Galvez, J.; Aleman, P. A.; Martin-Algarra, R. V.; Anton-Fos, G. M. *J. Med. Chem.* **2006**, *49*, 3667.

(191) Randić, M. In *Handbook of Proteomics Methods*; Michel Conn., P., Ed.; Humana Press, Inc.: Towota, NJ, 2003; pp 429−450.

(192) Randić, M.; Estrada, E. *J. Proteome Res.* **2005**, *4*, 2133.

(193) Randić, M.; Zupan, J.; Vikić-Topić, D.; Plavšić, D. *Chem. Phys. Lett*. **2006**, *431*, 375.

(194) Kvasnicka, V.; Pospichal, J. *Int. J. Quantum Chem* **1990**, *38*, 253.

(195) Kvasnicka, V.; Pospichal, J. *Theor. Chim. Acta* **1991**, *79*, 65.

(196) Koca, J.; Kratichvil, M.; Kvasnicka, V.; Matyska, L.; Pospichal, J. *Lecture Notes in Chemistry*; Springer-Verlag: Heidelberg, 1989; Vol. 51, pp 1−207.

(197) Estrada, E.; Rodriguez-Velazques, J. A. *Phys. Rev. E* **2005**, *71*, 056103.

(198) Estrada, E.; Uriarte, E.; Vilar, S. *J. Proteome Res.* **2006**, *5*, 105.

(199) Estrada, E. *J. Phys. Chem. A* **2004**, *108*, 5468.

(200) Matamala, A.; Estrada, E. *J. Phys. Chem. A* **2005**, *109*, 9890.

(201) Matamala, A.; Estrada, E. *Chem. Phys. Lett.* **2005**, *410*, 343.

(202) Estrada, E.; Matamala, A. *J. Chem. Inf. Model.* **2007**, *47*, 794.

(203) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.

(204) Niemi, G. J.; Basak, S. C.; Veith, G. D.; Grunwald, G. *Environ. Toxicol. Chem*. **1992**, *11*, 893.

(205) Katrizky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794 and 1806.

(206) Diudea, M. V.; Ursu, O. *Ind. J. Chem*. **2003**, *41A*, 1283.

(207) Estrada, E. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844.

(208) Estrada, E. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 320.

(209) Estrada, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23.

(210) Estrada, E.; Gonzáles, U. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 75.

CR0780006