On Characterization of Chemical Structure

Milan Randić

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311

Received December 31, 1996[⊗]

We briefly review characterization of chemical structure as evolving from the early work on the connectivity index to the latest work on a characterization of 3-D structures, including characterization of the folding of model proteins.

"It would certainly be a serious illusion on my part if I hoped that my remarks have carried general conviction, or even that they have been generally understood. Surely much more will be thought and written concerning these questions, for theorists are numerous and paper is patient." Max Planck[1]

INTRODUCTION

One of the central topics of chemistry is the structure−property relationship: How is molecular structure reflected in the diversity of molecular properties? In Figure 1 we listed isomers of C_8H_{18} and a single property, ΔH_v, the heat of vaporization. The first question is the following: Is there any regularity in the data of Figure 1? Providing there is regularity in the data, the next question is the following: Can we quantify the relationship between the structure and the property? Here we tacitly assumed that the selected physicochemical property is a reflection on individual molecules. This will be true for structure−activity studies where one is considering the interactions of a *single* molecule with a *single* receptor on a large molecule. However, most measured physicochemical properties are bulk properties and assumption that characterization of individual structures will parallel collective characterization of bulk is open for scrutiny.

REGULARITY IN DATA

Already 50 years ago Platt[2] has pointed to paths in molecules as potential molecular descriptors for structure−property studies. The contribution of Platt has been overlooked, and it would have been forgotten was it not for the development of chemical graph theory in the mid 1970s.[3−5] Since isomers have the same number of carbon atoms and the same number of CC bonds in searching for a structure−property relationship, we have to go beyond the count of atoms and bonds in order to arrive at a molecular characterization that will discriminate isomers. Consider paths of length two and paths of length three, i.e., the count of consecutive C−C−C bonds and consecutive C−C−C−C bonds, respectively. To each isomer now one can assign a pair of numbers, p_2 and p_3. The pair (p_2, p_3) can be viewed as Cartesian coordinates of a structure. Figure 2 illustrates the (p_2, p_3) space with isomers points on the p_2, p_3 grid.[6] If we now substitute the numerical values for selected property at the site of each isomer, we can immediately see a regular change in the ΔHv along p_2 and p_3 axis (Figure 3). This

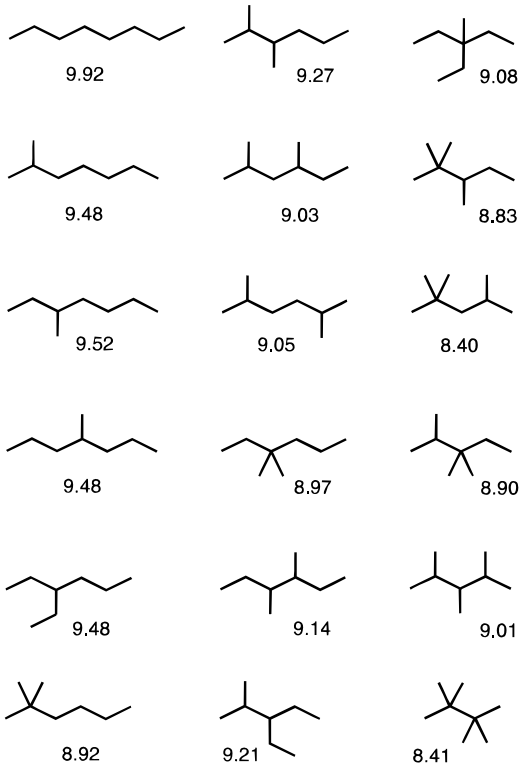


Figure 1. Carbon skeletons of the octane isomers and their heats of vaporization (in kcal).

remarkable parallelism between a partial ordering of structures (isomers) and properties clearly reflects the prophecy of Platt, who saw paths as promising molecular descriptors.

To quantify the relationship between a property and structure we need molecular descriptors. Sometimes a single descriptor may suffice; however, more often several descriptors are needed to get a useful structure−property regression. Hosoya's Z topological index was the first nontrivial single molecular descriptor suggested for expressing structure−property relationship.[7] A characterization of a molecule by a single mathematical descriptor is limited. It is not surprising to find that several structures have the same numerical magnitude for a descriptor. One speaks then of a *degeneracy* of an index. The size of the smallest graphs showing degeneracy is an indirect measure of the limitations of the particular descriptor to characterize diverse graphs. The degeneracy of several topological indices, including some common descriptors, is shown in Table 1. That different structures may have identical invariant is not surprising. What is surprising in fact is that so much

[⊗] Abstract published in *Advance ACS Abstracts,* June 1, 1997.

S0095-2338(96)00174-6 CCC: $14.00

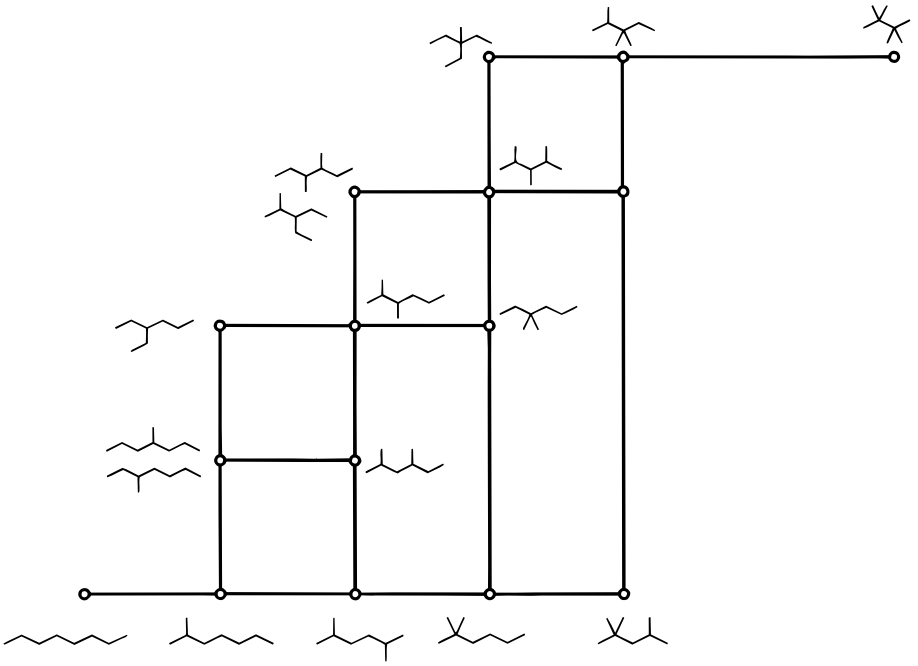


Figure 2. The (p_2, p_3) space with the octane isomers assignment.

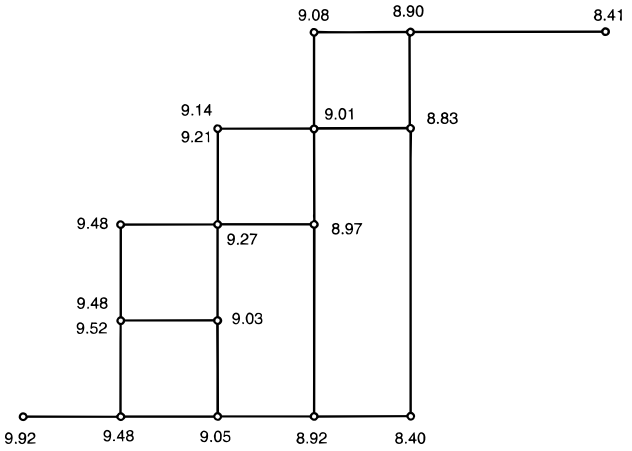


Figure 3. The regularity in isomeric variations of ΔH_v along p_2 and p_3 coordinate axes.

variations of diverse structural features in molecules can be condensed into a *single* mathematical parameter.

THE CONNECTIVITY INDEX

"But one must not imagine that advance is possible, even in the most exact of all natural sciences, without some view of the world, i.e., quite without some hypotheses not capable of proof." Max Planck[1]

In 1975 I proposed a mathematical descriptor χ, initially called the branching index,[8] that has been designed to parallel the boiling points of alkanes. Kier and Hall were not only the first to recognize the merits of the new descriptor but demonstrated its use on a wide class of compounds and properties.[9,10] Soon χ and the higher order connectivity indices received wide attention (for earlier bibliography see ref 10). The connectivity index is bond additive, the contributing terms depend on the type of the CC bond in a molecule. Bonds that are terminal, such as CC bond in CCH_3, make a greater contribution to those molecular properties that depend on molecular surface than the "buried" bonds in the molecular interior. Bond (m, n), having m and n nearest carbon atom neighbors, was assigned the weight $1/\sqrt{mn}$, since such an assignment induces the ordering of isomers of hexane and heptane that parallel the relative magnitudes for their respective boiling points.

HIGHER ORDER CONNECTIVITY INDICES

In order to obtain a satisfactory correlation many structure−property applications require use of two and more molecular descriptors. Already in 1947 Wiener used two mathematical descriptors, W and P in his notation, to obtain satisfactory regression of boiling points in paraffins.[11] Both these descriptors are related to the count of paths and weighted paths, P being the number of paths of length three, i.e., p_3, while W is the sum of paths of length L multiplied by L, with $L = 1, 2, ..., n$. Nevertheless, *ad hoc* selection of W and P does not represent a sufficiently broad basis that can be easily modified by inclusion of additional descriptors when required. The connectivity index χ allows a natural generalization and leads to a set of structurally related descriptors. Instead of bonds, which can formally be viewed as paths of length one, we now consider paths of length two. Paths of length two can be classified according to the valences of the vertices forming the paths. Here the word *valence* is used in the mathematical rather than chemical meaning, being synonymous to *vertex degree*. If we use the symbol d for valence of the vertex, then a path of length two involving vertices with the valences d_i, d_j, d_k will make the contribution $1/\sqrt{d_i d_j d_k}$. By summing the contributions from all paths of length two we obtain ${}^2\chi$.[12] Longer paths make analogous contributions to still higher order connectivity indices. In this way one arrives at a set of descriptors $\{{}^1\chi, {}^2\chi, {}^3\chi, {}^4\chi, ...\}$ that are structurally related and which offer a broader basis for characterization of molecules. Kier and Hall have further extended the above set of higher order connectivity indices by including ${}^0\chi$, which relates to individual atoms, and the descriptors $k\chi_C$ and $k\chi_{PC}$, which correspond to the so called cluster and path-cluster fragments, respectively.[9]

Table 1

topological index	real/integer	symbol	n	N_{n-1}	ref
Wiener index	I	W	7	13	*a*
leading eigenvalue	R	λ_1	7		*b*
Hosoya index	I	Z			*c*
Schultz index	I	MTI	8	22	*d*
connectivity index	R	χ	8	22	*e*
extended Wiener	I	W	9	40	*f*
hyper-Wiener number	I	WW	9	40	*g*
Balaban index	R	J	12	309	*h*
identification number	R	ID	15	3324	*i*
self-returning walk ID	R	SID	18	42 924	*j*
delta-connectivity paths	R	τ	19	103 447	*k*
prime number ID	R	ID_P	20	251 731	*l*
distance-based ID	R	BID	20>	618 050	*m*
weighted ID	R	WID	20>	618 050	*n*
weighted paths	R		20>	618 050	*o*
extended adjacency	R	EAID	22>	3 807 434	*j*

[a] References give the paper in which the degeneracy of the index is discussed. For introductory material see latter part of ref a. [a] Razinger, M.; Chretien, J. R.; Dubois, J. E. Structural selectivity of topological indexes in alkane series. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23−27. [b] Lovasz, L.; Pelikan, J. On the eigenvalue of trees. *J. Period. Math. Hung.* **1973**, *3*, 175−182. [c] Hosoya, H. Topological index. A newly proposed quantity characterizing topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339. [d] Muller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstic, N. Molecular topological index. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 160−163. [e] Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615. [f] Tratch, S. S.; Stankevich, M. I.; Zefirov, N. S. Combinatorial models and algorithms in chemistry. The expanded Wiener number−A novel topological index. *J. Comput. Chem.* **1990**, *11*, 899−908. [g] Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H. Wiener matrix: Source of novel graph invariants. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 709−716. [h] Balaban, A. T.; Quintas, L. V. *Math. Chem. (MATCH)* **1983**, *14*, 213. [i] Szymanski, K.; Muller, W. R.; Knop, J. V.; Trinajstic, N. On Randić's molecular identification number. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 413−415. [j] Hu, C. Y.; Xu, L. On highly discriminating molecular topological index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 82−90. [k] Hu, C. Y.; Xu, L. On Hall and Kier's topological state and total topological index. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1251−1258. [l] Szymanski, K.; Muller, W. R.; Knop, J. V.; Trinajstic, N. Molecular ID numbers. *Croat. Chem. Acta* **1986**, *59*, 719−723. [m] Muller, W. R.; Szymanski, K.; Knop, J. V.; Mihalic, Z.; Trinajstic, N. The walk ID numbers revisited. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 231−233. [n] Szymanski, K.;Muller, W. R.; Knop, J. V.; Trinajstic, N. *Int. J. Quant. Chem: Quant. Chem. Symp.* **1984**, *11*, 137. [o] Hu, C. Y.; Xu, L. A new topological index for CIAC-^{13}C NMR information system. *Anal. Chem. Acta* In press.

HETEROATOMS

Another important generalization of the connectivity indices $\{^1\chi, ^2\chi, ^3\chi, ^4\chi, ...\}$ considered by Kier and Hall is an extension of the approach to heteroatoms. In order to differentiate oxygen and nitrogen from carbon atom Kier and Hall modified vertex degrees by introducing a parameter that counts valence electrons.[9] Thus derived valence connectivity indices $\{^1\chi_m, ^2\chi_m, ^3\chi_m, ^4\chi_m, ...\}$ show variations with the change of atoms and offer a larger set of descriptors to be used when one considers heteroatomic molecules.

Do the valence connectivity indices represent an optimal solution to the characterization of heteroatomic molecules? They do not. However, they point to a need for a larger number of molecular descriptors when heteroatoms are present in a molecule. Finding optimal descriptors is a difficult task that has hardly been initiated.[13−18] The connectivity index can be viewed as derived from the row sums of the adjacency matrix. The row sums give for every atom in molecular graph the vertex degrees, i.e., d_i values. If we introduce a parameter x_C for carbon atom and x_O for oxygen atom as the diagonal entries of the adjacency matrix, then the adjacency matrix for ethyl alcohol (hydrogens being suppressed) becomes

$$\begin{pmatrix} x_C & 1 & 0 \\ 1 & x_C & 1 \\ 0 & 1 & x_O \end{pmatrix}$$

The new row sums are $\Sigma d_i + x_C$ and $\Sigma d_i + x_O$ for the carbon and oxygen atoms, respectively. They define the row sums of corresponding atoms making the contributions to the connectivity index for C−C and C−O bonds, respectively. Search for optimal parameters x_C and x_O which produce the smallest standard error for the boiling points for a family of 35 alcohols using the simple regression gives $x_C = 1.50$ and $x_O = -0.85$. With these parameters the standard deviation that has been 7.86 °C when oxygen and carbon were not differentiated was reduced by more than half, to the value 3.30 °C.

Are the parameters for heteroatoms transferable from one family (e.g., alcohols) to another family (e.g., ethers)? Are the parameters for heteroatoms transferable within a single family from one property (e.g., boiling points) to another property (e.g., heats of formation)? Is there some regularity in the empirical parameters for different heteroatoms? These are open questions that probably will be considered in the near future.

Optimal molecular descriptors are important from the statistical point of view. One would like to use as few descriptors as possible, preferably just a single descriptor. Use of fewer descriptors in multiple regression is reflected in the statistical quality of the regression analysis as measured by parameters like the correlation coefficients, the standard error, Fisher ratio, the probability errors for the coefficients of the regression equation, etc. For example, use of the traditional molecular descriptors of QSAR for a set of 18 clonidine-like imidazolidines, which included log P, and (log P)^2, pK, the hydrophobic constant π, the parachor, Taft's steric constant, π electron charge densities, HOMO (the highest occupied MO energies) in various combinations produced the best three-term expression with the correlation coefficient $r = 0.853$ and standard error $s = 0.544$. To obtain the correlation coefficient $r > 0.950$ (accompanied by the standard deviation smaller than 0.350) required five molecular descriptors.[19] This should be contrasted with the statistical parameters accompanying the regression based on three-term connectivity indices in which chlorine was assigned $x_{Cl} = -0.20$ (while x_C was kept zero). The regression gave $r = 0.977$ and $s = 0.222$.[20] This is considerably better than the five-term regression using traditional descriptors (combined with parameters derived from simple MO calculations). When compared with the best three-term regression of the traditional QSAR it shows a reduction of the standard error by half. Moreover, the value $x_{Cl} = -0.20$ is not the optimal heteroatom parameter, since no attempt was made to optimize x_C for carbon atom. The performance of the traditional descriptors here is dismal. Use of five descriptors on a set of only 18 structures in itself points to low quality of the descriptors employed regardless of the quality of the regression (which was rather limited).

The above illustrates that the *selection* of molecular descriptors in general is very important, if not critical. As

it is known, an exhaustive search for the best combination of *ad hoc* descriptors increases the risk of chance correlation.[21] However, in the case of the clonidine-like imidazolidines the connectivity indices produced a very good correlation coefficient and a small standard error that were not much altered in cross-validation ($r = 0.968$, $s = 0.248$). The result is particularly striking for this particular data set, because there are two extreme antihypertensive potency values that would be expected to give much trouble on cross-validation, but they did not.

One can arrive at a correlation of antihypertensive activities of the same clonidine-like compounds using only two graph theoretical descriptors. This reduction in the number of descriptors is a result of "compacting" information from several indices in fewer connectivity indices.[22]

One can argue, although not very convincingly, that the traditional approach may have failed because of questionable quality of computed quantum chemical parameters used. That argument is only diverting attention from the real question that needs to be addressed: Why are the graph theoretical descriptors so successful? Both questions, why the traditional descriptors have failed and why the topological indices produce good results, can be traced to the same root. Descriptors that fail apparently do not cover the critical part of the characterization of structural features important for the property considered. Conversely, descriptors that are successful are successful because they cover the important part of the characterization of structural features important for the property considered. In short, the descriptors that yield good regressions are able to reflect well the salient structural features of importance for the property considered.

MULTIPLE REGRESSION ANALYSIS (MRA)

Kier and Hall have examined structure−property relationships and structure−activity relationships for numerous molecules and numerous properties and searched for the best correlation using various combinations of the higher order connectivity indices.[9,10] Besides the connectivity indices based on paths they augmented the pool of indices by considering subgraphs associated with such fragments as cluster ($k\chi_C$) and path-cluster ($k\chi_{PC}$):

By optimizing the set of indices used in each application it is not surprising that different combinations of connectivity indices characterized different situations. When using different descriptors in different approaches it is not easy to make a comparison between different studies, either when it concerns the same property for different families of compounds or different properties for the same family. For example, to describe the heats of atomization of a set of 44 alkanes Kier and Hall used as descriptors n, the number of atoms, and the following connectivity indices: ${}^1\chi$, ${}^4\chi$, ${}^5\chi_C$, ${}^4\chi_{PC}$, ${}^5\chi_{PC}$, ${}^6\chi_{PC}$.[9] In a regression of the heat of formation for the same set of alkanes they used the descriptors: ${}^1\chi$, ${}^2\chi$, ${}^3\chi$, ${}^4\chi$, ${}^5\chi_C$, ${}^4\chi_{PC}$, ${}^5\chi_{PC}$, ${}^6\chi_{PC}$.

Using only a single descriptor, the number of carbon atoms n, the correlation of alkane heat of atomization and the correlation of alkane heat of formation are given respectively by[8]

$$\Delta H_a = 280.53n + 115.72$$

and

$$\Delta H_f = 6.087n + 12.24$$

Here the coefficients of the regression are 0.9999 and 0.9959, and the standard error are 1.455 and 1.32, respectively. The correlation coefficient is impressive but still misleading. The "success" of n as a single descriptor is, however, questionable because n clearly cannot differentiate among isomers and therefore reflects the dependence of the property on the size only.

With two descriptors the correlation of alkane heat of atomization and the correlation of alkane heat of formation are given respectively by

$$\Delta H_a = 283.33n - 6.321\,{}^1\chi + 115.72$$

and

$$\Delta H_f = 7.649n - 3.286\,{}^1\chi + 11.70$$

with the coefficients of the regression >0.9999 and >0.9971, and the standard error are 0.960 and 1.13, for ΔH_a and ΔH_f, respectively.

By including several additional descriptors the standard errors were reduced considerably. However, an interpretation of the equations

$$\Delta H_a = 286.15n - 12.08\,{}^1\chi + 0.92\,{}^4\chi + 1.50\,{}^5\chi - 2.44\,{}^5\chi_C + 0.86\,{}^4\chi_{PC} - 0.50\,{}^5\chi_{PC} - 1.42\,{}^6\chi_{PC} + 114.65$$

$$\Delta H_f = 1.15\,{}^1\chi - 2.52\,{}^2\chi + 7.63\,{}^3\chi_C - 12.02\,{}^4\chi_C - 1.72\,{}^5\chi_C + 0.89\,{}^4\chi_{PC} - 1.46\,{}^5\chi_{PC} - 0.28$$

becomes difficult. Even if we consider the same descriptor, such as ${}^1\chi$, ${}^5\chi_C$ or ${}^4\chi_{PC}$ and ${}^5\chi_{PC}$, they are combined in each equation with different descriptors to allow any meaningful comparison.

A comparative study of different properties for same compounds is hindered by using different descriptors each time, even if they may be the optimal choice for each case. From the given regression equations it appears as if path subgraphs are more important for ΔH_a and cluster fragments for ΔH_f. A direct correlation between ΔH_a and ΔH_f

$$\Delta H_a = 53.260\Delta H_f - 413.353$$

with

$$r = 0.9818, \quad s = 78.91, \text{ and } F = 1070$$

shown in Figure 4, also does not reveal a close relationship between the two properties. From Figure 4 we see that ΔH_a increases regularly with the size of molecules since for each set of isomers there is a regular increment in ΔH_a. In contrast the values of ΔH_f somewhat overlap between the isomers of different size.

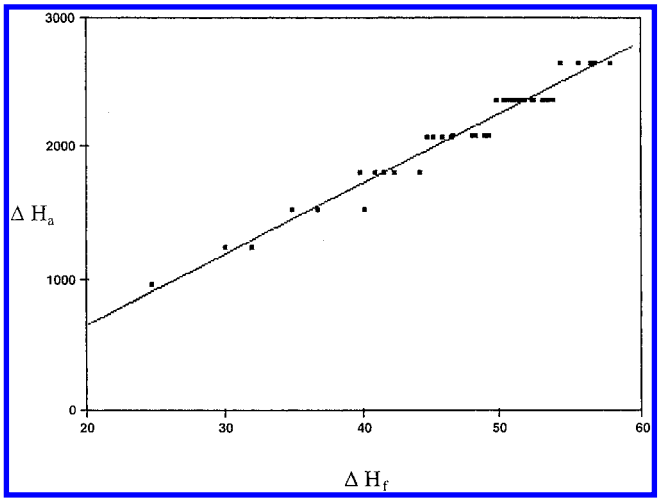


Figure 4. Plot of the heats of atomization (ΔH_a) against the heats of formation (ΔH_f) for alkanes (all in kcal).

For the properties considered the size of alkanes (described by n) is clearly the dominant structural factor. It is therefore of interest to examine how these properties vary for molecules of the same size. When only the 18 isomers of octane are considered we obtain a somewhat unexpected result (Figure 5):

$$\Delta H_a = \Delta H_f + 2308.12$$

i.e., for isomers of octane ΔH_a and ΔH_f are collinear! The same is true for other isomers. In the case of hexane isomers, for example, we have $\Delta H_a = \Delta H_f + 1751.14$. Hence, the two equations only differ in the constant. This may have been the reason that the close parallelism between ΔH_a and ΔH_f was obscured in the multiple regression applied to alkanes of different size. By using different descriptors not only the parallelism between the two properties could not be established but also the approach failed to point to the fact that for isomers the two properties reduce to the same common basis.

HOW TO CHOOSE DESCRIPTORS

Structure−property and structure−activity studies using MRA have sometimes been viewed as a mixture of "art" and "science." The "art" part is usually confined to the selection of descriptors. There are no *a priori* rules that apply to the selection of variables. In addition, one should be aware that more than one choice of descriptors may produce the results of the same quality.[23]

While there are no rules for selecting descriptors, often a "rule" has been imposed to exclude a descriptor. If a descriptor strongly correlates with another descriptor already used in a regression, such a descriptor in most studies has been discarded. For example $^1\chi$ and $^2\chi$ often strongly correlate and in many structure−property−activity studies $^2\chi$ has been discarded. This is not theoretically justified and despite the widespread practice should be stopped. Although two highly correlated descriptors overall depict the same features of molecular structure, it is important to recognize that even highly interrelated descriptors *differ* in some other structural traits. The difference between them may be relatively small but nevertheless very important for structure−property regression.

The criteria for inclusion or exclusion of descriptors should not be based on parallelism between descriptors even if overwhelming, but it should be based on whether the part in which two descriptors disagree is or is not relevant for the characterization of the property considered. If the part in which the second descriptors differ from the first, regardless of how small it is, is relevant for the property considered the descriptor should be included, otherwise it should be discarded.

Consider as an example the correlation of $^1\chi$ and $^2\chi$ with the molar refraction of octanes.[24] A simple regression for the molar refraction of octanes using the connectivity index $^1\chi$ as the descriptor gives practically no correlation: regression coefficient $r = 0.087$ and standard error $s = 0.187$. We may conclude that clearly $^1\chi$ is not a useful descriptor for this property. When we try $^2\chi$ (the second order connectivity index) we find a correlation with $r = 0.177$ and $s = 0.185$. Again this shows that $^2\chi$ is also not a good descriptor. The latter is not surprising because $^1\chi$ and $^2\chi$ have very high mutual correlation (above 98%).

Now consider the multiple regression using the two "proven" bad descriptors, $^1\chi$ and $^2\chi$. One finds

$$MR = 4.6951\ ^1\chi + 1.3720\ ^2\chi + 17.5482$$

to be a strong correlation of MR with both $^1\chi$ and $^2\chi$ taken together: $r = 0.971$ and $s = 0.047$ ($F = 114.8$)! Most people would discard both descriptors, $^1\chi$ and $^2\chi$, and yet when considered jointly we find that $^1\chi$ and $^2\chi$ result in an impressive correlation.

Why $^1\chi$ and $^2\chi$ work *together* and not separately can be understood. Figuratively speaking, both $^1\chi$ and $^2\chi$ point to the wrong "direction" in the structure−property space from the expected direction for the property considered. However, when combined $^1\chi$ and $^2\chi$ *span* the structure−property space and therefore are capable of giving good correlation. What happened is that the part in which $^1\chi$ and $^2\chi$ differ, even if very small, is here relevant. Apparently $^2\chi$ differs from $^1\chi$ precisely in the structural feature that is relevant for MR as the property. Pictorially this can be depicted by two vectors almost parallel, directed more or less along the x-axis (i.e., $^1\chi$-axis), while the y-axis is the direction that points to the desired structural element relevant for the property considered. Nevertheless, two vectors that are not strictly parallel, regardless of their direction, span the plane and can be used as basis vectors to represent any direction in the plane. On the other hand, suppose that the desired direction for a property of interest is, pictorially speaking, perpendicular to the plane defined by the two vectors representing $^1\chi$ and $^2\chi$. In that case we could say that indeed neither $^1\chi$ nor $^2\chi$ are important for the regression analysis considered, unless they can be combined with a descriptor that will span the third dimension of the structure space.

BASIS

"The chief problem in every science is that of endeavoring to arrange and collate the numerous individual observations and details which present themselves, in order that they may become part of one comprehensive picture." Max Planck[1]

Using the same set of the descriptors in different application has some advantages. Such a set can be viewed as a basis for the representation of molecules. How many descriptors are to be included in a basis will vary from case

to case, depending on the size of the sample considered. In the case just discussed we found that structure-space can be viewed as a two-dimensional space. The descriptors $^1\chi$ and $^2\chi$ have been found to span this space. Similarly we can view Wiener's *W* and *P* as descriptors that span well the structure space for numerous thermodynamic properties of alkanes, aliphatic acids, amines, and fatty acids. Not *any* pair of descriptors will satisfactorily span the relevant part (for the property considered) of the structure space. Hence, the search for novel descriptors translates into the search for a relevant basis (for the property considered) for structure space for compounds considered. The properties that mutually do not correlate are associated with different subspace of the structure space. Similarly to the principal component analysis (PCA),[25] if the search for descriptors is successful (and the same is true for successful PCA) we can find an upper bound for the number of descriptors necessary for acceptable regression for a predetermined predictive accuracy. One of the goals of the systematic search for structure−property correlations is to discern which descriptors can adequately characterize which property.

The distinction between a *collection* of descriptors and a *basis* based on the same set of descriptors is that the latter requires a prior *ordering* of the descriptors. In practice the difference is reflected in the distinction between the multiple regression analysis and the stepwise multiple regression analysis. In the former one looks for regression based on several descriptors d_i, in the latter one gradually increases the number of descriptors, the first step is given by using d_1, the second step by using d_1, d_2, then d_1, d_2, d_3, etc. It is desirable to use the same set of the ordered descriptors d_1, d_2, d_3, ... in different applications. In Table 2 we summarize the regression using the three connectivity indices $^1\chi$, $^2\chi$, $^3\chi$, as the basis for a comparative study of ΔH_v of octanes.[26] We restricted the analysis to octane isomers in order to eliminate the dominant role of the molecular size on the properties.

ORTHOGONAL BASIS

"...the experimenter cannot afford to close his eyes to a new discovery, obtained from another point of view, which will not fit in with his own ideas, nor must he treat it as unimportant, if not incorrect." Max Planck[1]

The importance of using a basis rather than a set of descriptors is that a basis can be orthogonalized. The result is that the mutual dependence of descriptors that has plagued MRA can thus be eliminated.[27−38] While the orthogonalization process does not alter the statistical parameters for the prediction of the property that is considered, the regression coefficient *r*, the standard error *s*, and Fisher ratio *F*, it leads to a stable regression equation, the coefficients of which are not affected by inclusion of additional descriptors or exclusion of some of the descriptors already used. Hence, the orthogonalized MRA approach for the first time allows one to *interpret* the contributions of the variables used in the regression analysis. In view of the long history of MRA and its widespread use not only in chemistry but also in physics, biology, medicinal sciences, psychology, sociology, and related disciplines this recent development was long overdue.

Orthogonal variables are no novelty in structure−property−activity studies. Already in 1933 Hotelling[25] introduced the Principal Component Analysis in which the final variables are mutually orthogonal linear combinations of the initial set of descriptors. The problem, however, is that while each individual descriptor may have a simple structural interpretation, their *linear combinations* have no clear interpretation. This is not only because they represent a superposition of several descriptors but also the descriptors that are so combined are mutually interrelated often strongly interrelated.

Table 2. Stepwise Regression Equations for the Heats of Vaporization for 18 Octane Isomers Using $^1\chi$, $^2\chi$, and $^3\chi$

	r	*s*
$\Delta H_v = -6.4224\ ^1\chi + 75.2462$	0.8504	0.6786
$\Delta H_v = +6.2484\ ^1\chi - 3.7949\ ^2\chi + 16.9911$	0.9302	0.4890
$\Delta H_v = +1.9200\ ^1\chi + 2.5539\ ^2\chi - 0.3829\ ^3\chi + 37.4108$	0.9311	0.5029

Table 3. Stepwise Regression Equations for the Heats of Vaporization for 18 Octane Isomers Using Orthogonal Descriptors $^1\Omega$, $^2\Omega$, and $^3\Omega$

	r	*s*
$\Delta H_v = -6.4224\ ^1\Omega + 75.2462$	0.8504	0.6786
$\Delta H_v = -6.4224\ ^1\Omega - 3.7949\ ^2\Omega + 75.2462$	0.9302	0.4890
$\Delta H_v = -6.4224\ ^1\Omega - 3.7949\ ^2\Omega - 0.3829\ ^3\Omega + 75.2462$	0.9311	0.5029

In Table 3 we illustrate the regression equations for the heats of formation for 18 isomers of octane (already considered in Table 2) with orthogonalized connectivity indices $^1\chi - {}^3\chi$. The following are the important features of Table 3 that need to be emphasized:

(1) The statistical parameters *r*, *s*, and *F*, the regression coefficient, the standard error, and the Fisher ratio, respectively, are the same for the orthogonalized descriptors and nonorthogonalized descriptors.

(2) The coefficients of the orthogonalized variables do not change when an additional descriptor is introduced.

(3) The coefficients of Table 3 have appeared also in Table 2; each time the new variable is added in the stepwise regression.

(4) Each time an additional variable is introduced in MRA, the standard error for the coefficients of the regression equation worsen for nonorthogonalized procedure, but they improve for the orthogonalized procedure.

(5) The stability of the regression equation, i.e., the constancy of the coefficients, allows assigning the relative role to individual orthogonal components.

Orthogonal MRA has been recognized as the important development in structure−property−activity studies and has been extended and applied.[39−51]

COMPARISON OF GRAPH THEORETICAL AND THE TRADITIONAL QSAR DESCRIPTORS

For many years criticism has been raised in some circles against chemical graph theory and against topological indices in particular, primarily because the apparent inability of graph theory to handle three-dimensional molecular structure. A cry "Topology is dead" reflects some such concerns, but to this attitude one can reply by using the words of Mark Twain: "News about my death are highly exaggerated". Topology (i.e., topological indices) is alive and has seen important developments in the last few years.[52−54] Before we list some of the new directions in the development of topological indices let us mention that the premise that graph

theory is not capable of capturing some aspects of 3-D molecular structure is false. Two examples will illustrate this point. By augmenting molecular graph with additional lines that connect selectively nonadjacent hydrogen atoms one can fully represent 3-D conformers of a molecular chain embedded on a diamond grid.[55] In another study it was noticed that besides the shortest-path metrics other metrics defined for graphs lead to graph invariants that have geometrical interpretations and correspond to classical geometric invariants.[56]

More recent developments of the "dead" topology include use of novel matrices associated with molecular graphs as a source of molecular descriptors,[57−66] novel indices,[67−75] indices that represent various generalizations of well-established indices,[76−79] in particular generalized Wiener index[80−84] and generalized Hosoya Z index,[85−88] general procedures for construction of a new generation of indices,[89,95] augmented basis descriptors,[96] and graph reconstruction from invariants.[97−106] In addition a novel fertile scheme was proposed for generating invariants from matrices the elements of which are qualified subgraphs.[107−109] Not only that mathematical descriptors have been found adequate for characterization of numerous molecular properties but also they have been found to surpass in quality the traditional QSAR descriptors. Several studies compared use of topological indices and the traditional QSAR descriptors in MRA and again and again found mathematical descriptors superior.[110−113]

3-D DESCRIPTORS

The criticism that topological indices are not applicable to three-dimensional structure is valid and holds. Topological indices cannot differentiate conformers such as chair and boat conformation of cyclohexane, *cis*−*trans* isomers, *gauche,* and *anti,* etc. However, the *methodology* that has developed with the growth of the chemical graph theory can be extended to three-dimensional structure! All that is required is from the start to consider the molecular geometry rather then molecular connectivity. Instead of the adjacency matrix of molecular graph one considers geometric matrix for a structure rigidly embedded in 3-D. However, in contrast to the main theme of Crippen's Distance Geometry[114,115] which is concerned with the constraints imposed by interatomic distances and how to minimize the inconsistencies that originate with experimental data, in the mathematical chemistry the emphasis is on structural *invariants* of the geometrical distance matrices.[116−136]

Among simple matrix invariants are the determinant of matrix, the spectrum of the matrix, the characteristic polynomial, the minimal polynomial, etc. However these well-known algebraic invariants often are not of interest in structure−property applications. Hence a search for 3-D structural invariants, that will parallel the search for topological indices of a decade ago, is only to be expected to flourish in the near future.

First 3-D descriptors were constructed analogously to the "classical" topological indices χ and W. Hence, we arrive at 3-D connectivity indices[117−119] and 3-D Wiener index.[120−127] More recently a scheme was outlined that generates mathematical descriptors, i.e., structural invariants, for 3-D structure, which are not only sufficiently general and apply to molecules of arbitrary geometrical forms but also offer a sizable set of structurally related invariants. These novel 3-D descriptors have been referred to as molecular profile.[129−136] Importantly, the approach has the necessary flexibility that it can be extended to molecules having heteroatoms, although at present only carbon molecular skeletal forms have been considered.

MOLECULAR PROFILE: THREE-DIMENSIONAL BASIS FOR STRUCTURES

Consider a molecular structure embedded in 3-D, or alternatively embedded in 2-D, i.e., a structure which is fully determined by the positions of its atoms in the space, or the plane, respectively. For such a structure one can derive its geometric distance matrix D, the (i, j) entry of which is the interatomic separation between atom i and atom j:

$$\mathrm{D_{i,j}} = \begin{array}{c} \text{Row Sum} \\ \begin{pmatrix} d_{1,1} & d_{1,2} & d_{1,3} & d_{1,4} & d_{1,5} \\ d_{2,1} & d_{2,2} & d_{2,3} & d_{2,4} & d_{2,5} \\ d_{3,1} & d_{3,2} & d_{3,3} & d_{3,4} & d_{3,5} \\ d_{4,1} & d_{4,2} & d_{4,3} & d_{4,4} & d_{4,5} \\ d_{5,1} & d_{5,2} & d_{5,3} & d_{5,4} & d_{5,5} \end{pmatrix} \end{array} \begin{array}{c} \\ \\ R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{array}$$

The row sums of the matrix $\Sigma_j d_{i,j} = R_\mathrm{i}$ can be viewed as local atomic structural invariant, while the average row sum: $(1/n)\Sigma_i\Sigma_j d_{i,j} = R$ represents molecular structural invariant. The latter, except for normalization factor, is equivalent to 3-D Wiener number for the particular structure. An alternative normalization, $(1/n^2)\ \Sigma_i\Sigma_j d_{i,j} = R/n$, gives the average $d_{i,j}$ entry, i.e., the average contribution of an atom to the particular molecular invariant. The latter is of interest when one is comparing molecules having different numbers of atoms or in general structures of different size or the same structure represented by different numbers of points (*vide infra*).

A single invariant is hardly representative of a 3-D structure. What one desires is to have a set of structurally related invariants. How are we to get additional structural invariants from the matrix element $d_{i,j}$? As outlined in the literature we consider a family of matrices ${}^kD$ the elements of which are kth powers of $d_{i,j}$. Each such matrix will produce the average row sum ${}^kR$, and thus the molecule is characterized by the sequence

$${}^1R, {}^2R, {}^3R, {}^4R, {}^5R, {}^6R, \ldots$$

As the exponent k increases the contributions of the most distant pairs of atoms are getting the prominence. The sequence ${}^1R, {}^2R, {}^3R, {}^4R, {}^5R, {}^6R$, ... therefore will reflect the molecular shape to some extent because for molecules of a same size (e.g., isomers) the structures that are spherical will have smaller components ${}^kR$ compared with elongated structures in which atoms at large separations will dominate the row sums for larger values of k.

Since larger separations plays the dominant role as k increases the sequence ${}^kR$ will diverge as k increases. A normalization factor $1/k!$ will curb the divergent nature of the power sequence and transform it into a convergent sequence by reducing gradually the polynomial growth of the individual entries of the $D_{i,j}$ matrix. The sequence

$${}^1R, {}^2R/2!, {}^3R/3!, {}^4R/4!, {}^5R/5!, {}^6R/6!, \ldots$$

has been referred to as a molecular profile. In Figure 6 we

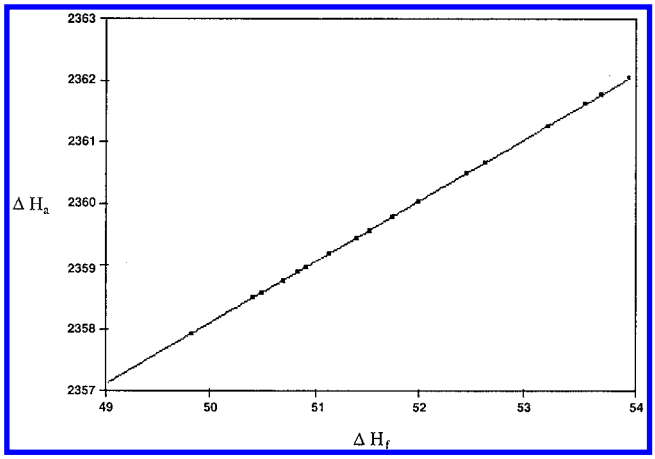


Figure 5. Plot of the heats of atomization (ΔH_a) against the heats of formation (ΔH_f) for octane isomers (all in kcal).

illustrate molecular profiles for the six conformers of 1,3,5-hexatriene. We assumed all CC bond lengths to be of length 1 and all angles to be 120°, i.e., the isomers are embedded on a graphite lattice.

MOLECULAR SHAPE

There are molecular properties that are determined mostly by the shape of the molecule, rather than its connectivity. An example is the smell and the taste. For example, the macrocyclic civetone and polycyclic sterol shown in Figure 7, as pointed out by Ružička and Prelog some time ago,[137] have similar musk odor. Most topological indices would fail to detect the apparent similarity between the above two compounds because most of the indices are derived from molecular connectivity or in the case of 3-D forms from the molecular spatial bonding pattern. However, the internal bonds do not participate in defining the shape, yet make contribution to computed graph or structural invariants, unless explicitly excluded. We can arrive at molecular *shape profiles* by considering only the contributions from atoms which are on the molecular periphery. In Table 4 we show the shape profiles for anthanthrene and coronene shown in Figure 8, both benzenoids having 18 carbon atoms on the molecular periphery. As one sees from Table 4 for $k < 4$ the profile components in coronene dominate those for anthanthrene but for $k \geq 4$ the opposite is the case. This reflects the fact that coronene is circular in shape and anthanthrene is elongated (oval).

BONDING PROFILE, CONTOUR PROFILE

In some problems a characterization of a structure by focusing attention solely to atoms does not suffice. In order to obtain useful description of a structure in such cases one has to consider also the bonding pattern. For example, an idealized *cis*−*cis*−*cis* hexatriene (embedded on a graphite lattice) and benzene ring have all six atoms at the same positions. Hence, all the atoms in both cases will have identical corresponding coordinates and consequently the same corresponding interatomic separations. However, the two systems differ in their connectivity. The connectivity is not reflected in the Euclidean distance matrix which records only interatomic separations. Consequently the two systems would yield an identical molecular profile (when based on carbon atoms only). In order to distinguish the two systems one can introduce m regularly spaced points along each CC bond and instead of considering 6×6

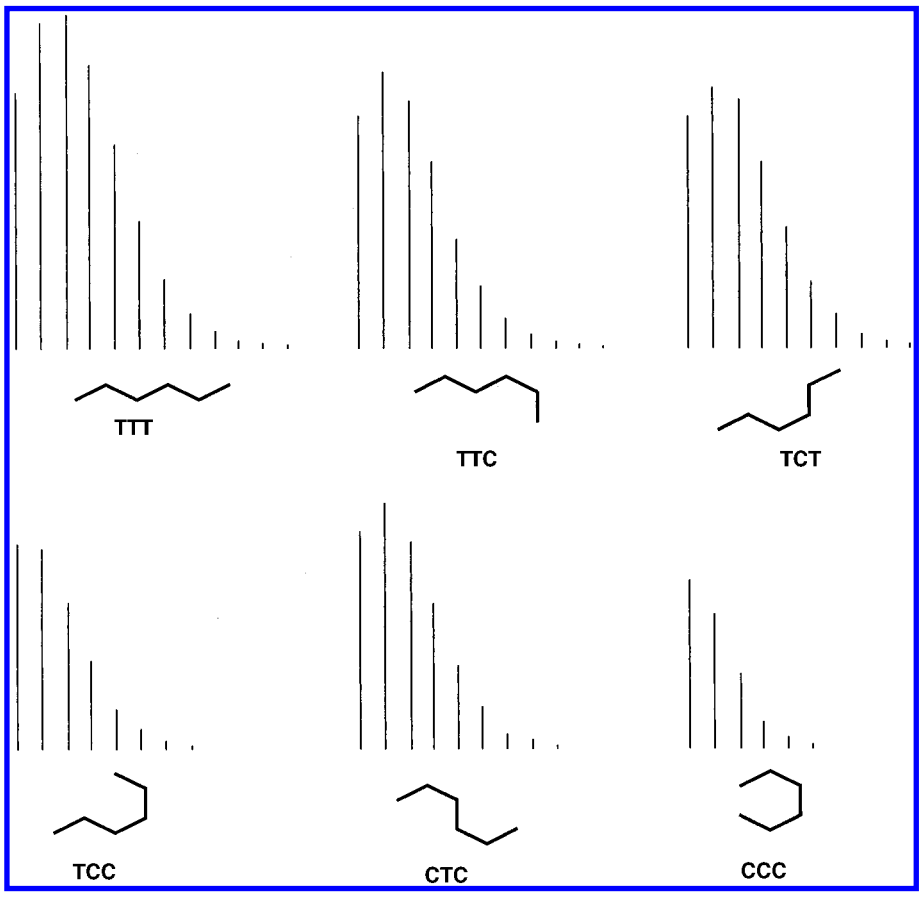


Figure 6. Molecular profiles for the six conformers of 1,3,5-hexatriene.

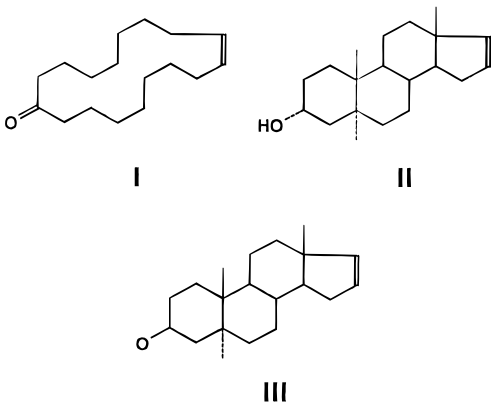


Figure 7. Macrocyclic civetone and two polycyclic sterols having the characteristic musk odor.

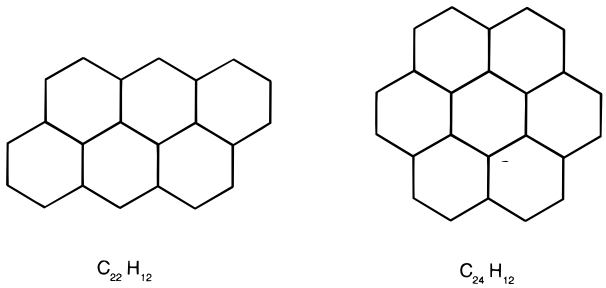


Figure 8. Carbon atom skeletons for anthanthrene $C_{22}H_{12}$ and coronene $C_{124}H_{12}$.

Table 4. Shape Profiles of Two Benzenoids of Figure 8

		anthanthrene	coronene
1	^1R	54.3	56.1
2	^2R/2!	103.5	108.0
3	^3R/3!	147.4	150.5
4	^4R/4!	170.3	165.0
5	^5R/5!	166.7	149.3
6	^6R/6!	142.3	115.1
7	^7R/7!	107.8	77.4
8	^8R/8!	73.6	46.1
9	^9R/9!	45.7	24.7
10	^{10}R/10!	26.0	12.0
11	^{11}R/11!	13.7	5.4
12	^{12}R/12!	6.7	2.2
13	^{13}R/13!	3.1	0.8
14	^{14}R/14!	1.3	0.3
15	^{15}R/15!	0.5	0.1
16	^{16}R/16!	0.2	
17	^{17}R/17!	0.1	

matrices we would have $(5m+6) \times (5m+6)$ and $(6m+6) \times (6m+6)$ matrices for the open chain and the closed ring systems, respectively. The procedure generates matrices of different sizes even for a single structure. The dimension of the matrix depends on m, the number of "ghost" sites introduced along the bonds. The contributions to the molecular profile should therefore be normalized to correspond to a single site contribution. If m is sufficiently large, the molecular bonding profiles, as we refer to this generalization of molecular profiles, apparently converge (Figure 9). Hence, profiles appear to be independent of m. This is illustrated in Table 5 for the case of cuboctahedron (Figure 10).

The molecular profile approach can be further generalized. Instead of using points along chemical bonds one can use m points located uniformly along the molecular contour, on the molecular periphery, or within the molecular volume. These will produce the following: contour profile, generalized

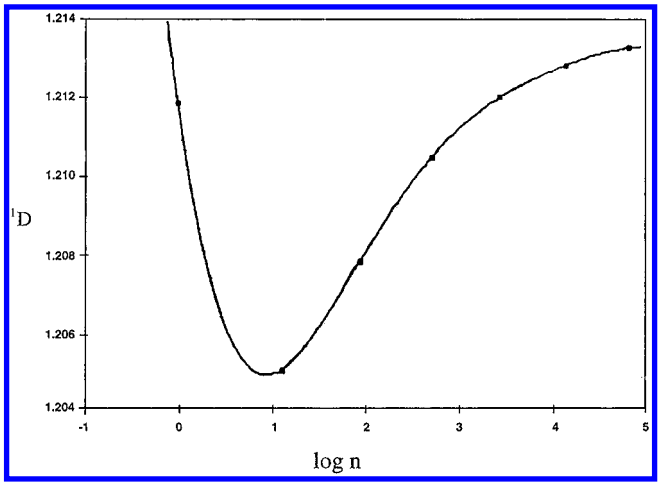


Figure 9. Converegence of the descriptor ^1D as the size of the matrix(N) increases.

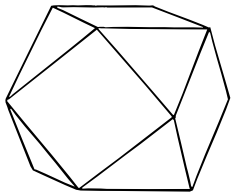


Figure 10. Cuboctahedron.

Table 5. The Cuboctahedron Profile as the Function of the Ghost Points Along Each Edge of the Polyhedron

	$m=0$	$m=1$	$m=3$	$m=7$	$m=15$	$m=31$	$m=63$
1		1.8043	1.7848	1.7783	1.7759	1.7749	1.7745
2		1.8585	1.8167	1.8032	1.7984	1.7965	1.7956
3		1.3804	1.3334	1.3186	1.3133	1.3112	1.3103
4		0.8144	0.7771	0.7654	0.7613	0.7597	0.7590
5		0.4018	0.3785	0.3713	0.3688	0.3678	0.3674
6		0.1712	0.1591	0.1555	0.1542	0.1537	0.1535
7		0.0644	0.0591	0.0574	0.0569	0.0567	0.0566
8		0.0218	0.0197	0.0190	0.0188	0.0187	0.0187
9		0.0067	0.0059	0.0057	0.0057	0.0056	0.0056
10		0.0019	0.0016	0.0016	0.0016	0.0016	0.0015

Table 6. Molecular Profiles for van der Waals Model of H_2O

	2-D model		3-D model	
1	0.39923		0.35901	
2	0.98066	10^{-1}	0.79687	10^{-1}
3	0.18187	10^{-1}	0.13441	10^{-1}
4	0.27513	10^{-2}	0.18609	10^{-2}
5	0.35395	10^{-3}	0.22034	10^{-3}
6	0.39756	10^{-4}	0.22890	10^{-4}
7	0.39716	10^{-5}	0.21241	10^{-5}
8	0.35779	10^{-6}	0.17843	10^{-6}

shape characterization, and generalized molecular characterization, respectively. In Table 6 the generalized molecular characterization for a water molecule H_2O is illustrated by using space-filling model based on van der Waals radii for oxygen and hydrogens. In Figure 11 is illustrated the distribution of 5000 random points (which simulate a "uniform" distribution) within van der Waals contours for H_2O, and in Figure 12 is illustrated the distribution of 5000 random points within 3-D model of H_2O.[122]

D/D MATRICES

A quite different approach to 3-D structures has been proposed by Randić, Kleiner, and DeAlba,[58] who were

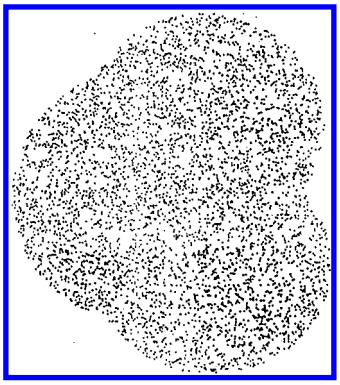
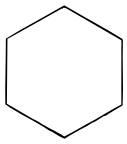
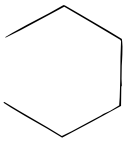


Figure 11. Representation of H_2O by van der Waals contour filled with 5000 random points.

interested in modifying the Euclidean distance matrix so that it reflects some features of the molecular connectivity. The result is the so-called D/D matrix, the elements of which are given by the ratio of Euclidean distance to the graph theoretical distance in a structure. For example, the D/D matrices for benzene ring and idealized *cis*−*cis*−*cis* conformation of hexatriene are now distinct



0	1	$\sqrt{3}/2$	2/3	$\sqrt{3}/2$	1
1	0	1	$\sqrt{3}/2$	2/3	$\sqrt{3}/2$
$\sqrt{3}/2$	1	0	1	$\sqrt{3}/2$	2/3
2/3	$\sqrt{3}/2$	1	0	1	$\sqrt{3}/2$
$\sqrt{3}/2$	2/3	$\sqrt{3}/2$	1	0	1
1	$\sqrt{3}/2$	2/3	$\sqrt{3}/2$	1	0



0	1	$\sqrt{3}/2$	2/3	$\sqrt{3}/4$	1/5
1	0	1	$\sqrt{3}/2$	2/3	$\sqrt{3}/4$
$\sqrt{3}/2$	1	0	1	$\sqrt{3}/2$	2/3
2/3	$\sqrt{3}/2$	1	0	1	$\sqrt{3}/2$
$\sqrt{3}/4$	2/3	$\sqrt{3}/2$	1	0	1
1/5	$\sqrt{3}/2$	2/3	$\sqrt{3}/2$	1	0

The derived D/D matrices can serve as a source for numerous structure invariants that are analogous to some of the topological indices derived from the graph adjacency matrix or graph distance matrix. These invariants include the (weighted) path numbers, molecular ID number, and local invariants such as atomic ID numbers (i.e., the row sums), and the traditional matrix invariants, such as the characteristic polynomial, determinant, and matrix eigenvalues.

Of particular interest appears to be the leading eigenvalue of the D/D matrix (i.e., the largest positive eigenvalue). In the case of trees (acyclic graphs) Lovasz and Pelikan had observed that the leading eigenvalue of the adjacency matrix can be viewed as a quantitative index of the branching of the tree.[138] Similarly, the leading eigenvalue of the D/D matrices appears to be a quantitative index of the bending

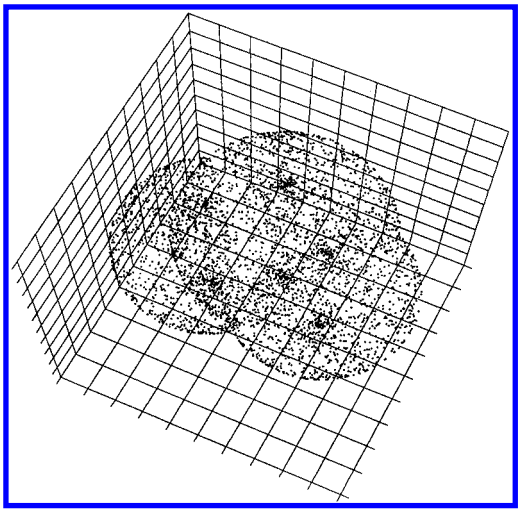


Figure 12. Representation of H_2O by van der Waals space filling model filled with 5000 random points.

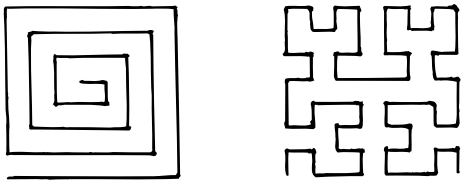


Figure 13. Spiral curve and Hilbert curve for the initial 64 points.

or the folding of a chain-like structure. If we order the hexatriene conformers according to the highest eigenvalue we obtain[139]

	TTT	TTC	TCT	CTC	TCC	CCC
λ	4.572	4.419	4.403	4.309	4.156	3.897
ϕ	0.7620	0.7365	0.7338	0.7182	0.6927	0.6495

Here λ is the leading eigenvalue and ϕ is normalized eigenvalue λ/n, which we take to represent the folding index for a chain-like structure. It is not difficult to see that for a straight line having n vertices we obtain $\phi = (n-1)/n$, which in the limit of an infinite chain gives $\phi = 1$. On the other hand, ϕ for a zigzag line that overlaps itself as the length increases, apparently approaches 0.[38,64] Hence, the index ϕ is limited to the intermediate values, the smaller ϕ corresponding to the more folded curves.

FOLDING OF MATHEMATICAL CURVES

In Figure 13 are shown two mathematical curves for which the corresponding folding index is listed in Table 7 as n, the number of points on the curve, increases. From the illustrations given we see that the folding index is quite sensitive

Table 7. The Folding Index as a Function of the Number of Points Representing a Curve for Spital Curve, Hilbert Curve, and Zigzag Curve (Having Self-Overlapping Sharp Turns in Alternate Directions)

n	spiral curve	n	Hilbert curve	n	zigzag curve
8	0.6099	8	0.6121	8	0.3026
16	0.5474	16	0.5952	16	0.1947
24	0.5010	24	0.4685	24	0.1467
32	0.4696	32	0.4113	32	0.1190
40	0.4404	40	0.3712	40	0.1008
48	0.4181	48	0.3464	48	0.0878
56	0.4014	56	0.3217	56	0.0780
64	0.3860	64	0.3026	64	0.0704

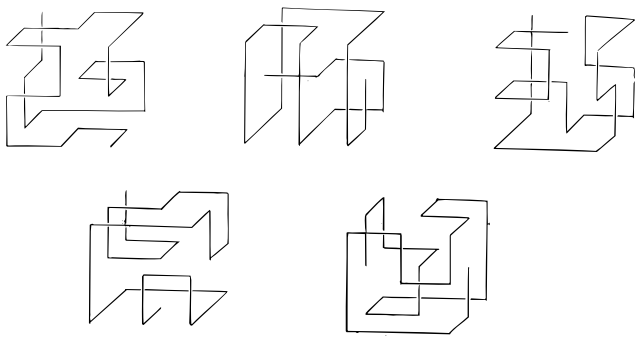


Figure 14. The geometrical form of the model proteins of C. Tang and co-workers.

to geometrical detail of the curves. The simple spiral, as it grows, is built from larger and larger straight line segments, and it is expected therefore that although as n increases the index ϕ will decrease it will decrease at a slower rate than in the case of fractals. Hilbert curve, which illustrates a step in construction of a Hilbert fractal, as it grows apparently becomes more and more folded. Hence the rate of change of ϕ for Hilbert curve is expected to be greater than for the spiral. Table 7 confirms the expectations and thus supports the interpretation of ϕ as a folding index.

Single index may not suffice for detailed characterization of molecular folding. It is possible to construct additional indices associated by folding by considering the higher order D/D matrices. These are obtained from the D/D matrix by raising elements (i, j) of D/D to the power k. In this way we obtain folding profile ${}^1\phi$, ${}^2\phi$, ${}^3\phi$, ${}^4\phi$, ${}^5\phi$, ${}^6\phi$, ... which we will illustrate in the next section on model of "toy" proteins of C. Tang and co-workers.[140]

FOLDING OF PROTEINS

Folding of proteins continue to dominate the efforts of many laboratories interested in clarifying the relationship between the structure and the function of these complex molecules. Much effort was made in trying to relate the amino acid sequence of the primary structure to the tertiary structure of a protein. Our goal here is very modest: Rather than looking for factors that determine the tertiary structure (such as van der Waals interactions, electrostatic interactions, interactions of hydrogen bonds, the role of solvents, polarizability, etc.) we will assume that tertiary structure of a protein is given and will consider a characterization of such 3-D structure. In other words, we are interested to assign to each protein a characterization, starting by a single number and followed by a sequence of invariants, that has some structural interpretation. As we will see it is possible to derive not only a molecular profile for a protein but also an index of folding.

In Figure 14 we illustrate five model proteins studied by C. Tang and co-workers as reported in the literature.[141−143] The amino acids were classified as either hydrophobic or polar. Observe that in all five cases the proteins are embedded within a $2 \times 2 \times 2$ cube. Hence, all five model proteins would have the same distance matrix if all amino acids are suitably labeled. The model assumes that each vertex represents an amino acid. The five structures differ only in the spatial pattern of connectivity, i.e., the way of linking of the 27 points that define the exterior and the interior of the cube. In order to differentiate the five protein structures we have to consider bonding pattern, i.e., we have to augment each structure by m "ghost" points along individual links. Already a single point inserted between any two amino acids discriminates among the five cases illustrated in Figure 14. In Table 8 we show convergence of the profiles as m, the number of inserted points along each link, increases.

Table 8. Convergence of the Profiles for Protein A as the Number of Sites Along Each Link Increases

	$k = 1$	$k = 3$	$k = 7$	$k = 15$	$k = 31$	$k = 63$
1	1.8043	1.7848	1.7783	1.7759	1.7749	1.7745
2	1.8585	1.8167	1.8032	1.7984	1.7965	1.7956
3	1.3804	1.3334	1.3186	1.3133	1.3112	1.3103
4	0.8144	0.7771	0.7654	0.7613	0.7597	0.7590
5	0.4018	0.3785	0.3713	0.3688	0.3678	0.3674
6	0.1712	0.1591	0.1555	0.1542	0.1537	0.1535
7	0.0644	0.5905	0.5743	0.5688	0.5668	0.0566
8	0.0218	0.0197	0.0190	0.0188	0.0187	0.0187
9	0.0067	0.0059	0.0057	0.0057	0.0056	0.0056
10	0.0019	0.0016	0.0016	0.0016	0.0016	0.0015

Table 9. The Profiles for the Model Proteins

power	A	B	C	D	E
1	1.7745	1.8124	1.7922	1.7934	1.8121
2	1.7956	1.8713	1.8313	1.8328	1.8713
3	1.3103	1.3909	1.3482	1.3493	1.3912
4	0.7590	0.8192	0.7870	0.7877	0.8195
5	0.3674	0.4025	0.3836	0.3840	0.4028
6	0.1535	0.1704	0.1612	0.1614	0.1706
7	0.0566	0.0636	0.0597	0.0598	0.0637
8	0.0187	0.0213	0.0199	0.0199	0.0213
9	0.0056	0.0064	0.0060	0.0060	0.0065
10	0.0016	0.0018	0.0017	0.0020	0.0018

In Table 9 we listed the profiles for the five model proteins as obtained when $m = 63$. This implies characterization of the structures considered not by 27×27 matrices but rather as 833×833 matrices (i.e., the 27×27 matrices augmented by 31 points along each link). We have ordered proteins by the relative magnitude of the first entry in their respective profiles. By inspection of the Table 9 it is immediately clear that among the model proteins considered B and E are similar and also C and D appear to be similar. The quantitative similarity measure points to the pair B, E as the most similar and the pair A, E as the least similar.[144]

It yet remains to be seen to what extent this quantitative characterization of proteins based on the average distances and their powers will parallel their properties. The molecular folding profile is based on the average ratios of Euclidean to "topological" (i.e., graph theoretical) distance. By inspection of the Figure 14 we can qualitatively rank the protein A as the most folded and B as the least folded by the count of the number of "long" line segments (i.e., the segments of length two). There are only two such fragments in A and eight in B, while the remaining structures have four or five such fragments. The quantitative characterization leads to the following values for the folding index:[144]

	A	B	C	D	E
ϕ	0.345	0.360	0.348	0.363	0.353

As we see A has been found the most folded of the five structures of Figure 14, while B and D have been found the least folded. The role of the eigenvalues of higher order D/D matrices has yet to be better studies before their use as additional structure descriptors.

CRITICS, SKEPTICS, AND ENEMIES

"A new truth always has to contend with many difficulties; if it were not so, it would have been discovered sooner." ...Max Planck[1]

There have been two groups of critics of mathematical chemistry, chemical graph theory, and chemometrics: outsiders and insiders. Outsiders have been loud in denouncing topological indices, but as time passed the size of this group must have diminished and their vehemence subsided. Topological indices, despite the misnomer in the label, are today accepted as valid mathematical descriptors of chemical structure.[145] Advantages of mathematical descriptors, as contrasted with the traditional descriptors of QSAR, are becoming more evident with time.

Apparently a group of skeptics consider MRA not as a suitable tool for data reduction approach. Extreme views were heard that MRA is but outright wrong. Most of these critics appear to be more interested in promoting their own methodology, such as the principal component analysis, the partial least square approach, the neural networks, the pattern recognition, etc. The above criticism is based on the fact that the regression equation shows notorious instabilities. Inclusion of an additional descriptor to MRA can completely change the coefficients previously employed. However, the orthogonalization process has cured this fatal illness of MRA. Instead of rejoicing that one of the oldest data reduction approach, MRA, has been *completely rehabilitated* and is available to supplement other alternative approaches, all of which are to some degree complementing one another, apparently some critics, some skeptics, and the enemies continue their campaign against MRA, and particularly against MRA in QSAR when using mathematical descriptors.

The simple least square fit can be traced back to Newton, Boscovich, and others some 300 years ago, hence it is probably the oldest data reduction procedure in science. In a generalized form it leads to multivariate regression analysis in which a single property is correlated with several variables. The method has been widely used in such diverse disciplines as econometrics, psychology, chemistry, biology, sociology, etc. In its long history, which benefited from the developments of statistics, there are in my opinion but three important giant steps for the regression analysis:

(1) development of the Principal Component Analysis (PCA) by Hotelling in 1933;[25]

(2) development of Partial Least Square Methodology initiated by H. Wold in 1974;[145] and

(3) development of Orthogonal Molecular Descriptors in 1991.[27−35]

These three approaches are complementary and should be combined whenever this is advantageous for structure−property−activity studies.

The PCA yields the number of statistically significant variables in a multiple regression analysis (given by the number of the leading eigenvectors of the correlation matrix). It produces linear combinations of the original descriptors (some of which may be statistically irrelevant although by an iterative procedure these can be eliminated) as the new variables. The new variables are mutually orthogonal, i.e., do not correlate among themselves. However, these eigenvectors do not offer a meaningful quantitative interpretation, and often even qualitative interpretation is too vague and somewhat arbitrary. There are two reasons for the failure of PCA to offer rigorous interpretation of the leading eigenvectors: (1) Even if one could interpret individual descriptors entering the analysis, there is neither simple nor elaborate interpretation for *linear combinations* of *ad hoc* descriptors contributing to the leading eigenvectors. (2) Even if the individual descriptors may have simple structural interpretation they are all mutually *interrelated*, often strongly interrelated, which makes it impossible to quantitatively assign the relative role of different contributions from the derived analysis.

This later objection holds also fully for multivariate regression analysis (MRA) and is reflected in the numerical instability of the regression equation. Two totally different regression equations using different descriptors may produce regressions which will have the same regression coefficient r and the same standard error s. Moreover, an already satisfactory regression may be improved in a stepwise regression by introducing an additional descriptor (when justified from the statistical point of view). While introduction of the additional descriptor will not dramatically alter the regression coefficient r and the standard error s it may dramatically change the coefficients of all the variables hitherto used. This well-known ill-behavior of the regression equations (which is numerically expressed by excessive probability errors for the coefficients of a regression equation that may exceed the numerical values of the coefficients themselves) has frustrated investigators ever since it was first recognized. Even though MRA remains a widely used data reduction method, particularly in QSAR,[9,10,147] a critic apparently lost patience proclaiming the following:

"In general, the results from this kind of approach are terrible. Correlation coefficients are excellent, computed standard deviations were very small, but when the regression equations were examined, it was found that the standard deviations of the regression coefficients were usually much larger than the values of the coefficients themselves. It was gradually recognized that such equations were useless for predictive purposes, and statistically invalid for correlating data."

We have already seen how this fatal deficiency of MRA has been eliminated by using orthogonalized descriptors. If the critic who has been alarmed by the instability of the regression equations is correct when suggesting to "abandon the ship," why do most of the traditional studies in QSAR continue to ignore the cry? If the critic is wrong (as I think that he or she is) the traditional studies in QSAR should continue, but to disarm the critic they ought to adopt the available remedy, the orthogonalized descriptors. MRA based on orthogonal descriptors not only is free of the fatal flaws associated with the numerical instability of the regression equations but also allows a meaningful interpretation of the results. It can be simply implemented by a stepwise regression approach which will yield all the coefficients entering the stable regression equation. This is not a place to counter argue with anonymous critics. If they wish to be taken seriously they should first come publicly with their criticisms and not use a disguise of anonymity as referees of papers that, despite their objections, appear in print. Apparently they failed to convince other referees and editors!

Critics and skeptics apparently have not made comment on orthogonal descriptors, while others continue to use the "flawed" old approach, that does not lead to interpretation of the regression equation. Why? Some are apparently

unaware of the development of MRA using orthogonal descriptors and are likely to adopt the approach as soon as they became familiar with the technique. Others may be reluctant for different reasons. The orthogonalized MRA force one to prioritize the descriptors. One of the disadvantages of the traditional QSAR, as practiced by Hansch and his followers, is that there is no natural preference for ordering the variables used in MRA. Orthogonalization process (just as the stepwise regression analysis or Gram-Schmidt orthogonalization of vectors in Linear Algebra) requires ordering of variables, such as log *P*, Taft's sigma, etc. It may be difficult to convincingly argue which of the traditional descriptors "deserves" to be the first, which to be the second, third, etc. However this dilemma can be overcome by adopting a pragmatic approach: Examine first all the descriptors in a pool (to be used) as a *single* variable in simple regression analysis. Select as the *first* descriptor one which gives the best statistical parameters for the simple regression (i.e., the best correlation coefficient, the best standard error, or the best Fisher ratio, *F*, depending on your preference). In the next step search for the *second* best descriptor using the already selected descriptor in a multiple regression analysis. The process continues till all the descriptors considered are exhausted (providing the statistics justifies use of that many descriptors). This approach, which represents use of the greedy algorithm approach in making the decision, does not ensure that the final set is necessarily the optimal, but the final set is well defined and will offer reasonable solution. This approach (i.e., so performed stepwise regression) has been referred to as DCA, the Dominant Component Analysis when the partial steps are interpreted as steps for construction of regression equation corresponding to orthogonal descriptors.[36] The implied final variables in such stepwise regression are mutually orthogonal, just as in PCA and PLS. However, in contrast to PCA and PLS there are no ambiguities in the *interpretation* of the final variable, because they are constructed as a linear combination of orthogonalized descriptors.

Critics of MRA, and some of the promoters of PCA and PLS, argue that "due to the statistical limitations of regression analysis, one biological activity variable at a time is usually analyzed" [148] as if this represents a serious limitation of MRA. One can argue also to the opposite, that it is an advantage of MRA that it is confined to a single biological activity. When considering several molecular properties jointly often the factors that influence data collection vary with the property considered. Similarly the sensitivity and the accuracy of data (i.e., the quality of data) pertaining to different activities may vary considerably. For example, atom−atom distances are measured to different accuracy in X-ray experiments, in microwave spectroscopy, in NMR, etc. Moreover not only that they are measured with different accuracy but also the interatomic distances determined by different methodologies are often inherently conceptually distinct even if numerically similar.

CONCLUDING REMARKS

"...a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows that is familiar with it ..." Max Planck[1]

Clearly in different situations different approaches may have an advantage, and rather than wasting time in praising one approach over another all valid statistical schemes should be used. In summary, when studying complex QSAR problems we may say that chemical graph theory, mathematical chemistry in general, and chemometrics, all of which strongly overlap and often cover much of the same territory, are likely to approach the complex phenomena of structure−property−activity relationship from somewhat individualistic positions. Traditional QSAR school will continue to be traditional, topological indices will continue to be tested on different families of compounds, often hydrocarbons, PCA will continue to struggle with interpretation of the principal components, and PLS will continue to be promoted as panacea, as if nobody recollects the saying of Benjamin Franklin: "We must all hang together, or assuredly we shall all hang separately!".

Our efforts ought to be directed not in the "fight" of one method against another but as a fight of all methods against the complex phenomena that dominate the structure−activity and drug design studies. Novel techniques are emerging. Yesterday it was neural networks,[149,150] genetic algorithms,[151−153] and cell automata;[154−158] tomorrow may be something else. New methodologies will continue to emerge, often not to displace the existing schemes but rather to supplement them. The possibility of using orthogonalized descriptors combined with other available approaches is likely to upgrade and empower the existing tool. Use of orthogonal descriptors in structure-similarity has pointed to the importance of having strictly independent descriptors.[159] It is to be expected that in short time we will see use of orthogonalized descriptors in neural networks, in pattern recognition, and even in PCA and PLS. This "injection" of orthogonal descriptors in these widely used methods can only sharpen their power not only by making the interpretation of the results easier but also possibly simplifing some of the "hidden" computational steps, which may result in "transferability" of some molecular descriptors. We will, hopefully, soon find out if indeed these optimistic expectations are founded or not as the number of researches who use orthogonal descriptors apparently grows.

ACKNOWLEDGMENT

I would like to thank Professor A. T. Balaban (Bucharest, Romania) for his comments that lead to an improved presentation of the material.

REFERENCES AND NOTES

(1) Planck, M. *Survey of Physical Theory*; Dover Publ.: New York, 1960.
(2) Platt, J. R. Prediction of isomeric differences in paraffin properties. *J. Chem. Phys.* **1952**, *56*, 328−336.
(3) *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976.
(4) Randić, M. Characterization of atoms, molecules and classes of molecules based on paths enumerations. MATCH, **1979**, *7*, 5−64.
(5) Trinajstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, Fl, 1992.
(6) Randić, M.; Wilkins, C. L. Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. *J. Phys. Chem.* **1979**, *83*, 1525−1540.
(7) Hosoya, H. Topological index. A newly proposed quantity characterizing topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(8) Randić, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(9) Kier, L. B.; Hall, H. L. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(10) Kier, L. B.; Hall, H. L. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: Letchworth, 1986.

(11) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(12) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V: Connectivity series concept applied to density. *J. Pharm. Sci.* **1976**, *65*, 1226−1230.
(13) Randić, M.; Basak, S. C. Search for empirical descriptors for heteroatoms in QSAR. Work in progress.
(14) Randić, M. On computation of optimal parameters for multivariate analysis of structure−property relationship. *J. Comput. Chem.* **1991**, *12*, 970−980.
(15) Randić, M.; Dobrowolski, J. Cz. Optimal molecular connectivity descriptors for nitrogen containing molecules. *J. Math. Chem.* Submitted for publication.
(16) Balaban, A. T. Chemical graphs. 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122.
(17) Randić, M. Linear combination of path numbers as molecular descriptors. *New J. Chem.* Submitted for publication.
(18) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava, V. K.; Trinajstić, N. On the distance matrix of molecules containing heteroatoms. In *Chemical Applications of Topology and Graph Theory*; King R. B., Ed.; Elsevier: Amsterdam, 1983; pp 222−230.
(19) Timmermans, B. M. W. M.; van Zweiten, P. A. Quantitative structure−activity relationship in centrally acting imidazolidines structurally related to clonidine. *J. Med. Chem.* **1977**, *20*, 1636−1644.
(20) Randić, M. Novel graph theoretical approach to heteroatoms in quantitative structure−activity relationships. *Chemometrics Intel. Lab. Systems* **1991**, *10*, 213−227.
(21) Topliss, J. G.; Edwards, P. Chance factors in studies of quantitative structure−activity relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(22) Randić, M. Compact molecular descriptors. *Acta Pharm.* Submitted for publication.
(23) Kohn, M. C. Strategies for computer modeling. *Bull. Math. Biol.* **1986**, *48*, 417−426.
(24) Randić, M. Comparative regression analysis. Regressions based on a single descriptor. *Croat. Chem. Acta* **1993**, *66*, 289−312.
(25) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417−441 and 489−520.
(26) Needham, D. E.; Wei, I.-C.; Seybold, P. G. Molecular modeling of the physical properties of the alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186-4194.
(27) Randić, M. Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−320.
(28) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517−525.
(29) Randić, M. Search for Optimal Molecular Descriptors. *Croat. Chem. Acta* **1991**, *64*, 43−54.
(30) Randić, M. Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct. (Theochem)* **1991**, *233*, 45−59.
(31) Randić, M. Chemical Structure—What is she? *J. Chem. Educ.* **1992**, *69*, 713−718.
(32) Randić, M.; Seybold, P. G. Molecular shape as a critical factor in structure−property−activity studies. *SAR and QSAR* **1993**, *1*, 77−85.
(33) Randić, M.; Trinajstic, N. Isomeric variations in alkanes: boiling points of nonanes. *New J. Chem.* **1994**, *18*, 179−189.
(34) Randić, M. Fitting of nonlinear regressions by orthogonalized power series. *J. Comput. Chem.* **1993**, *14*, 363−370.
(35) Randić, M. Curve-fitting paradox. *Int. J. Quant. Chem: Quant. Biol. Symp.* **1994**, *21*, 215−225.
(36) Randić, M. Similarity methods of interest in chemistry. In *Mathematical Methods in Contemporary Chemistry*; Kuchanov, S. I., Ed.; Gordon & Breach: Amsterdam, 1996.
(37) Randić, M. Quantitative structure-property relationship. Boiling points of planar benzenoids. *New J. Chem.* **1996**, *20*, 1001−1009.
(38) Randić, M.; Razinger, M. On characterization of three-dimensional molecular structure. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 159−236.
(39) Pogliani, L. Modeling with special descriptors derived from a medium-sized set of connectivity indices. *J. Phys. Chem.* **1996**, *100*, 18065−18077.
(40) Pogliani, L. Modeling purines and pirimidines with the linear combination of connectivity indices. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1082−1091.
(41) Pogliani, L. A Strategy for molecular modeling of a physicochemical property using linear combination of connectivity indexes. *Croat. Chem. Acta* **1996**, *69*, 95−109.
(42) Pogliani, L. Molecular modeling by linear combinations of connectivity indices. *J. Phys. Chem.* **1995**, *99*, 925−937.
(43) Pogliani, L. Molecular connectivity descriptors of the physicochemical properties of the α-amino acids. *J. Phys. Chem.* **1994**, *98*, 1494−1499.
(44) Pogliani, L. Structure property relationships of amino acids and some dipeptides. *Amino Acids* **1994**, *6*, 141−153.
(45) Pogliani, L. *Curr. Top. Pept., Prot. Res.* **1994**, *1*, 119.
(46) Pogliani, L. Molecular connectivity model for determination of physicochemical properties of α-amino acids. *J. Phys. Chem.* **1993**, *97*, 6731−6736.
(47) Pogliani, L. Molecular connectivity: treatement of electronic structure of amino acids. *J. Pharm. Sci.* **1992**, *81*, 967−969.
(48) Šoškić, M.; Plavšić, D.; Trinajstic, N. 2-Difluoromethylthio-4,6-bis-(monoalkylamino)-1,3,5-triazines as inhibitors of Hill reaction: A QSAR study with orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 146−150.
(49) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D. The structure-property models can be improved using orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532−538.
(50) Amić, D.; Davidović-Amić, D.; Jurić, A.; Lučić, B.; Trinajstić, N. Structure-Activity correlation of flavone derivarives for inhibition of cAMP phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034−1038.
(51) Amić, D.; Davidović-Amić, D; Trinajstić, N. Calculation of retention times of anthocyanins with orthogonalized topological indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 136−139.
(52) *Mathematical Chemistry*; Klein, D. J., Randić, M., Eds.; VCH: Weinheim, 1990.
(53) Randić, M. In search of structural invariants. *J. Math. Chem.* **1992**, *9*, 97−146.
(54) Randić, M.; Trinajstić, N. In search for graph invariants of chemical interest. *J. Mol. Struct. (Theochem)* **1993**, *300*, 551−571.
(55) Randić, M. Graphical enumeration of conformations of chains. *Int. J. Quant. Chem: Quant. Biol. Symp.* **1980**, *7*, 187−197.
(56) Zhu, H.-Y.; Klein, D. J. Graph-geometric invariants for molecular structure. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1067−1075.
(57) Tratch, S. S.; Stankevich, M. I.; Zefirov, N. S. Combinatorial models and algorithms in chemistry. The expanded Wiener Number—A novel topological index. *J. Comput. Chem.* **1990**, *11*, 899−908.
(58) Hall, L. H. Computational aspects of molecular connectivity anbd its role in structure-property modeling. *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Scientific: New York, 1990; pp 202−233.
(59) Randić, M. Novel molecular descriptor for structure−property studies. *Chem. Phys. Lett.* **1993**, *211*, 478−483.
(60) Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H. Wiener matrix: Source of novel graph invariants. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 709−716.
(61) Klein, D. J.; Randić, M. Resistance distance. *J. Math. Chem.* **1993**, *12*, 81−95
(62) Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H.; Naylor, L. Wiener matrix invariants. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 361−367.
(63) Randić, M. Hosoya matrix - a source of new molecular descriptors. *Croat. Chem. Acta* **1994**, *67*, 415−429.
(64) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277−286. (For errata see ref 133).
(65) Randić, M. Restricted random walks on graphs. *Theor. Chim. Acta* **1995**, *92*, 97−106.
(66) Amić, D.; Trinajstić, N. On the detour matrix. *Croat. Chem. Acta* **1995**, *68*, 53−62.
(67) Jiang, Y.; Zhu, H. Evaluation of pattern indices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 377−380.
(68) Trinajstić, N.; Babić, D.; Nikolić, S.; Plavšić, D.; Amić, D.; Mihalić, Z. The Laplacian matrix in chemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 368−376.
(69) Plavšić, D.; Nikolić, S.; Trinajstić, N.; Mihalić, Z. On the Harary index for characterization of chemical graphs. *J. Math. Chem.* **1993**, *12*, 235−250.
(70) Estrada, E. Graph theoretical invariant of Randić revisited. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1022−1025.
(71) Yao, Y. Y.; Xu, L.; Yang, Y. Q.; Yuan, X. S. Study on structure−activity relationship of organic compounds. Three new topological indices and their application. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 478−483.
(72) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of topological indices. Part 4. Reciprocal distance matrix related local vertex invariants and topological indices. *J. Math. Chem.* **1993**, *12*, 309−318.
(73) Kier, L. B.; Hall, L. H. An atom-centered index for drug QSAR models. In *Advances in Drug Design*; Testa, B.; Ed.; Academic Press: 1992; Vol. 22.
(74) Estrada, E. Edge adjacency relationships and novel topological index related to molecular volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31−33.
(75) Xu, L.; Tao, Y. T.; Wang, H.-M. New topological index and prediction of phase transfer for protonated amines and tetraalkylamine ions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 45−49.

(76) Schultz, H. P. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 277−288.
(77) Müller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstić, N. Molecular topological index. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 160−163.
(78) Lukovits, I.; Linert, W. A novel definition of hyper-Wiener index for cycles. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 899−902.
(79) Khadikar, P. V.; Deshpande, N. V.; Kale, P. P.; Dobrynin, A.; Gutman, I; Domotor, G. The Szeged index and an analogy with Wiener index. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 547−550.
(80) Lukovits, I. The generalized Wiener index for molecules containing double bonds and the partition coefficients. *Rep. Mol. Theory* **1990**, *1*, 127−131
(81) Klein, D. J.; Lukovits, I.; Gutman, I. On the definition of the hyper-Wiener index for cycle-containing structures. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 50−52.
(82) Diudea, M. V. Wiener and hyper-Wiener numbers in a single matrix. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 833−836.
(83) Mohar, B.; Babić, D.; Trinajstić, N. A novel definition of the Wiener index for trees. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 153−154.
(84) Maier, B. J. Wiener and Randić topological indices for graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 87−90.
(85) Herman, A.; Zinn, P. List operations on chemical graphs. 6. Comparative study of combinatorial topological indexes of the Hosoya type. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 551−560.
(86) Randić, M.; Morales, D. A.; Araujo, O. Higher order Fibonacci numbers. *J. Math. Chem.* **1996**, *20*, 79−94.
(87) Plavšić, D.; Šoškić, M.; Landeka, I.; Gutman, I.; Graovac, A. On the relation between the path numbers ^1Z, ^2Z and the Hosoya Z index. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1118−1122.
(88) Stakhov, A. P. *Introduction into algorithmic measurement theory*; Soviet Radio Publ.: Moscow, 1977.
(89) Yang, Y.-Q.; Xu, L.; Hu, C.-Y. Extended adjacency matrix indices and their applications. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1140−1145.
(90) Diudea, M. V. Molecular topology. 16. Layer matrices in molecular graphs, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1064−1071.
(91) Diudea, M. V. Molecular topology. 17. Layer matrices of Walk degrees. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1072−1078.
(92) Kirby, E. C. Sensitivity of topological indices to methyl group branching in octanes and azulenes, or what does a topological index index? *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1030−1035.
(93) Balaban, A. T. Local versus global (i.e., atomic versus molecular) numerical modeling of molecular graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 398−402.
(94) Balaban, A. T. Lowering the intra- and intermolecular degeneracy of topological invariants. *Croat. Chem. Acta* **1993**, *66*, 447−458.
(95) Balaban, A. T. Using real numbers as vertex invariants for third generation topological indices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 28−23.
(96) Randić, M. Representation of molecular graphs by basic graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57−69.
(97) Baskin, I. I.; Gordeeva, E. V.; Devdariani, R. O.; Zefirov, N. S.; Palyulin, V. A.; Stankevich, M. I. Solving the inverse problem of structure-property relations for the case of topological indices. *Dokl. Akad. Nauk. USSR* **1989**, *307*, 613−617.
(98) Zefirov, N. S.; Palyulin, V. A.; Ranchenko, E. V. Problem of generation of structures with specified properties, solution of the inverse problem for Balaban Centric index. *Dokl. Acad. Nauk SSSR* **1991**, *316*, 921−924.
(99) Skvortosova, M. L.; Baskin, I. I.; Slovokhtova, O. L.; Palyulin, V. A.; Zefirov, N. S. Inverse problem in QSAR/QSPR studies for the case of topological indices characterizing molecular shape (Kier indices). *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 630−634.
(100) Hall, L. H.; Kier, L. B. Generation of molecular structure from a graph based QSAR equation. *Quant. Struct.-Act. Relat.* **1993**, *12*, 60−66.
(101) Kier, L. B.; Hall, L. H.; Frazer, J. W. Design of molecules from quantitative structure-activity relationship models. 1. Information transfer between path and vertex degree counts. *J. Chem. Inf. Comput. Sci.* **1993**, *363*, 143−147.
(102) Kier, L. B.; Hall, L. H.; Frazer, J. W. Design of molecules from quantitative structure-activity relationship models. 2. Derivation and proof of information transfer relating equations. *J. Chem. Inf. Comput. Sci.* **1993**, *363*, 148−152.
(103) Kier, L. B.; Hall, L. H.; Frazer, J. W. Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: Path 3. *J. Chem. Inf. Comput. Sci.* **1993**, *363*, 598−603.
(104) Kvasnička, V.; Pospichal, J. Simulated annealing construction of molecular graphs with required properties. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 516−526.
(105) Lukovits, I. Toward reconstruction of trees by using graph invariants. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 287−289.
(106) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Evolutionary design of molecules with desired properties using genetic algorithm. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 188−195.
(107) Randić, M.; Plavšić, D.; Razinger, M. Double invariants. *Math. Chem (MATCH)* In press.
(108) Randić, M. On molecular branching. *Acta Chim. Sloven.* In press.
(109) Randić M. *J. Math. Chem.* Submitted for publication.
(110) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indices vs. electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.
(111) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph theoretic indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.
(112) Basak, S. C.; Grunwald, G. D. Molecular similarity and estimation of molecular properties. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 366−372.
(113) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054−1060.
(114) Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformations*; Research Studies Press/Wiley: Taunton, Sommerset, England, 1988.
(115) Ghose, A. K.; Crippen, G. M. Distance geometry QSAR: Current achievements and future extensions. In *QSAR and Strategies in the Design of Bioactive Compounds*; Seydel, J. K., Ed.; VCH: 1985; pp 116−119.
(116) Balasubramanian, K. Geometry-dependent characteristic polynomial of molecular structures. *Chem. Phys. Lett.* **1990**, *169*, 224−228.
(117) Randić, M. Molecular topographic descriptors. *Stud. Phys. Theor. Chem.* **1988**, *54*, 101−108.
(118) Randić, M. On the characterization of three-dimensional structures. *Int. J. Quant. Chem: Quant. Biol. Symp.* **1988**, *15*, 201−208.
(119) Randić, M.; Jerman-Blažič, B.; Trinajstić, N. Development of 3-dimensional molecular descriptors. *Comput. Chem.* **1990**, *14*, 237−246.
(120) Bogdanov, B. On the three-dimensional Wiener number. *J. Math. Chem.* **1989**, *3*, 299−309.
(121) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the three-dimensional Wiener number. A comment. *J. Math. Chem.* **1990**, *5*, 305−306.
(122) Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavšić, D.; Trinajstić, N. The distance matrix in chemistry. *J. Math. Chem.* **1992**, *11*, 223.
(123) Pogliani, L. On a graph theoretical characterization of cis/trans isomers. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 801−804.
(124) Diudea, M. V.; Horvath, D.; Graovac, A. Molecular topology. 15. 3D distance matrices and related topological indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 129−135.
(125) Estrada, E. Three-dimensional molecular descriptors based on charge density weighted graphs. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 708−713.
(126) Nikolić, S.; Trinajstić, N.; Mihalić, Z.; Carter, S. On the geometric distance matrix and the corresponding structural invariants of molecular systems. *Chem. Phys. Lett.* **1991**, *176*, 21−28.
(127) Schultz, H. P.; Schultz, E. B.; Schultz, T. P. Topological organic chemistry. 10. Graph theory and topological indices of conformational isomers. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 996−1000.
(128) Randić, M.; Razinger, M. Molecular topographic indices. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 140−147.
(129) Randić, M. Molecular shape profiles. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373−382.
(130) Randić, M. Molecular profiles. Novel geometry-dependent molecular descriptors. *New J. Chem.* **1995**, *19*, 781−791.
(131) Randić, M. On characterization of the conformations of nine-membered rings. *Int. J. Quant. Chem: Quant. Biol. Symp.* **1995**, *22*, 61−73.
(132) Randić, M. Molecular bonding profiles. *J. Math. Chem.* **1996**, *19*, 375−392.
(133) Randić, M. Quantitative structure−activity relationship. Boiling points of planar benzenoids. *New J. Chem.* **1996**, *20*, 1001−1009.
(134) Randić, M.; Krilov, G. Bond profiles for cuboctahedron and twist cuboctahedron. *Int. J. Quant. Chem: Quant. Biol. Symp.* **1996**, *23*, 1851−1863.
(135) Randić, M.; Krilov, G. On characterization of molecular surface. *Int. J. Quant. Chem: Quant. Biol. Symp.* **1997**, *24* In press.
(136) Randić, M. Fullerene profiles; Fullerene Sci., Techn. Submitted for publication.
(137) Ružička, L.; Prelog, V. *Helv. Chim. Acta* **1944**, *27*, 66.
(138) Lovasz, L.; Pelikan, J. On the eigenvalue of trees. *Period. Math. Hung.* **1973**, *3*, 175−182.
(139) In ref 63 in Table 2 and Table 3 the labels CCC and CTC have to be exchanged.
(140) Li, H.; Helling, R.; Tang, C.; Wingreen, N. Emergence of preferred structures in a simple model of protein folding. *Science* **1996**, *273*, 666−669.

(141) Borman, S. Protein folding model focuses on "designability," *Chem., Eng. News*, **1966**, August 12, p 36.
(142) Kardar, M. Which came first, protein sequence or structure? *Science* **1966**, *273*, 610.
(143) Randić, M.; Krilov, G. On Characterization of 3-D structure of proteins. *Chem. Phys. Lett.* In press.
(144) Randić, M.; Krilov, G. On characterization of the folding of proteins. Presented at 7th International Conference on Mathematical Chemistry. Girone, Spain, May 26−31.
(145) Randić, M. *Topological indices*. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Editor-in-Chief; J. Wiley and Sons: London, In press.
(146) Wold, H. *Eur. Econ. Rev.* **1974**, *5*, 67
(147) Hansch, C.; Leo, A. *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, D.C., 1995.
(148) Skagerberg, B.; Dunn III, W. J.; Hellberg, S.; Wold, S. The PLS data analytic method in QSAR. In *QSAR and Strategies in the Design of Bioactive Compounds*; Seydel, J. K., Ed.; VCH: 1985; pp 305−310.
(149) Zupan, J.; Geisteger, J. Neural Networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **1991**, *248*, 1−30.
(150) Zupan, J.; Geisteger, J. *Neural Networks for Chemists*; VCH Publishers: New York, 1993.
(151) Holland, J. *Adaptation in Natural and Articial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
(152) Devillers, J. Genetic algorithms in computer-aided molecular design. *Genetic Algorithms in Molecular Modeling*; Academic Press: Devillers, J., Ed.; London, 1996; pp 131−157.
(153) Hopfinger, A. J.; Patel, H. C. Application of genetic algorithms to the general QSAR problem and to guiding molecular diversity experiments. *Genetic Algorithms in Molecular Modeling*; Academic Press: Devillers, J., Ed.; London, 1996; pp 131−157.
(154) Kier, L. B.; Cheng, C.-K. A cellular automata model of water. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 647−652.
(155) Kier, L. B.; Cheng, C.-K. A cellular automata model of aqueous solution. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1334−1337.
(156) Kier, L. B.; Cheng, C.-K. A cellular automata model of hydrophobic effect. *Pharm. Res.* **1995**, *12*, 615−620.
(157) Kier, L. B.; Cheng, C.-K. A cellular automata model of disolution. *Pharm. Res.* **1995**, *12*, 1521−1525.
(158) Cheng, C.-K.; Kier, L. B. A cellular automata model of oil-water partitioning. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1054−1059.
(159) Randić, M. Orthosimilarity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1092−1097.

CI960174T