Resolution of Ambiguities in Structure–Property Studies by Use of Orthogonal Descriptors[†]

MILAN RANDIĆ

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

Received August 8, 1990

We initiate here a systematic analysis of molecular descriptors, the quantities used to represent molecules in structure–property and structure–biological activity studies. We illustrate the approach with Hosoya's Z index, which is based on the enumeration of disjoint edges in molecular graphs. The purpose of this analysis is to discern similarities and differences among the molecular descriptors used in regression analysis. In this way, we hope to facilitate the rational use and perhaps the rational design of such descriptors. The main tool in this analysis is the recently described methodology of construction of orthogonalized molecular descriptors in multivariate regression analysis. The analysis points to critical parts of descriptors which are responsible for their performance in a regression.

INTRODUCTION

Studies of relationships between structure and properties or between structure and activity, the latter relating to biological responses of chemicals, are of central interest in various branches of chemistry and even physics, including particularly physical chemistry, medicinal chemistry, chemometrics, chemical graph theory, mathematical chemistry, chemical physics, computational chemistry, mathematical modeling, computer-assisted manipulation of structures, chemical documentation, artificial intelligence, and so on. Various aspects of structure–property and structure–activity studies are of interest in different applications. In chemical documentation and computer retrieval of structures, for example, it is critical that structures are represented by unique codes, while for computer-assisted manipulation, it is desirable that codes be operational, i.e., that meaningful alterations of codes lead to other structures or intermediates of interest. We are here interested primarily in molecular descriptors that are useful in structure–property and structure–activity regression studies. Questions of interest in this connection include:

1. How and why do similar structures exhibit similar properties?
2. How can structures be characterized quantitatively?
3. How can searches be carried out for an optimal structure within a family of structures?
4. How can new "lead" structures be identified?
5. How can the "best" representation for a structure be determined?
6. How can the *pharmacophore* (the critical local molecular fragment) that is responsible for a given molecular activity be identified?
7. How can a partial order, or hierarchy, be derived for a set of structures?
8. How can a global property be partitioned within an additive model?

It should be observed that answering such questions extends beyond the capabilities of quantum chemistry. Quantum chemistry provides insight into the electronic "nature of the chemical bond", discussed by Linus Pauling in his celebrated book of the same title.[1] Quantum chemical calculations may help to clarify important mechanistic questions involving drug–receptor interactions and the nature of various metastable transition states. Questions pertaining to "the nature of chemical structure"[2] belong to chemical graph theory[3] which is concerned with the identification, from comparative studies of several molecules simultaneously, of the crucial structural factors involved in structure–property relationships. In this paper, we continue such studies, with an emphasis, however, on a *systematic* study of the molecular descriptors employed in research into the relationships between structure, property, and activity.

Graph theoretical characterizations of structures often allow them to be ranked, provided that standards—e.g., structures with known properties—are available. Graph theory alone usually does not produce numerical values for molecular properties; information on selected properties must be supplied, at least for a few standard structures, upon which comparisons can be based. Such standards serve as "target compounds", the compounds that other compounds ought as much as possible to imitate, mimic, follow, or approach, and most desirably, surpass in quality, the properties of interest. Successful filtering will reduce a relatively large pool of 50–100 structures to a few structures, say half a dozen, which *qualify* as *survivors* in which a desirable property (e.g., therapeutic potency or lack of toxicity) is somewhat enhanced. Graph theory does not normally resolve cause-and-effect relationships although, by focusing attention on selected structural features, it may suggest such relationships.[4]

Our prime purpose is to *develop* methods for the analysis of molecular descriptors, including properties used as descriptors, and perform a *critical analysis* of their role on regression. In this paper, we propose a method that allows one to investigate if a descriptor is of potential interest in a structure–property–activity study or if it can be discarded as irrelevant, i.e., unable to account for important structural features, or as redundant, i.e., duplicative of information available from other descriptors.

CLASSIFICATION OF MOLECULAR DESCRIPTORS

Structure–property–activity studies can conveniently be classified as structure-cryptic, structure-implicit, and structure-explicit.[5]

In **structure-cryptic methods**, selected molecular properties can be used in attempts to express more complicated properties, such as biological activities, in terms of simpler and better known properties. Many molecular properties are interrelated and such interrelationships can be quantitatively studied with advanced statistical methods. Cramer,[6] for example, has pursued this approach on a diverse set of compounds with properties such as aqueous solvation, octanol–water partition, molar refraction, boiling points, molar volumes, and heats of vaporization. Principal component analysis has shown that two dominant components account for most of the observed variations. The structural features that are involved are not necessarily easily identified, but it is clear that the above six properties depend on the *same* structural factors. Such ap-

[†] Dedicated to Herbert Hauptman (Nobel Prize in Chemistry, 1985) for making mathematical chemistry visible.

0095-2338/91/1631-0311$02.50/0 © 1991 American Chemical Society

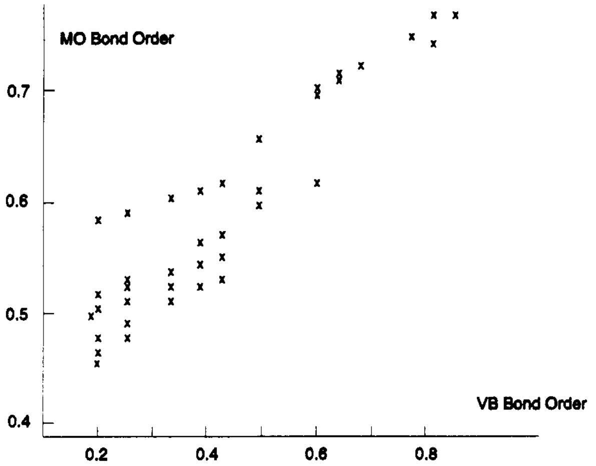


Figure 1. Plot of Coulson's MO bond orders against Pauling's VB bond orders for a number of smaller benzenoids.

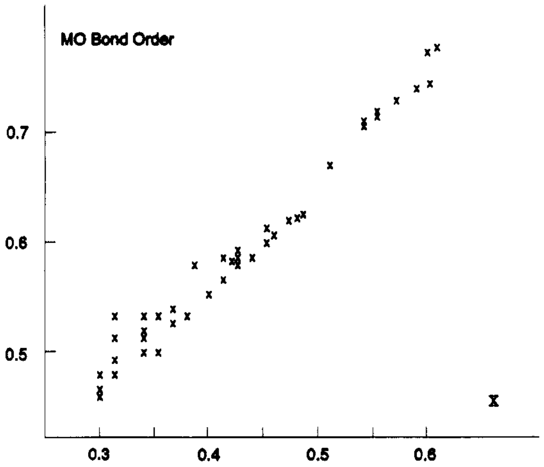


Figure 2. Plot of Coulson's MO bond orders against the connectivity bond weights (i.e., bond contributions to the connectivity index) for the same set of smaller benzenoids used in Figure 1.

plications of factor analysis belong to the structure-cryptic type of analysis,[5] which also include the traditional quantitative structure–activity relationship (QSAR) studies.[7] A classical illustration of a property–property relationship is Kováts' correlation[8] between the boiling points and the chromatographic retention volumes exhibited by alkanes. This demonstrated the "equivalence", in alkanes at least, of chromatographic parameters and thermodynamic characteristics such as boiling points. Such methods have been shown to have some utility in the prediction of structures possessing specific desired properties, but they do not identify or utilize the structural components that are dominant for the property in question and so they are termed structure-cryptic. The use of such schemes leads, at best, to property–property correlations. Some properties are better known than others, and consequently it is useful to be able to express less familiar phenomena in terms of data that are better known and better understood.

Structure-implicit methods are exemplified by quantum chemical computations which, being geometry-dependent, involve structure indirectly. As a result of such computations, it is possible to obtain—very precisely in some cases—data pertaining to a molecule. Some of the computed properties of a molecule can be processed, i.e., partitioned in various components, to assist in the building of a model of the chemical structure. Quantum chemical computations shed no light on the reasons for the magnitude of a particular computed property, e.g., the short length of the central carbon–carbon bond in naphthalene, and will not identify the structural factors that may be critical to this bond length. The accuracy of the prediction of a bond length will be determined by the approximations that are used and in principal, in the limit of ambitious ab initio calculations, a unique result is anticipated. Such analysis leads to structure–structure relationships, and to illustrate this, there is shown in Figure 1 a correlation between bond orders in smaller conjugated hydrocarbons (precursors to computed bond lengths) computed by molecular orbital methods and those calculated for the same species by valence bond methods. It has been seen that property–property correlation can afford some indirect insights into structure and that, similarly, structure–structure correlation can sometimes provide some indirect insights into properties and are of considerable interest in modeling of chemical structures. The computed bond orders may be regarded as mathematical properties of a molecule, and then the distinction between property–property and structure–structure correlations becomes semantic. It should be noted that mathematical properties, such as bond orders, are hypothetical (i.e., nonobservable) while quantum chemistry can supply observable properties derived from the total wave function for a system.[9] Nonobservable quantities such as bond orders, potential curves, hybrids, hybridization, Kekulé valence structures, molecular orbitals, and associated parameters, such as HOMO and LUMO energies, and other commonly used concepts of theoretical chemistry offer a bridge to chemical graph theory, and may in some situations make the distinction between the two disciplines ambiguous, if not arbitrary.

Structure-explicit methods are typified by approaches developed in chemical graph theory and other disciplines of mathematical chemistry. The descriptors that are used here have defined structural interpretations, and efforts are made to relate molecular properties to selected descriptors of this sort. When a good correlation is found with a single descriptor, the claim is usually made that the structural feature encoded by the descriptor is *relevant*, i.e., it accounts for the dominant part of the correlation. There is no cause-and-effect relationship, merely a parallelism. Descriptors however have a direct structural interpretation, and as a consequence, it becomes possible to advance structural models and design structures that may have desirable properties.

An illustration of a graph theoretical correlation is shown in Figure 2 in which MO bond orders, as defined by Coulson,[10] are plotted against the bond contribution to the connectivity index.[11] The connectivity index[12] X for a molecule is defined as a bond additive property in which bonds are first classified as various (m,n) bond types, where m and n indicate the valencies, i.e., the number of immediate neighbors of the atoms forming the bond. Each bond makes its own contribution of $(m,n)^{-1/2}$. These weights simulate the relative "importance" of different types of bonds; terminal bonds play a larger role than a bond buried inside a molecule. The weighting algorithm was derived as a solution to the system of inequalities based on the *ordering* of isomers of smaller alkanes that parallels the relative magnitudes of their boiling points. It may be seen that the scatter of points in Figure 2 is visibly less than that in Figure 1.

Why should an apparently arbitrarily defined quantity, such as the bond connectivity index, correlate so well with a quantum chemically defined quantity, and certainly better than a quantum chemical quantity correlates with another quantum chemical descriptor, such as the Pauling bond order (illustrated in Figure 1)? Coulson bond orders have well-justified interpretations based upon bond overlap, but they do not tell us why the central bond in naphthalene is the shortest. The computed orbital overlaps add significantly in shorter bonds, but why their *coefficients* are the largest is hidden in the intricacies of an eigenvalue problem. In the valence bond picture, the fact that some bonds have a high C-C double bond character is conventionally rationalized by observing that such bonds are more often represented as double bonds in the collection of Kekulé valence structures. In large polycyclic structures, however, it is not obvious which bonds will have a large bond order and which will have a small bond order, and the computational complexity of the determination of Pauling bond order in very large systems can be prohibitive. Buckminsterfullerene, a highly symmetrical cage structure with 60 carbon atoms, was found by Klein et al.[13] to possess some 12 500 Kekulé structures! Clearly, in this case, determination of bond orders requires novel tools such as the transfer matrix approach.[14] The graph theoretical recognition of short C-C bonds is simple: in naphthalene the short internal bond is classified as a (3,3) type and the external bonds are classified either as (2,3) or (2,2) types. Why (3,3) bonds should be shorter remains a quantum chemical problem but the *simple* structural feature, the number of nearest neighbors for each atom forming a bond, is the critical structural factor that determines the bond types and, consequently, the bond lengths. Such observations lead in chemical graph theory to the development of models that relate more complicated properties to structurally simpler ones.

The classification of descriptors as structure-cryptic, -implicit, or -explicit is useful because it emphasizes one of the dominant aspects of distinct approaches to chemical structure, but it should not be viewed as absolute. The distinction between some descriptors is less clear than it may appear. Pauling bond orders, for example, can be viewed equally as graph theoretical quantities or as quantum chemical properties. In fact the concepts of graph "factors" and "perfect matching",[15] which are the mathematicians' equivalents of Kekulé valence structures,[16] have been developed independently in graph theory, just as the concepts of alternants (in conjugated hydrocarbons) and bipartite graphs (in mathematics) are equivalent. Many intimate relationships between graph theoretical and quantum chemical concepts are known. These include the subtle relationship between MO bond orders and VB bond orders, as delineated by Ruedenberg, Ham, and Platt,[17] a relationship between the determinant of Hückel Molecular orbital (HMO) and the number of Kekulé valence structures indicated by Dewar and Longuet-Higgins,[18] the relationship between the coefficients of nonbonding molecular orbitals (NBMO) and the number of Kekulé valence structures for related systems, pointed out by Platt,[19] and the relationship, discussed by Ruedenberg[20] and Heilbronner,[21] between inverse adjacency matrices and Kekulé valence structures.

Apparently diverse descriptors can be combined in specific applications, but it is not obvious how a choice between alternative descriptors might be made or how duplications resulting from the use of multiple descriptors might be avoided. These questions, which have not generally been addressed to date, motivated our interest in a systematic analysis of molecular descriptors. The dilemmas have been summarized by Kohn:[22]

> "Modeling is a ubiquitous and often misunderstood enterprise in which data from diverse disciplines are analyzed by techniques from other diverse disciplines in an attempt to confirm or falsify a set of hypotheses about the real world."

This article, rich in guidelines for model design, including requirements for the credibility of models and suggestions for their validation, concludes with:

> "It is unusual for only one model to be compatible with experimental observations. Often the data are not sufficiently extensive to discriminate amongst rival models and new experiments must be designed to answer outstanding questions. The statistical, graph theoretical, and sensitivity analysis methods ... can identify the areas for further investigation that are likely to produce significant new results."

Hence graph theory is recognized as a desirable approach to model building. We will restrict our efforts here to the *upgrading* of graph theoretical descriptors. The task appears to be feasible as a consequence of the availability of a novel tool for multivariate regression analysis: orthogonalization of descriptors.[23] This allows clarification of the ambiguities that have plagued earlier applications of multivariate regression analysis, whether they used descriptors of graph theoretical origin or not.

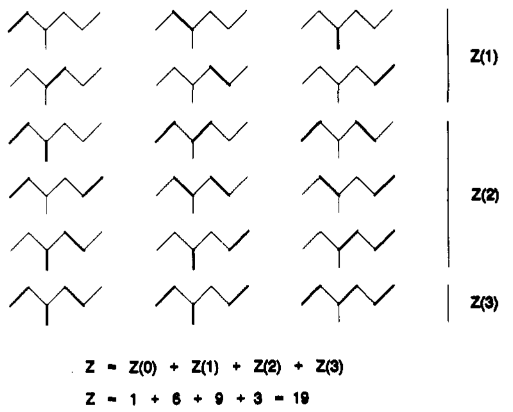


Figure 3. Construction of Hosoya's Z index illustrated with the carbon skeleton of 3-methylhexane.

ANALYSIS OF MOLECULAR DESCRIPTORS

Only a handful of graph theoretical molecular descriptors have found wide use in structure–property and QSAR studies, but the proliferation of such descriptors continues to obscure their role in structure–property and structure–activity studies. We will outline the approach by examining for a set of isomers of heptane, the role played by Hosoya's Z Index,[24] which is defined by the count of nonadjacent edges in a molecular graph. In Figure 3, we illustrate the construction of Z for 3-methylhexane. This particular index has been used in several structure–property correlations,[25] and some of its mathematical properties have received attention.[26] The methodology outlined, however, applies equally, not only to heptanes and other alkanes, but to any family of compounds and any other topological indices as descriptors, including molecular properties.

Relatedness among topological indices has received considerable attention[27] and is perceived by some as a major deterrent to their extensive use. At the same time, relatedness among physicochemical properties as descriptors, although it introduces a fully equivalent problem and is equally disadvantageous, appears to have been generally overlooked (but see Motoc[28]). *Approximate* notions of orthogonality and collinearity may be quite misleading, as was shown recently,[23] whether applied to physicochemical properties or topological indices as descriptors.

BASIC DESCRIPTORS

To evaluate descriptors, a *basis* set of descriptors, B_1, B_2, B_3, B_4, ... must first be selected. Once a basis is chosen, it must be decided whether additional, particularly additional ad hoc, descriptors should be added or whether they are redundant. Graph theoretical invariants, many of which are often and customarily though somewhat incorrectly referred to as topological indices, can be conveniently divided into "individualistic" and "family" types. The former can be viewed as conceptually independent of one another, i.e., each is defined independently of the others. When several indices are derived from essentially the same general definition, one speaks of a

Table I. Graph Theoretical Invariants and Topological Indices Classified as "Individualistic" and "Family" Types[a]

notation	description	authors	ref
	Individualistic Descriptors		
Z	nonadjacent numbers	Hosoya	22
x_i	Eigenvalue index	Lovasz & Pelikan	*b*
W	Wiener (distance sum)	Wiener	*c*
C	centric index	Balaban	*d*
p^1/p	edge path order	Randić	*e*
	Family-Type Descriptors		
	family descriptors	Kier, Murray	27
^1X	higher connectivities	Randic, Hall	27
^1P	path numbers	Platt, Randić	28
^1K	shape indices	Kier	*f*

[a] For information on various topological indices, see Balaban et al.[40] [b] Lovasz, L.; Pelikan, J. *Period. Math. Hung.* **1973**, *3*, 175. [c] Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17. [d] Balaban, A. T. *Theor. Chim. Acta* **1979**, *53*, 355. [e] Randić, M. *J. Math. Chem.* **1991** (in press). [f] Kier, L. B. *Acta Pharm. Jugosl.* **1986**, *36*, 171.

"family" of indices. Both these types are illustrated in Table I. The former is represented, for example, by Hosoya's Z counting index and the latter by the connectivity index,[11] and higher connectivity indices.[29] A natural way of associating an index with various "higher" connectivities is governed by the relative size of the underlying subgraphs.

As the basis (descriptors B_1, B_2, B_3, B_4, ...) in the following, we will adopt the connectivity indices ^1X, ^2X, ^3X, ^4X, ..., which offer a basis for multiparametric representation of a structure. Other "naturally" ordered graph theoretical invariants such as higher path numbers ^kP,[30] weighted paths $^k\Pi$,[31] self-returning walks of different lengths,[32] quantities derived from distance,[33] and other matrices[2] can be used as a basis.

With a basis ^kX, we can answer specific questions on the relevance of other molecular descriptors for the characterization of structures of interest. To qualify as structurally useful, additional descriptors must account for some structural features that the (incomplete) basis set fails to cover. In this way, an ad hoc descriptor that complements deficiencies of the basis can qualify as relevant and be incorporated into the basis. If a descriptor is found to have no significant influence upon the regression, it is discarded from the augmented basis. It may be added that a finite list of graph (and structural) invariants, of which molecular descriptors and topological indices are examples, is generally believed not to form a *complete* basis for the characterization of a structure. However, notwithstanding the fact that the basis set will always be incomplete, our efforts are not directed to approaching the limit of completeness, but rather to define ***minimal basis sets***, i.e., basis sets that are based on as small a number of descriptors as possible but which can still characterize most of the observations of interest. This effort is analogous to the results of factor analysis, in which the apparent dimensionality of a problem under consideration is reduced as much as possible so that the major features of interest of the system are well represented within a space of lower dimension.

To proceed, we must *specify* a statistical threshold which can be regarded as acceptable. As a rule, and somewhat arbitrarily, we will view correlations as suggestive of statistically significant regression if their coefficient of regression, *R*, is above approximately 0.80. Regressions with lower *R* will be viewed as deficient or uninteresting.

EXPANSION OF A DESCRIPTOR IN A CONNECTIVITY BASIS

In order critically to examine the merits or the redundancy of a descriptor for QSAR and other physicochemical studies, we will formally view such a descriptor as a "property". This approach will be illustrated with a multivariate regression of the "property" Z (Hosoya's Index) against the connectivity basis. The Hosoya Index was taken as an illustration because this index, as well as the connectivity index, has been used in the past in a few structure–property correlations, and hence it has shown some use. The nine heptane isomers shown in Table V were selected as a test sample because the size of the molecules is sufficiently small that the various computational steps can easily be checked. The same analysis, however, applies equally to large samples (e.g., the 18 isomers of octane, the 35 of nonane, or the 75 of decane) in which the use of three or more descriptors would be justified, so as to minimize the hazard of a chance correlation. Factors which influence the statistics, but which can be regarded as "external", involving sample size and the number of parameters used in the regression, have received some attention,[34,35] but because external factors are understood (even if often not respected) we will assume here that they do not cause problems in this illustrative case. If they do, this can be established independently of our analysis, by extending the analysis to higher alkanes. Here, we focus attention on what we refer to as the *internal* or inherent factors of a multiple regression. Table II shows the

Table II. Connectivity Index ^1X and Higher Connectivity Indices for the Nine Heptane Isomers with the Value of Hosoya's Z Index

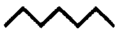
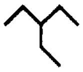
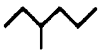
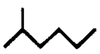
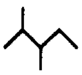
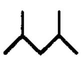
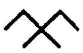
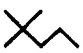
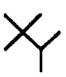
compound	structure	^1X	^2X	^3X	^4X	Z
n-heptane		3.4142	2.0607	1.2071	0.6768	21
3-ethylpentane		3.3461	2.0908	1.7321	0.8660	20
3-methylhexane		3.3081	2.3021	1.4784	0.6969	19
2-methylhexane		3.2701	2.5361	1.1350	0.6124	18
2,3-dimethylpentane		3.1807	2.6295	1.7820	0.4714	17
2,4-dimethylpentane		3.1259	3.0234	0.9428	0.9428	15
3,3-dimethylpentane		3.1213	2.8713	1.9142	0.2500	16
2,2-dimethylpentane		3.0607	3.3107	1.0000	0.7500	14
2,2,3-trimethylbutane		2.9434	3.5207	1.7321	0.0000	13

Table III. Various Combinations of Connectivity Indices Used in a Regression against Hosoya's Z Index and the Values of R (Coefficient of Multiple Regression) and S (Standard Error of Estimate)[a]

descriptors	R	S
${}^1X$	0.9869	0.4721
${}^2X$*	0.9929	0.3486
${}^3X$	0.0478	2.9244
${}^4X$	0.4441	2.6231
${}^1X{}^2X$	0.9951	0.3125
${}^1X{}^3X$	0.9879	0.4898
${}^1X{}^4X$*	0.9963	0.2722
${}^2X{}^3X$	0.9935	0.3607
${}^2X{}^4X$	0.9934	0.3636
${}^3X{}^4X$	0.6125	2.4997
${}^1X{}^2X{}^3X$	0.9987	1.7520
${}^1X{}^2X{}^4X$	0.9962	0.3024
${}^1X{}^3X{}^4X$*	0.9990	0.1531
${}^2X{}^3X{}^4X$	0.9974	0.2516
${}^1X{}^2X{}^3X{}^4X$*	0.9991	0.1682

[a] The best descriptors are indicated with an asterisk (*).

connectivity indices for isomers of heptane used in a somewhat ambitious multivariate regression between Z (the "property", shown in the last column) and the ${}^kX$ connectivity indices. Four indices, ${}^1X$, ${}^2X$, ${}^3X$, and ${}^4X$, have been used which, in view of the size or the number of carbon atoms of the molecules, seem to be a reasonably extensive basis. In larger molecules, even higher connectivity indices, or path numbers, or weighted path numbers could be added to the basis. The question of interest here is: Is the gradual improvement in both the regression coefficient, R, and the standard error of estimate, S, statistically significant *internally*? Put another way: Are the individual descriptors responsible for different structural characteristics of isomeric variations?

Table III shows a summary of the R and S values for various combinations of the connectivity indices used. Many combinations have been used and it should be noted that this causes us to work "against" the external statistical factors discussed by Topliss.[35] The gradual inclusion of higher connectivity indices, understandably, increases R values (from $R = 0.987$ for ${}^1X$ to $R = 0.999$ when all four of the leading connectivity indices are used), while simultaneously decreasing S (from 0.472) to 0.168 when all four indices are used. A threefold reduction in the standard error may appear to be an important improvement in the regression—provided it is not spurious, due to the excessive number of parameters used. Examination of the R and S values, however, immediately indicates the regression based on all four descriptors as being of doubtful quality. This conclusion, as will be seen, does not involve *external* factors as arguments, although some of the same results could be drawn from such.

First, it can be seen that the best single descriptor for the isomeric variation in Z among the heptanes is ${}^2X$, although ${}^1X$ also produces an impressive regression. However, even though ${}^2X$ produces the best regression, the best pair of descriptors does not include ${}^2X$. Among the six pairs of descriptors, the combination of ${}^1X$ and ${}^4X$ produces the best regression. This shows that "greedy" algorithms, in which one successively adds new descriptors to the best combination from a previous step, would miss the optimal solution in this case, the combination of ${}^1X$ and ${}^4X$. Instead they would suggest a combination of ${}^2X$ and ${}^3X$, which gives $R = 0.993$ and $S = 0.361$, as the best pair. Interestingly, this pair does not include either of the descriptors which actually do comprise the best pair.

Among combinations involving three descriptors, the combination ${}^1X$, ${}^3X$, and ${}^4X$ gives an impressive regression with $R = 0.999$ and $S = 0.153$, which is even better than the regression involving all four connectivity indices. This immediately suggests that a fourth descriptor is unnecessary. It is noteworthy that the three descriptors of the best regression do not include ${}^2X$, which was initially found to be the best single descriptor. In fact, addition of this "best single descriptor" to the set of three causes the regression to deteriorate. Conflicting messages of this sort from alternative multivariate regressions hamper evaluations of the significance of individual descriptors and make interpretation of the results of many regressions cumbersome, to say the least. After all, the descriptor ${}^2X$ cannot be as irrelevant as is suggested here; it should be noted that the second best set of three descriptors, almost as good as the best, in terms of R and S, does include ${}^2X$.

How can we disentangle this perplexing situation and make some rational choice of descriptors in some objective manner? The answer is by use of *orthogonal* descriptors, a procedure recently outlined by this author,[32,36] and elaborated further here.

ORTHOGONAL DESCRIPTORS

Rather than using the four connectivity indices ${}^1X$–${}^4X$ as a basis for the characterization of the heptane isomers, we will use four "omega" descriptors, ${}^1\Omega$–${}^4\Omega$, which are derived from the nonorthogonal connectivity indices in a step-by-step process that is analogous to the process in linear algebra of constructing an orthogonal vector basis. Before we describe the orthogonalization process, it may be useful to present an informal outline of the concept of orthogonality. In the case of vectors, the concept of orthogonality can be interpreted geometrically: The simplest case is that of two vectors **a** and **b** in a plane. If **a** and **b** can be drawn in such a way that one is *perpendicular* to the other, they can be described as *orthogonal* [Greek: ορϑος = right and γονια = angle]. If the two vectors are not orthogonal, they can be decomposed so that one vector remains as is while the other produces components which either coincide with the first or which are orthogonal to it. Accordingly, nonorthogonal vectors can be characterized as vectors which have a common (albeit perhaps small) common component. In this spirit, the idea of orthogonality can be generalized to quantities and objects other than vectors. Hence if two molecular descriptors have no component in common, i.e., they do not duplicate one another, we can refer to them as orthogonal. If, on the other hand, they possess a common component, we can view them as nonorthogonal. In applications using multivariate regression, "duplication" becomes "collinearity" and orthogonality of two descriptors in such a context is reflected by a lack of mutual relatedness between them. This suggests the use of residuals between two descriptors in a linear regression to construct novel descriptors that will be orthogonal, because residuals are those portions of the descriptors that cannot be extracted from the regression.

In Table IV, the first phase in the construction of an orthogonal basis for the connectivity indices ${}^1X$–${}^4X$ is illustrated. The outline applies equally to any set of nonorthogonal (i.e., ordinary) molecular descriptors, including quantum chemical quantities and even molecular properties used as descriptors.

We start by selecting the first descriptor, here ${}^1X$, as the first orthogonal descriptor. This is analogous to the first step in the orthogonalization of vectors.[37] Next, we want to make ${}^2X$ orthogonal to ${}^1X$, simulating the process known for vectors, which assumes that the scalar product is defined and requires one to project out the common component of the second vector from the first. Projection of one descriptor on another corresponds to duplication but here is interpreted as that part of the regression between the two descriptors which the first descriptor can fully account for. Hence we proceed by constructing a regression of ${}^2X$ against ${}^1X$ and then use the

Table IV. Orthogonal Descriptors Derived by Making the Higher Connectivity Indices Orthogonal to ^1X

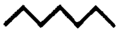
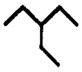
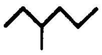
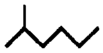
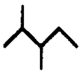
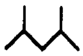
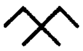
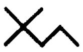
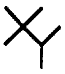
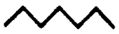
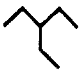
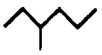
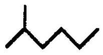
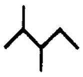
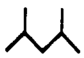
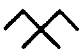
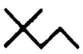
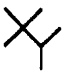
compound	structure	$^1\Omega$	$^2\Omega^1$	$^3\Omega^1$	$^4\Omega^1$
n-heptane		3.4142	0.0910	−0.1745	−0.1533
3-ethylpentane		3.3461	−0.1093	0.3334	0.1127
3-methylhexane		3.3081	−0.0265	0.0703	−0.0136
2-methylhexane		3.2701	0.0790	−0.2826	−0.0554
2,3-dimethylpentane		3.1807	−0.1295	0.3421	−0.0958
2,4-dimethylpentane		3.1259	0.0789	−0.5109	0.4374
3,3-dimethylpentane		3.1213	−0.0886	0.4594	−0.2502
2,2-dimethylpentane		3.0607	0.1457	−0.4700	0.3181
2,2,3-trimethylbutane		2.9434	−0.0408	0.2328	−0.2998

Table V. Set of Orthogonal Descriptors Based on ^1X, ^2X, ^3X, and ^4X and Used in That Order in a Stepwise Orthogonalization Procedure

compound	structure	$^1\Omega$	$^2\Omega$	$^3\Omega$	$^4\Omega$
n-heptane		3.4142	0.0910	0.1403	−0.0036
3-ethylpentane		3.3461	−0.1093	−0.0445	0.1809
3-methylhexane		3.3081	−0.0265	−0.0212	−0.0119
2-methylhexane		3.2701	0.0790	−0.0093	−0.1637
2,3-dimethylpentane		3.1807	−0.1295	−0.1058	−0.0902
2,4-dimethylpentane		3.1259	0.0789	−0.2379	0.0044
3,3-dimethylpentane		3.1213	−0.0886	0.1531	0.0739
2,2-dimethylpentane		3.0607	0.1457	0.0338	0.1908
2,2,3-trimethylbutane		2.9434	−0.0408	0.0916	−0.1206

calculated "^2X" as parts to be subtracted from the original "experimental" or "observed" $^2X_{obs}$. The reason for this is that these very parts of ^2X can be derived from ^1X and hence represent "duplication". The residual, the difference between $^2X_{calc}$ and $^2X_{obs}$ which represents the scatter or imperfection and is actually the "error" of ^2X in the regression against ^1X, becomes the new orthogonal descriptor $^2\Omega$. If a correlation between two descriptors gives a perfect regression ($R = 1$, $S = 0$), the difference between the two descriptors is zero. They represent complete duplicates or, in the language of vector algebra, they are collinear. If a correlation between two descriptors is totally random ($R = 0$, S = the mean deviation of the second variable) the descriptors have no common characteristics; they are already orthogonal.

Table IV illustrates the orthogonal descriptors derived for the higher connectivities ^2X, ^3X, and ^4X, obtained in the first step of orthogonalization. The process continues once one of the orthogonal descriptors derived in this way is selected as a *second* orthogonal descriptor—the first such descriptor being ^1X itself. Then the two remaining descriptors are made orthogonal to the descriptor $^2\Omega$, which was just selected. In the next step, which in this case is the final step, one of the remaining two descriptors is taken as the third orthogonal descriptor and the other is made orthogonal to it. As in vector algebra, successive steps do not upset previously established orthogonalities, and so this process leads to an orthogonal basis, here four omega molecular descriptors.

The orthogonal descriptors derived in this way for the nine isomers of heptane and based upon the successive use of ^1X, ^2X, ^3X, and ^4X, in that order, are given in Table V. The resulting orthogonal descriptors are called *omega descriptors*, the implication being that such descriptors are ultimate descriptors in the characterization of molecules for multivariate regression. Instead of using ^1X directly as our first orthogonal descriptor, we could have subtracted from each ^1X the mean ^1X value. This would make the $^1\Omega$ descriptors sum zero, as is the case with other orthogonal descriptors, and consequently, the constant of the regression would be regarded as the coefficient of the $^0\Omega$ descriptor, which has 1 as its components. Just as in vector algebra, the basis that is constructed will

Table VI. Coefficients in the Regression Equations Using the Orthogonal Connectivity Descriptors of Table V and the Nonorthogonal Connectivity Indices of Table II[a]

descriptor	R	S	descriptor	
$^{1}\Omega$	0.9869	0.4721	^{1}X	
$^{1}\Omega^{2}\Omega$	0.9951	0.3125	$^{1}X^{2}X$	
$^{1}\Omega^{2}\Omega^{3}\Omega$	0.9987	0.1752	$^{1}X^{2}X^{3}X$	
$^{1}\Omega^{2}\Omega^{3}\Omega^{4}\Omega$	0.9991	0.1682	$^{1}X^{2}X^{3}X^{4}X$	
		Coefficients of Regression Equations		
$^{1}\Omega$	$^{2}\Omega$	$^{3}\Omega$	$^{4}\Omega$	constant
17.96691				−40.43492
17.96691	−3.47050			−40.43492
17.96691	−3.47050	1.87545		−40.43492
17.96691	−3.47050	1.87545	−0.56093	−40.43492
^{1}X	^{2}X	^{3}X	^{4}X	constant
17.9669				−40.4349
6.2334	−3.4705			6.4615
28.6314	3.0156	1.8755		−85.3765
22.0204	0.9347	1.0786	−0.5609	−57.1671

[a] Note that both bases are associated with the same R and S values, the coefficients of the orthogonal descriptors are stable, and the "tail" parts of the regressions based on nonorthogonal descriptors give the same coefficients as the corresponding orthogonal descriptors.

depend upon the *choice* of structures and on the *ordering* of vectors (descriptors).

ORTHOGONAL VERSUS NONORTHOGONAL DESCRIPTORS

Statistics based on the use of orthogonal descriptors of Table V, starting with $^{1}\Omega$ and successively adding higher connectivity indices are given in Table VI. The resulting regression equations are compared with the results derived from the nonorthogonal (i.e., ordinary) connectivity indices which were given in Table II. Several important results can be observed in Table VI:

(i) The coefficients of multiple regression R and the standard error of estimate S are the *same* whether we use a set of nonorthogonal descriptors or the corresponding set of orthogonal descriptors. This is not surprising because the latter are derived as a combination of the former and cannot have more information content than the former.

(ii) The distinction of orthogonal descriptors lies in the *stability* of the equation coefficients. It should be noted that each time a new descriptor is added, the correlation coefficients that have already been established do not change. In contrast, the coefficients in equally satisfactory regressions (same R and S) that are based on nonorthogonal descriptors fluctuate without regularity.

(iii) The constancy of the coefficients in the regression equations makes interpretation of the coefficients possible. This is a significant advantage which alone would mandate use of the orthogonal descriptors, but as will be seen later, orthogonal descriptors have other, even more important properties that make their use in future applications of regression analysis mandatory.

(iv) Finally, there is an additional, very important observation to be made from Table VI. Each of the "diagonal" coefficients in the lower part of the table, which describe the connectivity indices ^{k}X added to the regression of a previous step, has the *same* value as the coefficient of the corresponding orthogonal descriptor, $^{k}\Omega$. Hence one can construct the regression equation corresponding to a set of orthogonal descriptors by deriving stepwise regression equations for ordinary (nonorthogonal) descriptors and then using the diagonal coefficients as those of the sought-after regression equation. The constant term of the equation is given by the first (single descriptor) regression equation.

The exploration of combinations of descriptors, illustrated in Table III, can guide one in the selection of descriptors and in the construction of orthogonal bases. But Table VI and the process used to derive it do not indicate that the basis so calculated, whose R and S values are shown at the top of Table VI, is necessarily a statistically significant result. To obtain information on its statistical validity, we must consider the roles of the individual descriptors. This is done in the next section.

INHERENT QUALITY OF INDIVIDUAL DESCRIPTORS AND SEARCH FOR OPTIMAL REGRESSION

The order in which the subsequent descriptors are added, i.e., the ordering of the basis descriptors B_1, B_2, B_3, B_4, ..., fixes the orthogonalization process. In the previous illustration, the ordering of the descriptors was preselected, and as a result, the best combination of descriptors, in this case, $^{1}\Omega$, $^{2}\Omega$, $^{4}\Omega$, was not detected. Clearly, a strategy is required for the detection of optimal regressions, but before addressing this problem, we will consider how to evaluate the role that each descriptor plays *individually* in a regression. Each descriptor is independent of the others, and consequently, we can examine each of them separately and determine whether or not they make a statistically significant contribution to the regression under investigation.

We start by examining regressions based on single descriptors. In our case, ^{1}X (R = 0.987, from Table III) and ^{2}X (R = 0.993) appear to offer some promise. Being "greedy", we select ^{2}X as our first choice, although we are aware that "greedy" algorithms need not yield optimal solutions. The problem we now face is to decide which additional orthogonal descriptors will make a statistically significant improvement upon the already selected descriptor for the regression which, it will be recalled, was between the connectivity indices and Hosoya's Z, used as a prototype of a property. The emphasis here is on the phrase "statistically significant".

Because of the observed additivity of contributing descriptors to the regression equation, i.e., the constancy of the coefficients of individual descriptors, it is possible to determine the role of the second descriptor, using as a *source*, the derived regression based on the first descriptor. First, we compute Z_{calc} as predicted by the regression based on the first descriptor, ^{2}X. The residual—the difference between the computed Z_{calc} and the actual Z_{obs}—represents that part of the correlation of Z that ^{2}X fails to "explain". Hence it is the part that descriptors orthogonal to ^{2}X ought to account for. Direct regressions based on ^{2}X and descriptors that are orthogonal to it ($^{1}\Omega^{2}$, $^{3}\Omega^{2}$, and $^{4}\Omega^{2}$) against Z give the following values for R and S, respectively:

$$\begin{array}{lll} R = 0.995 & R = 0.998 & R = 0.996 \\ S = 0.331 & S = 0.199 & S = 0.276 \end{array}$$

Here, the symbol $^{i}\Omega^{j}$ is used to represent ^{i}X orthogonal to ^{j}X. From the above data there can be no doubt that $^{3}\Omega^{2}$ (the central column) is the best of the three considered. But from such an overall regression, we still have no precise idea of the *significance* of the improvement and, consequently, of the significance of the corresponding descriptor. Does the additional descriptor, for example, make an improvement as significant as that produced by the first descriptor, indicated by R = 0.993 and S = 0.349?

To answer this, we ought to consider 1:1 regressions of the orthogonal descriptors $^{1}\Omega^{2}$, $^{3}\Omega^{2}$, and $^{4}\Omega^{2}$, taken one at a time, against the *residual* of Z. The results of such one-descriptor regressions for selected descriptors are summarized in Table

Table VII. Statistical Parameters Based on Various Combinations of Orthogonalized Descriptors

descriptor	R	S	residual
${}^{1}\Omega^{2}$	0.558	0.289	Z(2)
${}^{3}\Omega^{2}$	0.287	0.334	
${}^{4}\Omega^{2}$	0.259	0.337	
${}^{2}\Omega^{1}$	0.790	0.289	Z(1)
${}^{3}\Omega^{1}$	0.921	0.184	
${}^{4}\Omega^{1}$	0.840	0.256	
${}^{2}\Omega^{4,1}$	0.136	0.254	Z(1,4)
${}^{3}\Omega^{4,1}$	0.854	0.133	
${}^{2}\Omega^{3,1}$	0.592	0.148	Z(1,3)
${}^{4}\Omega^{3,1}$	0.515	0.158	

VII, from which one can deduce the role of individual descriptors independently of any other such descriptor, already used or to be used later. This is the most important aspect of regression analysis based on orthogonal descriptors. It promises to revitalize the entire field of regression analysis by offering an *objective*, statistically based measure of the relevance of any descriptor considered.

When each of the orthogonal descriptors ${}^{1}\Omega^{2}$, ${}^{3}\Omega^{2}$, and ${}^{4}\Omega^{2}$ is individually correlated with the residual Res(2)—the residual remaining after the use of ^{2}X—we obtain, respectively, the following R values:

$$R = 0.558 \qquad R = 0.287 \qquad R = 0.259$$

This shows that none of the higher connectivities orthogonal to the ^{2}X descriptor add anything that is statistically significant to the regression of Z based on ^{2}X. Hence, while ^{2}X by itself is a good descriptor ($R = 0.993$), it leads to a dead-end. Put another way, one cannot improve upon regression based on ^{2}X within the basis of connectivity indices at least not at the similar level of statistical significance.

If, however, we consider a less "greedy" approach and adopt ^{1}X ($R = 0.987$) as our first descriptor, then from Table VII it can be seen that all three of the orthogonal descriptors ${}^{2}\Omega^{1}$, ${}^{3}\Omega^{1}$, and ${}^{4}\Omega^{1}$ show a respectable regression coefficient R for the residual Res(1), the part of Z not accounted for by regression against ^{1}X. In view of the fact that ^{3}X and ^{4}X show no correlation or statistically unacceptably low regression with Z (see Table III), no effort was made to use either of them as the "leading" descriptor. From the upper part of Table VII, we conclude that only ${}^{3}\Omega^{1}$ ($R = 0.921$) and ${}^{4}\Omega^{1}$ ($R = 0.840$) have any promise for further investigation. We excluded ${}^{2}\Omega^{1}$ not because of its marginally lower correlation coefficient ($R = 0.790$) or because of its marginally poorer associated standard error ($S = 0.289$), which is anyway the same as that of the regression based upon the pair ^{2}X and ${}^{1}\Omega^{2}$, but because we have already recognized the regression based on this pair of descriptors to be unacceptable. It should be observed that a correlation on two nonorthogonal descriptors, such as ^{1}X and ^{2}X, can lead to two alternative orthogonalizations, depending upon which of the two vectors is selected as the first descriptor. Because ^{2}X in our illustration is better as a single descriptor, its orthogonal mate ${}^{1}\Omega^{2}$ registers less improvement ($R = 0.559$, $S = 0.289$) than ${}^{2}\Omega^{1}$ makes to ^{1}X ($R = 0.790$, $S = 0.289$). The combined effect of the two descriptors should be the same, and therefore one can use the statistics associated with alternatives to gauge the significance of either of the combinations.

In the next step, summarized in the lower part of Table VII, we explore the role of a third additional descriptor, which is to be added to ${}^{1}\Omega$ and either ${}^{3}\Omega^{1}$ or ${}^{4}\Omega^{1}$. The additional descriptors will be designated as ${}^{i}\Omega^{j,k}$, which signifies that ^{i}X has been made orthogonal to the descriptor ${}^{j}\Omega^{k}$, which in turn is ^{j}X that is already orthogonal to ^{k}X. The orthogonalized descriptors ${}^{2}\Omega^{3,1}$ and ${}^{4}\Omega^{3,1}$ were derived from a regression of ^{2}X and ^{3}X, respectively, against the residuals Res(1,3). The residual Res(1,3) is obtained from a regression of ${}^{3}\Omega^{1}$ against ${}^{1}\Omega$. Similarly, ${}^{2}\Omega^{4,1}$ and ${}^{3}\Omega^{4,1}$ are obtained from a regression of ^{2}X and ^{3}X, respectively, against Res(1,4). It will be seen that the "greedy" approach, attempting to improve upon the best combination ^{1}X, ${}^{3}\Omega^{1}$ ($R = 0.921$) apparently fails to produce statistically significant regression coefficients, the corresponding R values being less than 0.60. If, however, we try to improve on the basis ^{1}X, ${}^{4}\Omega^{1}$ ($R = 0.840$), we see that ${}^{3}\Omega^{4,1}$ accounts for most of the residual Res(1,4) that is left "behind" ^{1}X and ${}^{4}\Omega^{1}$, which we viewed as statistically significant ($R = 0.854$). Having thus established the best combination of descriptors (^{1}X, ^{3}X, ^{4}X) one may now examine the alternative routes to the final combination.

In a general case, the orthogonalization process continues until one has exhausted the possibilities that make a significant contribution to the residuals of the regression in the previous step. In our case, since we already know from Table III that the regression based on all four descriptors is statistically unacceptable, we are in a position to terminate the search for the optimal regression.

We can now summarize the process of construction of optimal regression in three steps, as follows:

(1) Test various combinations of descriptors to produce results analogous to those listed in Table III

(2) Guided by the R and S values, filter out unpromising combinations, eliminating in particular those combinations whose S value is larger than the S values of combinations with fewer descriptors

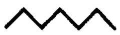
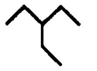
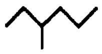
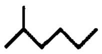
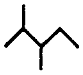
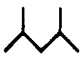
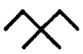
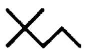
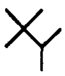
(3) Test accepted descriptors for their R and S values against residuals from previous steps based upon fewer descriptors to determine if they qualify as statistically significant

If the above "rules" are applied to the data in Table III, for the property Z that is under consideration, we would in the first step eliminate as "unpromising" the single descriptors ^{3}X and ^{4}X because both show low R values and excessive S values. Similarly, the pairs of descriptors ^{3}X, ^{4}X and ^{1}X, ^{3}X would be eliminated because their S values are larger than those accepted in the previous step. One could use an even more stringent rule and disqualify combinations whose S value is worse than the "best" S value from the previous step on the grounds that a pair of descriptors should do better than a single descriptor. A more radical filter of this sort eliminates as unpromising the pairs ^{2}X,^{3}X ($S = 0.361$) and ^{2}X,^{4}X ($S = 0.364$) in spite of their impressive R values because ^{2}X as a single descriptor has already given an S value of 0.349. Thus, the only combinations of pairs of descriptors that remain as promising are ^{1}X,^{2}X ($S = 0.313$) and ^{1}X,^{4}X ($S = 0.272$). These two combinations can lead to ^{1}X, ^{2}X, ^{3}X; ^{1}X, ^{2}X, ^{4}X; and ^{1}X, ^{3}X, ^{4}X but not ^{2}X, ^{3}X, ^{4}X, even though it is associated with an acceptable S value of 0.252. From these three possibilities, only the first ($S = 0.175$) and the last ($S = 0.153$) qualify as promising, i.e., could lead to further improvement. However, because the S value associated with all four connectivity indices disqualifies the combination ^{1}X, ^{2}X, ^{3}X, the search ends here with ^{1}X, ^{3}X, ^{4}X selected as the optimal combination. To obtain the optimal orthogonal descriptors, one must trace the history of the final, winning combination by backtracking through the steps that led to its formation and establish the order in which the descriptors were introduced. In our case, the order is ^{1}X followed by ^{4}X followed by ^{3}X, and this defines the orthogonalization process.

CONCLUDING REMARKS

We have outlined a process of construction of orthogonal descriptors by illustration of the regression of Hosoya's Z index, here treated as a property, against connectivity indices. Rather than constructing at each step all the orthogonal combinations that are to be tested in the next step of the

Table VIII. Residuals in Regressions of Z against Various Connectivity Indices and Their Orthogonal Combinations Viewed as Novel Molecular Descriptors

compound	structure	Z(1)	Z(2)	Z(3)	Z(1,2)	Z(1,3)	Z(1,4)	Z(1,2,3)	Z(1,4,3)
n-heptane		0.0921	0.6204	4.0807	0.4840	0.2844	−0.1369	0.1449	0.1232
3-ethylpentane		0.3166	−0.2216	2.8956	−0.0627	−0.0508	0.4849	0.0208	0.1336
3-methylhexane		−0.0007	−0.1133	1.9851	−0.0925	−0.0781	−0.0211	−0.0527	−0.0649
2-methylhexane		−0.3178	0.1139	1.1061	−0.0435	−0.0064	−0.4005	−0.0260	−0.1352
2,3-dimethylpentane		0.2870	−0.3959	−0.1220	−0.1624	−0.0890	0.1440	0.0360	−0.0455
2,4-dimethylpentane		−0.7277	−0.3302	−1.8262	−0.4537	−0.1648	−0.0742	−0.0076	−0.0292
3,3-dimethylpentane		0.3545	−0.1278	−1.1686	0.0470	−0.1517	−0.0194	−0.2400	−0.1756
2,2-dimethylpentane		−0.5557	0.1764	−2.8464	−0.0501	−0.0379	−0.0805	−0.1135	0.0291
2,2,3-trimethylbutane		0.5517	0.2782	−4.1044	0.4100	0.2952	0.1037	0.2382	0.1645

analysis, a more efficient procedure consists of combining the results of Table III with those of Table VI. First, combinations of descriptors are filtered, and only the promising combinations are retained for the next step in the analysis. Then, by examining the history of the optimal solution, one can reconstruct the corresponding orthogonal combinations to which one then applies a statistical test for the residual of the correlation using fewer descriptors. This test will determine if addition of the new descriptor is a statistically significant and therefore permissible step.

The present methodology makes most of the past regression analyses outdated, or at least incomplete. Moreover, the analysis outlined makes it possible to test the statistical significance of past results independently of alternative statistical criteria, such as cross-validation[38] and bootstrapping.[39] The stepwise inclusion of (nonorthogonal) descriptors in published multivariate regressions may now be reexamined. As published, they should generally give the corresponding regression equation for the implied orthogonal basis because it can be constructed using the "diagonal" coefficients. Such an orthogonalized basis also allows a reinterpretation of the results.

A question posed, but not yet answered is: Do we need molecular descriptors beyond a well-selected basis? More specifically: Do we need a Z index to describe the various properties of heptane isomers?

We can formally view the regressions of connectivity indices against Z as an "expansion" of Z in terms of the connectivity indices. It appears from the foregoing analysis that "most" of the Z property can be "captured" by the three connectivity indices ^1X, ^3X, and ^4X, but in spite of this, Z contains structural characteristics that the connectivity indices do not fully account for. The emphasis here is on "fully", which means that Z possesses some structural characteristics that are beyond the descriptive capabilities of connectivity indices. It is misleading to think that because "most" of one descriptor is accounted for by other descriptors that such a descriptor is redundant. This is seen clearly if one considers the "unexplained" residual of such regressions between descriptors as a *new* descriptor. For instance, we can view the residual of Z with respect to the regressions against orthogonalized ^1X, ^4X, and ^3X as novel descriptors: $^Z\Omega$, index orthogonal to the respective connectivity indices ^1X–^4X. Obviously then the connectivity indices have nothing in common with the Z descriptor that has been orthogonalized in this way. In Table VIII, we show the residuals of Z that were obtained by the use of various combinations of connectivity indices. These residuals can be regarded as descriptors made orthogonal to the corresponding connectivity indices ^1X, ^2X, and ^3X, respectively (the first part of Table VIII), and as orthogonal to $^2\Omega^1$, $^3\Omega^1$, $^3\Omega^{2,1}$, and $^3\Omega^{4,1}$, respectively (the second part of Table VIII). Thus, for example, Z(*i*) and Z(*ij*) describe residuals or orthogonal Z descriptors to ^iX and $^j\Omega^i$, respectively, and in general, Z(*i,j*) ≠ Z(*j,i*). Thus a Z index—and many other indices—are needed, despite their mutual interrelatedness, if they are reflections of *distinctive* structural characteristics in order to supplement the deficiencies of the basis that uses only connectivity indices.

Finally, we should add that the question "Do we need additional descriptors beyond the connectivity indices?" could be restated as "Do we need the connectivity indices beyond the available ad hoc descriptors?" The expected answer will be reciprocal, i.e., if a base B_1, B_2, B_3, B_4, ... does not need descriptors D_1, D_2, D_3, D_4, ..., then the same will be true when D_i is used as a base instead of B_i. It may be a matter of opinion as to which basis is preferred, but simplicity of structural interpretation should be the final guide. Future applications will show which descriptors may best account for the properties of interest in different situations, and at the same time the methodology described in this paper will allow critical evaluation not only of the various reported regressions but also of the various molecular descriptors.

This article has delivered some bad news and some good news for those engaged in multivariate analysis. The bad news is that most previously reported regressions in structure–property and structure–activity now appear to be incomplete, to say the least. The good news is that previously reported stepwise regressions—those in which one descriptor is added to a smaller set that has already been examined—will allow the reconstruction, using diagonal coefficients, of the corresponding multivariate regressions based upon orthogonal descriptors, and this will allow a meaningful interpretation of the coefficients.

ACKNOWLEDGMENT

I thank Dr. George W. A. Milne (NIH, Bethesda), the Editor, for his interest in this work and for his numerous comments and suggestions on the manuscript which resulted

in a significantly improved presentation of the material.

REFERENCES AND NOTES

(1) Pauling, L. *The Nature of the Chemical Bond*; Cornell University Press: Ithaca, NY, 1960.

(2) Randić, M. *Proc. Galveston Conf. Math. Chem.* April 1990; *J. Math. Chem.* **1991**, *4*, 157.

(3) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983.

(4) Klopman, G.; Kalos, A. N. *J. Comput. Chem.* **1985**, *6*, 492.

(5) Trinajstić, N.; Randić, M.; Klein, D. J. *Acta Pharm. Yugosl.* **1986**, *36*, 267.

(6) Cramer, R. D., III. *J. Am. Chem. Soc.* **1980**, *102*, 1837.

(7) Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86*, 1616. Hansch, C. *Acc. Chem. Res.* **1969**, *2*, 232.

(8) Kováts, E. *Z. Anal. Chem.* **1961**, *181*, 351.

(9) Dirac, P. A. M. *Quantum Mechanics*; Oxford University Press: London, 1958.

(10) Coulson, C. A. *Proc. R. Soc. London, A* **1939**, *169*, 413.

(11) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6009.

(12) Initially the index was named the "branching index", being sensitive to molecular branching. The index, however, applies equally to linear chains and cyclic structures and a better name, suggested by L. B. Kier, is the "connectivity index": Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. *J. Pharm. Sci.* **1975**, *64*, 1971.

(13) Klein, D. J.; Schmalz, T. G.; Hite, G. E.; Seitz, W. A. *J. Am. Chem. Soc.* **1986**, *108*, 1301. Klein, D. J.; Seitz, W. A.; Schmalz, T. G. *Nature* **1986**, *232*, 6090. Schmalz, T. G.; Klein, D. J.; Hite, G. E. *J. Am. Chem. Soc.* **1988**, *110*, 1113.

(14) Montroll, E. E. *J. Chem. Phys.* **1941**, *9*, 706. Klein, D. J.; Hite, G. E.; Schmalz, P. G. *J. Comput. Chem.* **1986**, *7*, 443.

(15) Essam, J. W.; Fisher, M. E. *Rev. Mod. Phys.* **1970**, *42*, 272.

(16) Cyvin, S. J.; Gutman, I. *Kekule Structures in Benzenoid Chemistry*; Lecture Notes in Chemistry, Vol. 46; Springer: Berlin, 1988.

(17) Ham, N. S.; Ruedenberg, K. *J. Chem. Phys.* **1958**, *29*, 1215, 1229. According to K. Ruedenberg (private communication), it was J. R. Platt who recognized the novel bond orders of Ham and Ruedenberg as Pauling bond orders which had already been described.

(18) Dewar, M. J. S.; Longuet-Higgins, H. C. *Proc. R. Soc. London, A* **1952**, *214*, 482.

(19) Platt, J. R. In *Encyclopedia of Physics*; Flügge, S., Ed.; Springer Verlag: Berlin, 1961; Vol. 37, Part 2. W. C. Herndon, who, in a series of papers, advocated this approach of enumeration of Kekulé valence structures, also drew attention to Platt's early work.

(20) Ruedenberg, K. *J. Chem. Phys.* **1954**, *22*, 1878.

(21) Heilbronner, E. *Helv. Chim. Acta* **1962**, *45*, 1722.

(22) Cohn, M. C. *Bull. Math. Biol.* **1986**, *48*, 417.

(23) Randić, M. *New J. Chem.* **1991**, in press.

(24) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.

(25) Hosoya, H.; Kawasaki, K.; Mituzani, K. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 3415. Narumi, H.; Hosoya, H. *Bull. Chem. Soc. Jpn.* **1980**, *53*, 1228.

(26) Gutman, I.; Polansky, O. *Mathematical Concepts in Organic Chemistry*; Springer Verlag: Berlin, 1986.

(27) Balaban, A. T.; Motoc, I. *Match* **1979**, *5*, 107. Motoc, I.; Balaban, A. T. *Rev. Roum. Chim.* **1981**, *26*, 593. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. *Math. Modeling* **1987**, *8*, 302. Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. *Match* **1982**, *13*, 369. Razinger, M.; Chrětian, J. R.; Dubois, J.-E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23.

(28) Motoc, I. *Topics Curr. Chem.* **1983**, *114*, 93.

(29) Kier, L. B.; Murray, J. W.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226.

(30) Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419. Randić, M. *Match* **1979**, *7*, 5. Randić, M.; Wilkins, C. L. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1979**, *6*, 55. Wilkins, C. L.; Randić, M. *Theor. Chim. Acta* **1980**, *58*, 45. Wilkins, C. L.; Randić, M.; Schuster, S. M.; Marklin, R. S.; Steiner, S.; Dorgan, L. *Anal. Chim. Acta* **1981**, *133*, 637. Randić, M.; Kraus, G.; Jerman-Blazic, B. In *Chemical Applications of Topology and Graph Theory*; Elsevier: Amsterdam, 1983; p 192. Jerman-Blazic, B.; Randić, M. *Proc. Conf. Modeling Sim.* **1983**, *5*, 161.

(31) Randić, M. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1984**, *11*, 137. Randić, M. In *Molecular Basis for Cancer. Part A: Macromolecular Structure, Carcinogens and Oncogens*; Rein, R., Ed.; Alan R. Liss Inc.: New York, 1985; p 309. Randić, M.; Jerman-Blazic, B.; Grossman, S. C.; Rouvray, D. H. *Math. Modeling* **1986**, *8*, 571. Grossman, S. C.; Jerman-Blazic, B.; Randić, M. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1986**, *12*, 123. Randić, M.; Jerman-Blazic, B.; Rouvray, D. H.; Seybold, P. G.; Grossman, S. C. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1987**, *14*, 245. Randić, M.; Grossman, S. C.; Jerman-Blazic, B.; Rouvray, D. H.; El-Basil, S. *Math. Comput. Model.* **1988**, *11*, 837.

(32) Randić, M. *J. Comput. Chem.* **1980**, *1*, 368. Knop, J. V.; Müller, W. R.; Szymanski, K.; Randić, M.; Trinajstić, N. *Croat. Chem. Acta* **1983**, *56*, 405. Bogdanov, B.; Nikolic, S.; Sabjlic, A.; Trinajstić, N.; Carter, S. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1985**, *14*, 325.

(33) Rouvray, D. H. In *Mathematical and Computational Concepts in Chemistry*; Trinajstić, N., Ed.; Horwod: Chichester, 1986. Buckley, F.; Harary, F. *Distances in Graphs*; Addison-Wesley: Reading, MA, 1990.

(34) Wilson, E. B. *Introduction to Scientific Research*; McGraw-Hill: New York, 1952.

(35) Topliss, J. G.; Costello, R. G. *J. Med. Chem.* **1972**, *15*, 1066. Topliss, J. G.; Edwards, R. G. *J. Med. Chem.* **1979**, *22*, 1238. Stouch, T. R.; Jurs, P. C. *Quant. Struct.-Act. Relat.* **1986**, *5*, 57.

(36) Randić, M. *New J. Chem.* Submitted for publication. Randić, M. *Croat. Chem. Acta* **1991**, in press. Randić, M. *J. Mol. Struct.* **1991**, in press.

(37) Courant, R.; Hilbert, D. *Methoden der Mathematike Physik*; Springer Verlag: Berlin, 1931.

(38) Geisser, S. *J. Am. Statist. Soc.* **1975**, *70*, 328.

(39) Diaconis, P.; Efron, B. *Sci. Am.* **1984**, 116.

(40) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Top. Curr. Chem.* **1983**, *114*, 21.

The Beilstein Structure Registry System. 1. General Design

LÁSZLÓ DOMOKOS

Beilstein Institute, Varrentrappstrasse 40-42, D-6000 Frankfurt/Main 90, Germany

Received November 16, 1990

The rapidly growing Beilstein Online Database contains already more than 3.4 million organic compounds. Since it is a compound-oriented factual database, the structure registration plays a central role. The understanding of the basic features of the registration is important for the effective usage of the database. This paper describes the software and the philosophy of the structure-registry system embedded into the data processing from the data acquisition through to the dissemination of the data.

INTRODUCTION

The registration of chemical structures is fundamental to each large structural database and structure retrieval system. All larger systems have their own registration software. The best known is the CAS Chemical Registry System developed in the early 1960s. Registry III, its latest version, has been used since 1974.[1,2] These systems were designed to be optimal for the class of compounds to be processed, for the data structure, and for the software and hardware environment. As a consequence, these systems are usually not transportable and are not commercially available. To meet the requirements of the Beilstein database the Beilstein Structure Registry System was developed between 1986 and 1990.

The Beilstein Online Database receives the structural and factual data from three different sources: the Beilstein Handbook of Organic Chemistry, the file cards abstracted from the literature for the Handbook production, and the original literature covering all important journals of organic chemistry.[3] These three sources correspond to the literature periods 1830–1959, 1960–1979, and 1980 onwards, respec-

0095-2338/91/1631-0320$02.50/0 © 1991 American Chemical Society