PREDICTING PROPERTIES OF MOLECULES USING GRAPH INVARIANTS

Subhash C. BASAK and Gerald J. NIEMI
Center for Water and the Environment, Natural Resources Research Institute,
University of Minnesota, 5013 Miller Trunk Highway, Duluth, MN 55811, USA

Gilman D. VEITH
US Environmental Protection Agency, Environmental Research Laboratory–Duluth,
6201 Congdon Boulevard, Duluth, MN 55804, USA

Abstract

Topological indices (TIs) have been used to study structure–activity relationships (SAR) with respect to the physical, chemical, and biological properties of congeneric sets of molecules. Since there are many TIs and many are correlated, it is important that we identify redundancies and extract useful information from TIs into a smaller number of parameters. Moreover, it is important to determine if TIs, or parameters derived from TIs, can be used for global SAR models of diverse sets of chemicals. We calculated seventy-one TIs for three groups of molecules of increasing complexity and diversity: (a) 74 alkanes, (b) 29 alkylbenzenes, and (c) 37 polycyclic aromatic hydrocarbons (PAHs). Principal components analysis (PCA) revealed that a few principal components (PCs) could extract most of the information encoded by the seventy-one TIs. The structural basis of the first few PCs could be derived from their pattern of correlation with individual TIs. For the three sets of molecules, viz. alkanes, alkylbenzenes and PAHs, PCs were able to predict the boiling points reasonably well. Also, for the combined set of 140 chemicals consisting of the alkanes, alkylbenzenes and PAHs, the derived PCs were not as effective in predicting properties as in the case of individual classes of compounds.

1. Introduction

The last two decades witnessed an upsurge of interest in applications of graph theory in chemistry [1–21]. Constitutional formulae of molecules are chemical graphs where vertices represent the set of atoms and edges represent chemical bonds [13]. The pattern of connectedness of atoms in a molecule is preserved by constitutional graphs. A graph (or more correctly, non-directed graph) $G = [V, E]$ consists of a finite nonempty set V of points together with a prescribed set E of unordered pairs of distinct points of V [22]. A *structural model* assigns to the points of G a realization in some applied field and each element of E indicates a pair of entities (elements of the structural model) which are in the finite nonempty irreflexive symmetric binary relation described by G [23]. For example, when elements of the set V symbolize atomic cores without valence electrons and the elements of E represent covalent two-electron bonds, G is the molecular graph or constitutional graph of a covalent chemical species. Such a graph can represent structural formulae for a large number of organic compounds. Since more than 90% of chemical compounds described so far are either organic or contain organic ligands, such a graph has been found to be useful in chemistry [24]. The edge set need not always represent a covalent bond. In fact, elements of E may symbolize almost any type of bond, e.g. ionic bond, coordinate bond, hydrogen bond or weak bonds representing transition states of SN_2 reaction, etc. [25–27]. If the interaction between a pair of atoms is asymmetric, e.g. in the case of sufficiently polar covalent bonds, hydrogen bond donor acidity, hydrogen bond acceptor basicity or charge transfer complex formation, the bonding pattern can be represented by a binary relation which is antireflexive and asymmetric [6]. Further refinement could be achieved through the assignment of weights to the vertices or edges [13], and use of multiple edges between a pair of atoms held together both by *sigma* and *pi* bonds. Evidently, a weighted pseudograph appears to be the most general model capable of symbolizing the bonding pattern of a large number of organic and inorganic chemicals.

Chemists have long relied on visual perception to relate various aspects of constitutional graphs to observable phenomena. However, a clear and quantitative understanding of the structural basis of chemistry necessitates the use of precise mathematical techniques. In recent years, applications of matrix theory, group theory, and information theory to chemical graphs have produced results which are important in chemistry [4,6,9–11,13,19,28–31]. The power of graph-theoretic formalism in chemistry is evident from its successful applications in chemical documentation, isomer discrimination and characterization of molecular branching, enumeration of constitutional isomers associated with a particular empirical formula, calculation of quantum chemical parameters, structure–physicochemical property correlations, and chemical structure–biological activity relationships [1–21].

Mathematical characterization of a chemical structure may be accomplished by a matrix, a set of numbers or a single numerical index [13]. For example, the adjacency matrix $A(G)$, the distance matrix $D(G)$, and the incidence matrix $T(G)$ of a chemical graph G uniquely determine molecular topology. Among the different matrices used for the representation of chemical structure, the adjacency matrix has been more frequently used in chemistry [13]. However, the adjacency matrix poses a serious problem in chemical documentation because, as the size of the graph increases, they require a disproportionately large number of operations for the testing of graph isomorphism. Specifically, $n^2 \times n!$ operations are required to determine whether two graphs G_1 and G_2 with n vertices are isomorphic [13]. Another limitation of matrices is that they cannot be used as structural descriptors in the correlation or prediction of properties. Therefore, one of the cherished objectives of graph-theoretic research has been the discovery of a graph property, preferably a single numerical characteristic or a set of numbers derived from graphs, which would not only be easier to handle than the adjacency matrix itself, but also uniquely related to molecular topology. Unfortunately, in spite of numerous attempts, attainment of this goal has remained elusive.

In chronological order, Spialter [25–27] was the first to search for a graph invariant which could uniquely characterize molecular topology. A graph invariant is a graph-theoretic property which is preserved by isomorphism [13,22]. Spialter [25–27] asserted that the characteristic polynomial of the adjacency matrix or atom connectivity matrix of a molecule is uniquely related to its topology. This notion was, however, contradicted by later researchers, who found that nonisomorphic graphs may possess identical characteristic polynomials [32–34]. These graphs are called isospectral or cospectral graphs [13]. Later, Randić [35] conjectured that for tree graphs, collections of distance degree sequences (DDS) were sufficient to determine isomorphism. Subsequently, it was reported that neither DDS nor PDS (path degree sequence) could uniquely characterize the topology of tree graphs [36,37]. More recently, Randić [38] developed a single numerical index, the molecular identification number, which was successful in the unique characterization of the topology of a relatively diverse set of structures, including constitutional isomers and cyclic analogs. However, a counter-example, i.e. different chemical structures with the same identification number, has already been reported [38,39].

Under these circumstances, there are three distinct trends in chemical graph-theoretic index research: (a) the simultaneous use of more than one index, i.e. a superindex in order to better characterize chemical structure as compared to a single index [40], (b) development of indexes with high discriminating power [38,41,42], and (c) extraction of useful information from a collection of indices already defined in the literature [43–46].

2. System complexity, reductionism and structure–activity relationship (SAR)

Use of topological indices (TIs) in SAR may be clearly understood through a formal exposition of the *structure–property similarity principle* – the central paradigm of SAR [6,47,48]. Figure 1 represents an empirical property as a function $\alpha: C \to \mathbb{R}$ which maps the set C of compounds into the real line $\mathbb{R}$. A nonempirical SAR may be looked upon as a composition of a description function $\beta_1: C \to D$ mapping each chemical structure of C into a space of nonempirical structural descriptors (D) and a prediction function $\beta_2: D \to \mathbb{R}$ which maps the descriptors into the real line. When $[\alpha(C) - \beta_2\, \beta_1(C)]$ is within the range of experimental errors, we say that we have a good nonempirical predictive model. On the other hand, the property–activity relationship (PAR) is the composition of $\theta_1: C \to M$ which maps the set C into the molecular property space M and $\theta_2: M \to \mathbb{R}$ mapping those molecular properties into the real line $\mathbb{R}$. PAR seeks to predict one property (usually a complex property) of a molecule in terms of another (usually simpler) property. The latter group of properties may consist either of a number of experimentally determined quantities (e.g. melting point, boiling point, vapor pressure, partition coefficient) or substituent constants or solvatochromic parameters (e.g. steric, electronic, hydrophobic, charge transfer substituent constants, hydrogen bond donor acidity, hydrogen bond acceptor basicity).

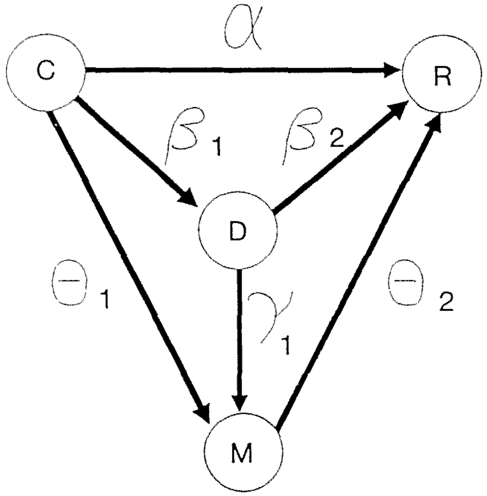

Fig. 1. Composition functions for structure–activity relationship (SAR) and property–activity relationship (PAR).

PAR using a calculated property (e.g. calculated partition coefficient, log P, octanol–water) may be looked upon as a mapping $\theta_2 \gamma_1 \beta_1 : C \rightarrow I\!R$, which is a composition of β_1, $\gamma_1 : D \rightarrow M$ mapping the descriptor space into the molecular property space (e.g. calculation of log P from fragments using additivity rule), and θ_2, as described above.

Both in drug design and in predictive toxicology, SAR is an explosive problem. In drug design, one can synthesize a large number of derivatives from a "lead" structure. It is not unusual that one has to test 200,000 or more chemicals to discover a molecule that is marketable [49]. In many cases, one might be interested to know the property of a molecule not yet synthesized. Then the only solution is to estimate properties using theoretical parameters which can be calculated for any arbitrary chemical structure, real or hypothetical. The situation is no different in the risk assessment of chemicals. More than nine million chemicals are listed in the Registry of Chemicals of the Chemical Abstract Service and out of these, about 76,000 are in daily use [50]. We need to know a large number of properties and activities (or endpoints) of these chemicals in order to perform a reasonable risk assessment. Table 1 gives a partial list of endpoints necessary for pharmacological/toxicological evaluation of chemicals. Although many of the properties listed in table 1 can be determined experimentally, the combination of these properties and the number of candidate chemicals is a combinatoric explosion! Cost and time limitations will not allow us to test a large fraction of existing chemicals in a rigorous way. Therefore, there is a need to develop procedures which can rapidly screen chemicals for their toxicological properties and allow us to focus scarce resources on chemicals with the greatest potential risk.

Table 1

Important SAR endpoints

Physicochemical	Pharmacological/Toxicological
Molar volume	Macromolecule level
Boiling point	: Receptor binding (K_D)
Melting point	: Michaelis constant (K_m)
Vapor pressure	: Inhibitor constant (K_i)
Water solubility	: DNA alkylation
Dissociation constant (pKa)	: Unscheduled DNA synthesis
Partition coefficient	Cell level
: Octanol–water (log P)	: Salmonella mutagenicity
: Air–water	: Mammalian cell transformation
: Sediment–water	Organism level (acute)
Reactivity (electrophile)	: Algae
	: Invertebrates
	: Fish
	: Birds
	: Mammals
	Organism level (chronic)
	: Bioconcentration
	: Carcinogenicity
	: Reproductive toxicity
	: Delayed neurotoxicity
	: Biodegradation
	Ecosystem level
	: ??

A closer look at the properties of table 1 will show that they belong to different levels of molecular and biological organization. Prediction of any endpoint using SAR techniques is based on the implicit reductionistic assumption, which may be expressed as [51]:

> "Given any physico-chemical system S, however complex, there exists a physically effective resolution of S (or a set of such resolutions) into fractional subsystems such that (i) each fractional subsystem in the resolution can be precisely characterized (in the customary physico-chemical sense), and (ii) any property of S can be reconstructed from the properties of the fractional subsystems."

It has been pointed out by Rosen [51] that systems with a high degree of controlability cannot in general be effectively fractionated into subsystems satisfying the above conditions.

On the other hand, there has been extensive discussion of "*emergent properties*", by which is meant properties of whole systems which cannot effectively be predicted from the properties of simpler subsystems or more precisely, that certain resolutions into fractional subsystems do not satisfy the above assumption. Indeed, some authors [52] have gone so far as to claim that most interesting biological properties are emergent in this sense, and are not reducible to, or derivable from, the kind of knowledge with which physicists have ordinarily been satisfied. Such claims mean, in essence, that if the above hypothesis is true at all, the required resolutions must be produced by entirely novel means.

In spite of the above limitations, we believe that in many cases systems can be resolved (at least to a first-order approximation) into non-interacting parts. Von Neumann [53], however, argued that there is a level of complexity below which the world behaves with regularity, but above which entirely new behaviors appear. In this sense, complex systems are counter-intuitive, i.e. their behavior deviates from what common sense suggests.

In SAR, where we are interested in explaining the behavior of complex systems (properties of molecules or organisms, or more complex reality like communities or ecosystems) in terms of behaviors of simpler subsystems (properties of substructures, properties of total molecular structure, or properties of simpler biological systems), the reductionist approach may lead to a few difficulties:

I. The chosen subsystems are incapable of predicting the behavior of the complex system; the predictive capability is highly uncertain.

II. There is reasonable predictability in a narrow range, but the model fails to account for situations which are diverse.

III. One class of subsystems is capable of predicting only one property and fails in predicting another property of the same system.

Situation I arises in the case of predicting carcinogenicity of chemicals. Even a combination of calculated structural parameters, physicochemical properties, and a range of genotoxic effects of chemicals at the level of macromolecules and simpler model biological systems does not adequately predict the carcinogenic effects of those chemicals in higher organisms [55]. This is probably because the gap in the level of complexity between the system property (carcinogenicity) and properties of predictor subsystems is so high that many aspects of system complexity are lost in the process of abstraction. A similar situation may arise in the assessment of ecotoxical risk of chemicals, where toxic effects on simpler systems, e.g. an organism, microcosm, or mesocosm, might not be able to predict population, community or ecosystem-level effects of chemicals.

Situation II is observable where there is reasonable success in predicting properties for a narrow range of systems, but the predictability gradually fades as the range broadens. An example may be the success of prediction of aquatic toxicity (LC_{50}) of a particular class of chemicals, e.g. narcotics, uncouplers of oxidative phosphorylation in fathead minnow, using lipophilicity [56]. When chemicals of different modes of action are taken together, the predictive power of lipophilicity diminishes.

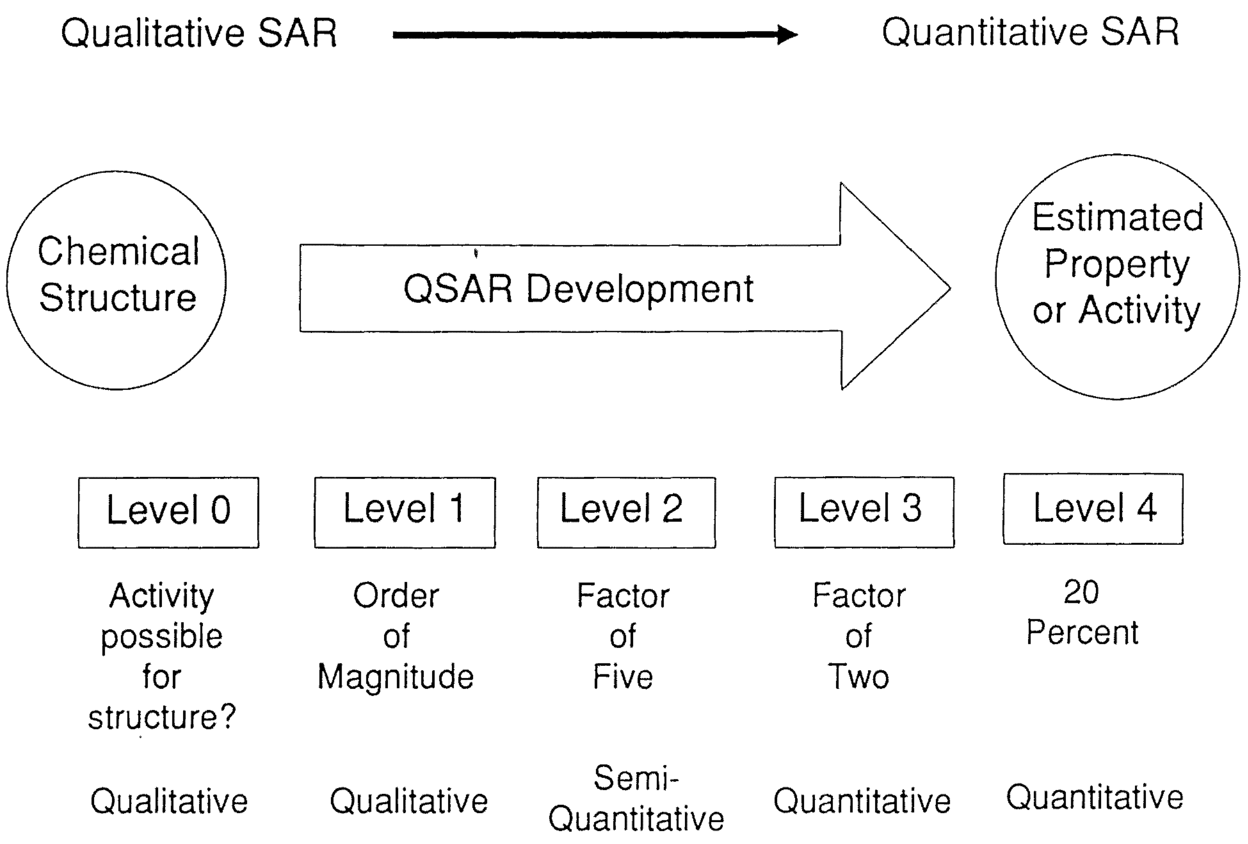

Fig. 2. Different levels of structure–activity relationship (SAR).

Situation III arises when a particular type of reductionist abstraction is useful in predicting a particular property *X* of the system, whereas another abstraction is necessary for another property *Y*. For example, connectivity index correlates well with the boiling point of alkanes, whereas Balaban's index (*J*) correlates highly with the octane number of alkanes. For a given set of alkanes, boiling point and octane number are not significantly correlated with each other [38]. This situation is explained by the fact that the selected set of invariants (subsystem) used for the calculation of the connectivity index is different from those used for calculating the *J* index, although they are derived from the same set of alkanes (system).

Prediction of a very complex property, e.g. the carcinogenic potential of chemicals, is zero level (or partial) SAR where structural variables comprise a subset of predictors and we have very little confidence in the success of the model. As we move to less complex properties like bioconcentration factor or acute toxicity, we are more successful in quantitative prediction of properties with greater confidence. Finally, with well-defined physicochemical properties, viz. boiling point or lipophilicity (K_{ow}, octanol–water partition coefficient), we have good structural models which might be called quantitative SAR or QSAR. Figure 2 gives a schematic view of the spectrum of SAR – from qualitative to quantitative SAR. At the level of QSAR, one would be interested to know the relationship between predictors and properties, i.e. whether the property is additive, multiplicative, constantive or derivative [57,58].

3. Quantitation of structure and graph invariants

In the context of molecular science, the various concepts of molecular structure (e.g. classical valence bond representation, various chemical graph-theoretic representations, ball and spoke model of a molecule, representation of the molecule by minimum energy conformation, semi-symbolic contour map of a molecule, or symbolic representation of chemical species by Hamiltonian operators) are model objects [59] derived through different abstractions of the same chemical reality or molecule [60,61]. In each instance, the equivalence class (concept or model of molecular structure) is generated by selecting certain aspects while ignoring some unique properties of those actual events. This explains the plurality of the concept of molecular structure and their autonomous nature, the word autonomous being used in the sense that one concept is not logically derived from the other [60].

Any concept of molecular structure is a hypothetical sketch of the organization of molecules. Such a *model object* is a general theory and remains empirically untestable. A model object has to be grafted onto a specific theory to generate a *theoretical model* [59] which can be empirically tested. For example, when it was suggested by Sylvester [62] in 1878 that the structural formula of a molecule is a special kind of graph, it was an innovative general theory without any predictive potential. When the idea of combinatorics was applied on chemical graphs (model object), it could be predicted that "there should be exactly two isomers of butane (C_4H_{10})" because "there are exactly two tree graphs with four vertices" when one considers only the nonhydrogen atoms present in C_4H_{10} [24]. This is a theoretical model of limited predictive potential. Although it predicts the existence of chemical species, given a set of molecules, e.g. isomers of hexane (C_6H_{14}), the model is incapable of predicting any property. This is because of the fact that any empirical property P maps a set of chemical structures into the set $\mathbb{R}$ of real numbers and thereby orders the set empirically. Therefore, to predict the property from structure, we need a nonempirical (structural) ordering scheme which closely resembles the empirical ordering of structures as determined by P [6,47]. This is a more specific theoretical model based on the same model object (chemical graph) and can be accomplished by using specific graph invariant(s).

The predictive potential of a theoretical model depends both on: (1) efficacy of representation of the relevant aspects of reality by the model object,.and (2) optimal treatment of the model object by appropriate mathematical theories. Effective prediction also depends on the quality of available data and the level of complexity of the property of interest.

In view of the pressing need of development of SARs for predicting a plethora of endponts, it is desirable to use easily calculable structural parameters as the first approach. In recent years, methods based on chemical graph theory have emerged as important tools in biomedicinal and toxicological SAR. Graph invariants have been used in isomer discrimination, ordering of structures, structure–property relationship, and chemical structure–biological activity relationships [1–21,63]. In the latter two areas, numerical graph invariants (or topological indices) have been used mainly in correlating properties of different groups of congeners.

However, some important questions regarding the nature and utility of topological indices (TIs) remain unanswered:

(a) Since many of the available TIs are intercorrelated, can we extract useful information from TIs for practical use?

(b) Can we use TIs to order molecules and predict properties of structurally diverse sets of chemicals?

In an attempt to answer these questions, we calculated a large number of TIs (*vide infra*) for three structurally diverse groups of molecules. We used principal components analysis (PCA) to extract useful information from TIs into a low-dimensional space consisting of principal components (PCs) that explain a large fraction of the variance in the original data. We studied the utility of TIs *vis-a-vis* PCs derived from them in correlating properties of different groups of molecules. Finally, we attempted to analyze which aspects of molecular structure are encoded by different PCs.

4. Calculation of parameters

4.1. CALCULATION OF WIENER INDEX (W) AND CONNECTIVITY INDICES

The topological parameters used in this paper may be conveniently derived from the adjacency matrix $A(G)$ or the distance matrix $D(G)$ of a chemical graph G.

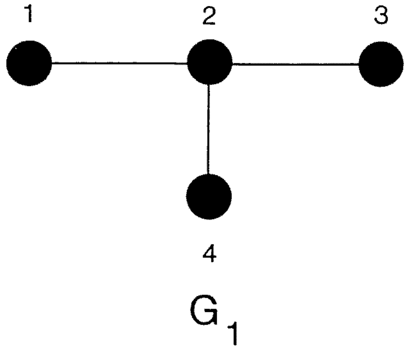

Fig. 3. Labeled hydrogen-suppressed graph of isobutane.

The adjacency matrix $A(G_1)$ and the distance matrix $D(G_1)$ of the labeled graph G_1 (fig. 3) of isobutane are given below:

$$A(G_1) = \begin{array}{c|cccc|} & (1) & (2) & (3) & (4) \\ 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 0 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 \end{array},$$

$$D(G_1) = \begin{array}{c|cccc|} & (1) & (2) & (3) & (4) \\ 1 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 2 & 1 & 0 & 2 \\ 4 & 2 & 1 & 2 & 0 \end{array}.$$

From the adjacency matrix of a graph with n vertices it is possible to calculate δ_i, the degree of the ith vertex, as the sum of all entries in the ith row:

$$\delta_i = \sum_{j=1}^{n} a_{ij} . \tag{1}$$

Zero-order connectivity index ${}^0\chi$ is defined as [12,64]:

$${}^0\chi = \sum_i (\delta_i)^{-1/2}. \tag{2}$$

Randić's connectivity index ${}^1\chi$ is defined as [4]:

$${}^1\chi = \sum_{\text{all edges}} (\delta_i \, \delta_j)^{-1/2}. \tag{3}$$

A generalized connectivity index ${}^h\chi$ considering paths of the type $v_0, v_1, \ldots, v_h$ of length h in the molecular graph is calculated as [12,64]:

$$ {}^h\chi = \sum(\delta_{v0}\,\delta_{v1}\ldots\delta_{vh})^{-1/2}, \tag{4} $$

where the summation is taken over all paths of length h.

Cluster, path-cluster, and cyclic types of simple connectivity indices are calculated using the method of Kier and Hall [12,64].

Valence connectivity indices are based on vertex-weighted graphs, where the weight δ_i^{v} of the ith vertex is calculated as follows [12,64]:

$$ \delta_i^{\mathrm{v}} = (Z_i^{\mathrm{v}} - h_i)/(Z_i - Z_i^{\mathrm{v}} - 1), \tag{5} $$

where Z_i^{v} is the number of valence electrons and Z_i is the atomic number of the atom represented by the ith vertex of the chemical graph and h_i is the number of hydrogen atoms attached to it. Valence connectivity indices ${}^h\chi^{\mathrm{v}}$ are calculated by replacing δ_i in eqs. (2)–(4) with δ_i^{v}. It is to be noted, however, that in the case of certain atoms, e.g. chlorine, bromine, iodine, fluorine, sulfur, etc., the δ_i^{v} values used are derived empirically through calibration with physicochemical properties [12,64]. The physical and/or graph-theoretic basis for these empirical adjustments remains far from clear.

The K_h ($h = 0, 1, \ldots, 10$) parameters used in this paper represent the number of occurrences of paths of length h in the hydrogen-depleted molecular graph of G. K_0 is the number of vertices and K_1 is the number of edges of G. Higher-order K_h terms can be calculated using graph-theoretic algorithms.

W is calculated as:

$$ W = \frac{1}{2}\sum_{i,j} d_{ij} = \sum_{h} h\cdot g_h\,, \tag{6} $$

where g_h is the number of unordered pairs of vertices whose distance is h.

4.2. CALCULATION OF MOLECULAR COMPLEXITY INDICES

Molecular complexity indices comprise another set of descriptors from molecular graphs [9,11,16,28–31,41]. The science of information theory has grown mainly out of the pioneering studies of Shannon [65], Wiener [66], Ashby [67], and Kolmogorov [68]. There is more than one version of information theory [69]. In Shannon's [65] statistical information theory, information is measured as reduced uncertainty of the system. In the algorithmic theory of Kolmogorov [68], the quantity of information is defined as the minimal length of a program which allows a one-to-one transformation of an object (set) into another. In applying information-theoretic formalism on chemical graphs, one looks upon the information content (or complexity) of a graph as a measure of its degree of variety or heterogeneity, as suggested by

Ashby [67]. An appropriate set S of n elements is derived from a molecular graph G depending on certain criteria. On the basis of an equivalence relation defined on A, the set A is partitioned into equivalence classes A_i of order n_i ($i = 1, 2, \ldots, h$, $\sum_i n_i = n$). A probability scheme is then assigned to the set of equivalence classes:

$$\begin{array}{l} A_1, A_2, \ldots, A_h, \\ p_1, p_2, \ldots, p_h, \end{array}$$

where $p_i = n_i/n$, n_i and n being the cardinalities of A_i and A, respectively. The mean information content (or complexity) of an element of A is defined by Shannon's [65] relation:

$$IC = -\sum_i p_i \log_2 p_i . \tag{7}$$

The logarithm is taken at base 2 for measuring the information content in bits. The total complexity of the set A is then n times IC.

It is to be noted that the complexity of a real object or a model object is not uniquely defined. While there could be more than one way of defining a model object [59,60] corresponding to the same piece of reality, complexity of the same model object may vary depending on the nature of the equivalence relation. In science, we deal with equivalence classes of events generated by grouping actual events and ignoring, at the same time, some unique properties of those events [60]. For example, when A represents the vertex set of a chemical graph G, two methods of partitioning have been widely used: (a) chromatic-number coloring of G, where two vertices of the same color are considered equivalent, and (b) determination of the transitive sets or orbits of the automorphism group of G, whereafter vertices are considered equivalent if they belong to the same orbit [70–73].

Rashevsky [74] symbolized molecules by simple linear graphs and calculated molecular complexity. In this approach, two vertices u and v of a graph G are said to be topologically equivalent if and only if for each neighboring vertex u_i ($i = 1, 2, \ldots, k$) of the vertex u there is a distinct neighboring vertex v_i of the same degree for the vertex v. Subsequently, various authors have computed complexity of molecules where linear graphs [28,41,70–74] or multigraphs [75] with indistinguishable vertices were used to symbolize the chemical species. On the other hand, to account for the unique nature of atoms and their bonding pattern in a molecule, Sarkar et al. [30], Roy et al. [31], and Basak et al. [20] calculated complexity of graphs on the basis of equivalence relations where both the nature of the atom (vertex) and the number and chemical nature of bonded neighbors of all atoms are taken into account. This was accomplished by defining open spheres for all vertices of the molecular graph [76]. If r is any nonnegative real number and v is a vertex of the graph G, then the open r-sphere $S(v, r)$ is defined as the subset of $V(G)$ consisting of all vertices v_i such that $d(v, v_i) < r$. Obviously, $S(v, 0) = \varnothing$, $S(v, r) = v$ for $0 < r < 1$, and $S(v, r) = (v) \cup \Gamma^1(v) = N^1(v)$ for $0 < r < 2$. One can construct open r-spheres of each vertex of G for all integral values of r, $0 \le r \le \rho$. For a particular value of r, the collection of all such open spheres $S(v, r)$, where v runs over the entire vertex set V, forms a neighborhood system of the vertices of G. A suitably defined equivalence relation can then partition V into disjoint subsets based on the equivalence of nature, connectedness, and bonding pattern of neighbors up to rth order neighborhoods [31]. It is noteworthy that this approach incorporates the effects of distant neighbors (i.e. neighbors of immediately bonded neighbors) on an atom or a reaction center. After partitioning of the vertices for a particular order (r) of neighborhood, IC_r is calculated by eq. (7). Subsequently, Basak, Roy and Ghosh [29] defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$SIC_r = IC_r/\log_2 n, \tag{8}$$

where IC_r is calculated by eq. (7) and n is the total number of vertices of the graph. It is noted that SIC_r is related to Brillouin's [77] measure of redundancy of a system. Another information-theoretic invariant, complementary information content (CIC_r) was defined as [78]:

$$CIC_r = \log_2 n - IC_r. \tag{9}$$

The Wiener index W [75], and the information-theoretic indices I_D^W and $\overline{I_D^W}$ are calculated from the distance matrix of chemical graphs [13]. The set of topological indices used in this paper are shown in table 2. Topological parameters were calculated by the computer program POLLY [79] where SMILES line notation [80] is the input.

Table 2

Definition and symbols for graph invariants

W	Half-sum of the off-diagonal elements of the distance matrix of a graph.
I_D^W	Information index for the magnitudes of the distances between all possible pairs of vertices of a graph.
$\overline{I_D^W}$	Mean information index for the magnitude of the distance.
IC_r	Mean information content or complexity of a graph based on the rth ($r = 0, 1, \ldots, 6$) order neighborhood of vertices in a graph.
SIC_r	Structural information content of a graph based on the rth ($r = 0, 1, \ldots, 6$) order neighborhood of vertices.
CIC_r	Complementary information content of a graph G calculated from the rth ($r = 0, 1, \ldots, 6$) order neighborhood of vertices.
${}^h\chi$	Path terms of hth order ($h = 0, 1, \ldots, 6$) calculated from δ values.
${}^h\chi_c$	Cluster terms of hth order ($h = 3, \ldots, 6$) based on δ values.
${}^h\chi_{PC}$	Path-cluster terms of hth order ($h = 4, \ldots, 6$) calculated for simple graphs.
${}^h\chi_{CH}$	Chain or cycle terms of different order ($h = 3, \ldots, 6$) based on δ values.
${}^h\chi^v$	Valence connectivity type path terms of hth order ($h = 0, 1, \ldots, 6$) calculated from δ^v values.
${}^h\chi^v_c$	Valence connectivity type cluster terms of hth order ($h = 3, \ldots, 6$) based on δ^v values.
${}^h\chi^v_{PC}$	Valence connectivity type path-cluster terms of hth order ($h = 4, \ldots, 6$).
${}^h\chi^v_{CH}$	Valence connectivity type chain or cycle terms of hth order ($h = 3, \ldots, 6$).
K_h	Number of paths of length h ($h = 0, 1, \ldots, 10$) in the hydrogen depleted graph.

5. Statistical analysis

We calculated TIs (table 2) for three different sets of compounds consisting of 37 polycyclic aromatic hydrocarbons, 29 alkylbenzenes, and 74 alkanes, respectively. Two statistical methods, viz. multiple regression analysis and principal components analysis (PCA) were used for analysis of the data.

5.1. MULTIPLE REGRESSION ANALYSIS

We used the maximum $\mathbb{R}^2$ improvement method [81] to identify prediction models. This method finds the "best" one variable model, the "best" two variable model and so forth for the prediction of properties. Analyses of the algorithmically-defined variables were exploratory in the sense that several models (e.g. combinations of variables) were examined to identify combinations of variables with good prediction capabilities. In all regression models developed, we examined a variety of statistics associated with the residuals, e.g. the Wilks–Shapiro test for normality [82] and Cook's D statistic for outliers [83], to obtain the most reasonable model.

5.2. PRINCIPAL COMPONENTS ANALYSIS (PCA)

The data for each of the four sets of chemicals analyzed in this paper may be viewed as n (number of chemicals) vectors in p (number of calculated parameters) dimensions. The data for each set can be represented by a matrix X which has n rows and p columns. For all the compounds, the number of calculated parameters was 71 (TIs of table 2). Each chemical is therefore represented by a point in $\mathbb{R}^{71}$. If each chemical could be represented in $\mathbb{R}^2$, then one could plot and investigate the extent of relationship between individual parameter and molecular properties. In $\mathbb{R}^{71}$, such a simple analysis is not possible. However, since many of the TIs are highly inter-correlated [1,43–45], the points in $\mathbb{R}^{71}$ can likely be represented by a subspace of fewer dimensions. The method of principal components analysis (PCA) or the Karhunen–Loève transformation is a standard method for reduction of dimensionality [84–86]. The first principal component (PC) is the line which comes closest to the points in the sense of minimizing the sum of the squared Euclidean distances from the points to the line. The second PC is given by projections onto the basis vector orthogonal to the first PC. For points in $\mathbb{R}^p$, the first r principal components give the subspace which comes closest to approximating the n points. The first PC is the first axis of the points. Successive axes are major directions orthogonal to previous axes. The PCs are the closest approximating hyperplane and because they are calculated from eigenvectors of a $p \times p$ matrix, the computations are relatively accessible. However, there are important scaling choices, because PCs are scale dependent. To control this dependence, the most commonly used convention is to rescale the variables so that each variable has mean zero and standard deviation one. The covariance matrix for these rescaled variables is the correlation matrix.

In addition to changes in linear scaling, nonlinear changes in scale such as logarithmic scale affect the PCA. Outliers will have a large effect on a best fitting plane when using Euclidean distance. For distributions which are positive and highly skewed by large values, a logarithmic transformation can be useful in reducing the importance of outliers. For the data in this investigation, the indices have been transformed by taking the log of the index plus one and then standardizing to mean zero and variance one.

No scaling convention for the original variables or the PCs is by definition correct. The task is to find a convention which works for the problem at hand. The choice made in this investigation was to standardize the log transformed indices and to use standardized principal components.

Another choice in the reduction of dimensionality is to choose the number of principal components retained. Using standardized variables, the variances of the unstandardized PCs are given by their eigenvalues, and the sum of the eigenvalues is p, the trace of the correlation matrix or the sum of the variances of the individual standardized variables. If all p standardized variables were uncorrelated, all eigenvalues would be 1.0. The eigenvalue of a PC divided by p is referred to as the variance explained by that PC. The cumulative variance explained by the first r PCs is the sum of their eigenvalues divided by p. The hope of PCA is to explain a large percentage of the total variance using a small number of PCs. In choosing the number of PCs retained, there are a number of possible conventions. The convention chosen here was to retain the PCs with eigenvalues greater than one [86].

6. Results

In this paper we analyzed three sets of chemicals in an attempt to investigate the extent algorithmically derived structural parameters are capable of predicting molecular properties.

6.1. ALKANES

Table 3 gives values of boiling points of 74 alkanes, the data being taken from the collection of Needham et al. [1]. Table 4 summarizes the result of PCA, which shows that the first three PCs explain 92.9% of the variance in the original data. The results of multiple regression analysis for the prediction of boiling point using PCs derived from TIs are shown in table 5. It is evident that PCs correlate highly with the boiling point data. Table 6 presents correlation coefficients of the first three PCs ($PC_1 - PC_3$) with the ten mostly correlated indices. The first PC is highly correlated with K_1, CIC_0, I_D^W, W, ${}^1\chi$, ${}^2\chi$, SIC_0 and K_0 parameters. These parameters are related to the size of the molecular graph. The second PC is highly correlated with a group of complexity parameters, viz. IC_2, IC_3, SIC_2, SIC_3, CIC_2, CIC_3, which measure the degree of heterogeneity of atomic environments in the molecule. The third PC is mainly correlated with SIC_1 and higher-order terms like ${}^3\chi_c$ and K_5.

Table 3

Boiling points and predicted boiling points for 74 alkanes

Observation	Name[a]	Boiling point [°C]	Predicted boiling point[b]
1	2	- 88.6	- 103.0
2	3	- 42.1	- 51.4
3	4	0.5	- 2.2
4	2M3	- 11.7	- 6.7
5	5	36.1	40.2
6	2M4	27.8	29.6
7	22MM3	9.5	24.0
8	6	68.7	75.8
9	2M5	60.3	62.0
10	3M5	63.3	61.9
11	22MM4	49.7	57.7
12	23MM4	58.0	64.0
13	7	98.4	104.9
14	2M6	90.0	94.4
15	3M6	91.8	94.6
16	3E5	93.5	88.0
17	22MM5	79.2	86.0
18	23MM5	89.8	90.5
19	24MM5	80.5	85.0
20	33MM5	86.1	86.6
21	223MMM4	80.9	86.2
22	8	125.7	128.6
23	2M7	117.6	118.6
24	3M7	118.9	121.0
25	4M7	117.7	118.1
26	3E6	118.5	119.0
27	22MM6	106.8	115.2
28	23MM6	115.6	116.4
29	24MM6	109.4	115.6
30	25MM6	109.1	112.2
31	33MM6	112.0	115.5
32	34MM6	117.7	114.4
33	23ME5	115.6	111.6
34	33ME5	118.3	110.8
35	223MMM5	109.8	107.8
36	224MMM5	99.2	104.2
37	233MMM5	114.8	108.5
38	234MMM5	113.5	117.5
39	2233MMMM4	106.5	106.3
40	9	150.8	149.1
41	2M8	143.3	139.8
42	3M8	144.2	142.1
43	4M8	142.5	141.6
44	3E7	143.0	142.0

. . . continued

Table 3 (continued)

Observation	Name[a]	Boiling point [°C]	Predicted boiling point[b]
45	4E7	141.2	141.5
46	22MM7	132.7	135.7
47	23MM7	140.5	137.8
48	24MM7	133.5	136.0
49	25MM7	136.0	137.1
50	26MM7	135.2	132.0
51	33MM7	137.3	139.6
52	34MM7	140.6	138.3
53	35MM7	136.0	137.3
54	44MM7	135.2	135.8
55	23ME6	138.0	138.4
56	24ME6	133.8	137.6
57	33ME6	140.6	138.5
58	34ME6	140.4	136.3
59	223MMM6	133.6	132.6
60	224MMM6	126.5	131.4
61	225MMM6	124.1	130.9
62	233MMM6	137.7	132.7
63	234MMM6	139.0	139.6
64	235MMM6	131.3	138.9
65	244MMM6	130.6	131.7
66	334MMM6	140.5	132.0
67	33EE5	146.2	130.0
68	223MME5	133.8	125.6
69	233MME5	142.0	128.4
70	234MEM5	136.7	133.0
71	2233(M)5	140.3	130.6
72	2234(M)5	133.0	133.1
73	2244(M)5	122.3	123.3
74	2334(M)5	141.6	132.5

[a] 2 = Ethane, 3 = Propane, etc.; M = methyl, E = Ethyl; 24ME6 = 2-Methyl-4-Ethyl hexane; structures are given in fig. 4.

[b] Predicted from the three-parameter model (involving the first three PCs) of table 5.

Table 4

Summary of principal components for alkanes

PC	Eigenvalue	Percent of variance	Cumulative percent
1	17.1	65.9	65.9
2	5.4	20.8	86.8
3	1.6	6.1	92.9

Table 5

Multiple regression analysis for prediction of boiling points of alkanes using principal components (PCs) derived from topological indices (TIs)

Step	Variables	n	R^2	F	Standard error of estimate
1	PC_1	74	95.9	1686	9.4
2	PC_1, PC_3	74	98.1	1828	6.5
3	PC_1, PC_2, PC_3	74	98.6	1608	5.7

Table 6

Correlation coefficients of TIs with principal components for alkanes

PC_1		PC_2		PC_3	
K_1	0.99	CIC_2	−0.95	SIC_1	0.68
CIC_0	0.99	SIC_2	0.90	$^3\chi_c$	0.62
K_0	0.99	CIC_3	−0.90	K_5	−0.53
SIC_0	−0.99	SIC_3	0.78	IC_1	0.41
I_D^W	0.99	IC_2	0.68	$^2\chi$	0.28
W	0.99	SIC_1	0.66	K_2	0.24
$^1\chi$	0.98	IC_3	0.63	K_4	−0.21 ns[a]
$^0\chi$	0.98	$^3\chi_c$	−0.57	CIC_1	−0.18 ns
IC_0	0.97	$^2\chi$	−0.37	$^1\chi$	−0.18 ns
CIC_1	0.95	K_5	0.36	W	−0.12 ns

[a] ns = not significant.

6.2. ALKYLBENZENES

The values of boiling points for alkylbenzenes are given in table 7, the data being taken from the collection of Mekenyan et al. [11]. Table 8 summarizes the result of PCA, which shows that the first five PCs explain 95.7% of variance in the original data. Results of multiple regression analysis using PCs are given in table 9. It is clear from the result of regression analysis that PCs derived from TIs are capable of predicting boiling points of this group of compounds satisfactorily. Table 10 presents correlations of the first five PCs with the ten individual TIs most highly correlated with the PCs. The pattern of correlation of TIs with the first two PCs seems to be analogous to that observed from alkanes. PC_3 is highly correlated with cluster and path/cluster terms, PC_4 is mainly correlated with higher order cluster terms ($^4\chi_c$, $^6\chi_c$, $^4\chi_c^v$, $^6\chi_c^v$). PC_5 is not highly correlated with any particular group of indices.

Table 7

Boiling points and predicted boiling points for alkylbenzenes

Observation	Name	Boiling point [°C]	Predicted boiling point[a]
1	Benzene	80.1	80.9
2	Toluene	110.6	110.7
3	Ethylbenzene	136.2	134.7
4	O-Xylene	144.4	142.0
5	m-Xylene	139.1	140.9
6	p-Xylene	138.4	140.5
7	n-Propylbenzene	159.2	156.0
8	1-Methyl-2-ethylbenzene	165.2	164.2
9	1-Methly-3-ethylbenzene	161.3	162.8
10	1-Methyl-4-ethlybenzene	162.0	163.2
11	1,2,3-Trimethylbenzene	176.1	171.2
12	1,2,4-Trimethylbenzene	169.4	170.2
13	1,3,5-Trimethylbenzene	164.7	168.3
14	n-Butylbenzene	183.3	175.8
15	1,2,-Diethylbenzene	183.4	186.4
16	1,3-Diethylbenzene	181.1	183.9
17	1,4-Diethylbenzene	183.8	182.8
18	1-Methyl-2-n-propylbenzene	184.8	184.9
19	1-Methyl-3-n-propylbenzene	181.8	182.7
20	1-Methyl-4-n-propylbenzene	183.8	182.1
21	1,2-Dimethyl-3-ethylbenzene	193.9	191.4
22	1,2-Dimethyl-4-ethylbenzene	189.8	190.5
23	1,3-Dimethyl-2-ethylbenzene	190.0	191.9
24	1,3-Dimethyl-4-ethylbenzene	188.4	191.1
25	1,3-Dimethyl-5-ethylbenzene	183.8	189.6
26	1,4-Dimethyl-2-ethylbenzene	186.9	190.8
27	1,2,3,4-Tetramethylbenzene	205.0	197.7
28	1,2,3,5-Tetramethylbenzene	198.2	197.1
29	1,2,4,5-Tetramethylbenzene	196.8	196.8

[a] Predicted from the three-parameter model of table 9.

Table 8

Summary of principal components for alkylbenzenes

PC	Eigenvalue	Percent of variance	Cumulative percent
1	34.2	57.0	57.0
2	15.2	25.3	82.3
3	4.6	7.6	89.9
4	2.4	4.0	93.9
5	1.1	1.9	95.7

Table 9

Multiple regression analysis for prediction of boiling points of alkylbenzenes using principal components (PCs) derived from topological indices (TIs)

Step	Variables	n	R^2	F	Standard error of estimate
1	PC_1	29	93.3	378	7.3
2	PC_1, PC_2	29	97.6	526	4.5
3	PC_1, PC_2, PC_3	29	98.8	668	3.3
4	PC_1, PC_2, PC_3, PC_5	29	99.1	632	2.9
5	$PC_1, PC_2, PC_3, PC_4, PC_5$	29	99.2	584	2.7

Table 10

Correlation coefficients of TIs with PCs for alkylbenzenes

PC_1		PC_2		PC_3		PC_4		PC_5	
$\overline{I}_D^W$	0.98	CIC_3	0.92	SIC_0	– 0.68	${}^5\chi_c$	– 0.49	K_9	0.59
${}^0\chi^v$	0.98	CIC_2	0.88	K_9	0.52	SIC_0	0.49	SIC_0	0.36 ns[a]
${}^0\chi$	0.98	SIC_3	– 0.85	K_8	0.51	${}^5\chi_c^v$	– 0.49	IC_0	0.24 ns
K_0	0.97	CIC_4	0.83	K_7	0.45	SIC_1	0.42	${}^5\chi_c$	0.24 ns
K_1	0.97	CIC_5	0.80	${}^3\chi_c$	– 0.45	CIC_1	– 0.36 ns[a]	${}^5\chi_c^v$	0.24 ns
I_D^W	0.96	CIC_6	0.80	${}^3\chi_c^v$	– 0.45	IC_1	0.36 ns	${}^6\chi^v$	– 0.23 ns
K_6	0.96	SIC_2	– 0.80	${}^4\chi_{PC}$	– 0.38	IC_0	0.35 ns	SIC_1	0.21 ns
K_4	0.96	CIC_1	0.79	${}^6\chi_{CH}^v$	0.37	CIC_5	0.31 ns	K_8	0.20 ns
W	0.95	SIC_4	– 0.75	${}^6\chi_{CH}$	0.37	CIC_6	0.31 ns	IC_1	0.18 ns
K_2	0.95	SIC_5	– 0.74	CIC_1	0.37	CIC_4	0.29 ns	${}^6\chi$	– 0.17 ns

[a] ns = not significant.

6.3. POLYCYCLIC AROMATIC HYDROCARBONS (PAHs)

Table 11 gives the values of boiling points for 37 PAHs (structures in fig. 4), the data being taken from the collection of Karcher [87]. Table 12 presents a summary of PCA which shows that the first four PCs explain 96.8% of the variance in the calculated TIs. Correlations of PCs with boiling points of PAHs are shown in table 13. In case of PAHs also, the PCs are significantly correlated with the boiling point data. Table 14 presents the data on correlations of TIs with the first four PCs. The pattern of correlation is almost the same, with a notable exception. In case of PAHs, the first PC is also highly correlated with ${}^3\chi_c$ (0.98), ${}^3\chi_c^v$ (0.97), ${}^5\chi_c$ (0.97), and ${}^5\chi_c^v$ (0.96) parameters. ${}^5\chi_c$ and ${}^3\chi_c$ terms are derived from a class of subgraphs shown in fig. 5. The preponderance of these subgraphs in the structures of PAHs might be the reason for this shift in the correlation pattern of TIs with PC_1 for PAHs.

Table 11

Boiling points and predicted boiling points for polycyclic aromatic hydrocarbons (PAHs)

		Boiling point	Predicted
Observation	Name	[°C]	boiling point[a]
1	Naphthalene	218	188
2	Acenaphthylene	270	277
3	Acenaphthene	279	258
4	Fluorene	294	320
5	Phenanthrene	338	340
6	Anthracene	340	327
7	4H-Cyclopenta(def)phenanthrene	359	361
8	Fluoranthene	383	398
9	Pyrene	393	410
10	Benzo(a)fluorene	403	411
11	Benzo(b)fluorene	398	370
12	Benzo(c)fluorene	406	411
13	Benzo(ghi)fluoranthene	422	460
14	Cyclopenta(cd)pyrene	439	441
15	Chrysene	431	446
16	Benz(a)anthracene	425	443
17	Triphenylene	429	441
18	Naphthacene	440	430
19	Benzo(b)fluoranthene	481	492
20	Benzo(j)fluoranthene	480	492
21	Benzo(k)fluoranthene	481	479
22	Benzo(a)pyrene	496	503
23	Benzo(e)pyrene	493	503
24	Perylene	497	496
25	Anthanthrene	547	551
26	Benzo(ghi)perylene	542	553
27	Indeno(1,2,3-cd)fluoranthene	531	522
28	Indeno(1,2,3-cd)pyrene	534	539
29	Dibenz(a,c)anthracene	535	523
30	Dibenz(a,h)anthracene	535	520
31	Dibenz(a,j)anthracene	531	520
32	Picene	519	529
33	Coronene	590	590
34	Dibenzo(a,e)pyrene	592	579
35	Dibenzo(a,h)pyrene	596	568
36	Dibenzo(a,i)pyrene	594	568
37	Dibenzo(a,l)pyrene	595	578

[a] Predicted from the three-parameter model of table 13.

Table 12

Summary of principal components analysis for polycyclic aromatic hydrocarbons

PC	Eigenvalue	Percent of variance	Cumulative percent
1	40.2	63.9	63.9
2	15.1	23.9	87.8
3	3.9	6.2	94.0
4	1.8	2.9	97.0

Table 13

Multiple regression analysis for predicting boiling points of polycyclic aromatic hydrocarbons (PAHs) using principal components derived from topological indices

Step	Variables	n	R^2	F	Standard error of estimate
1	PC_1	37	93.0	466	26.7
2	PC_1, PC_2	37	95.3	346	22.2
3	PC_1, PC_2, PC_3	37	97.5	422	16.6
4	PC_1, PC_2, PC_3, PC_4	37	97.5	316	16.6

Table 14

Correlation coefficients of TIs with PCs for polycyclic aromatic hydrocarbons

PC_1		PC_2		PC_3		PC_4	
${}^3\chi^v$	0.99	CIC_2	-0.96	${}^5\chi^v_{CH}$	0.74	IC_0	0.70
K_2	0.99	CIC_3	-0.96	${}^5\chi_{CH}$	0.71	SIC_0	0.63
${}^4\chi$	0.99	SIC_2	0.95	${}^6\chi^v_{CH}$	0.68	${}^6\chi_{CH}$	-0.49
K_3	0.99	SIC_3	0.95	${}^6\chi_{CH}$	0.59	${}^6\chi^v_{CH}$	-0.43
${}^3\chi$	0.99	SIC_4	0.93	SIC_0	0.56	${}^5\chi_{CH}$	-0.38
${}^2\chi^v$	0.99	SIC_5	0.92	IC_0	0.45	${}^5\chi^v_{CH}$	-0.31 ns[a]
${}^4\chi_{PC}$	0.99	CIC_4	-0.92	CIC_1	-0.42	CIC_0	-0.25 ns
${}^4\chi^v$	0.99	SIC_6	0.92	SIC_1	0.41	SIC_1	0.23 ns
${}^5\chi$	0.99	CIC_5	-0.90	CIC_0	-0.35	IC_1	-0.22 ns
K_4	0.99	CIC_6	-0.89	IC_1	0.30 ns[a]	CIC_1	-0.19 ns

[a] ns = not significant.

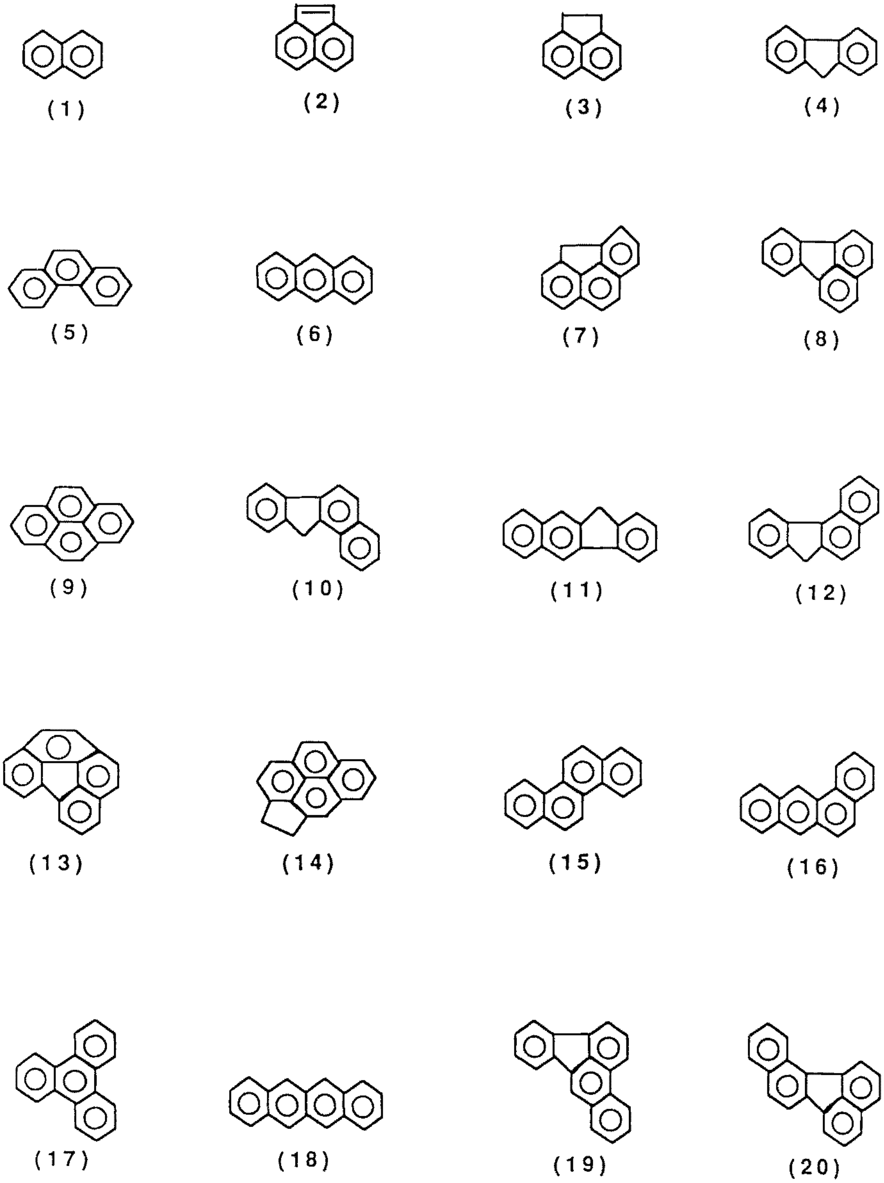

Fig. 4. (caption on following page).

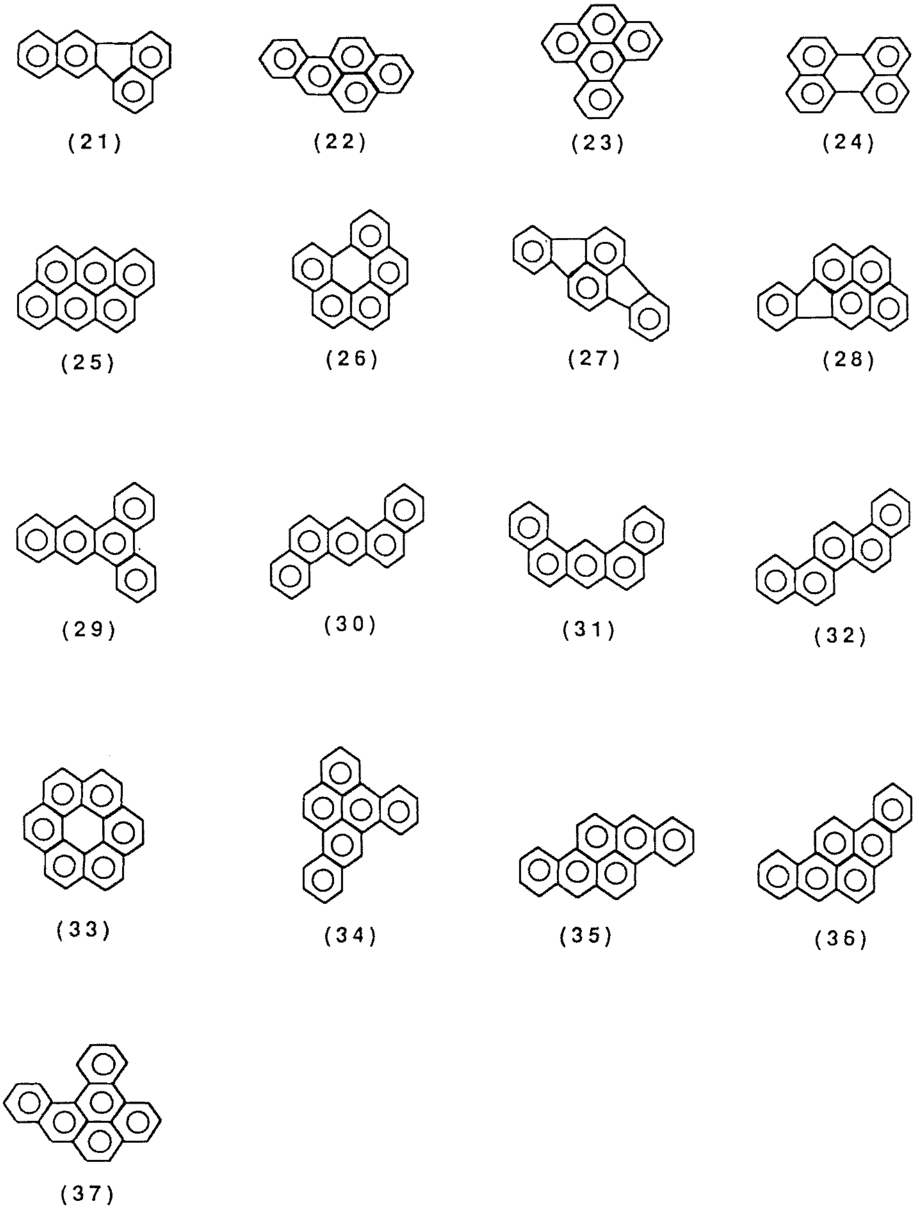

Fig. 4. Hydrogen-depleted graphs of polycyclic aromatic hydrocarbons.

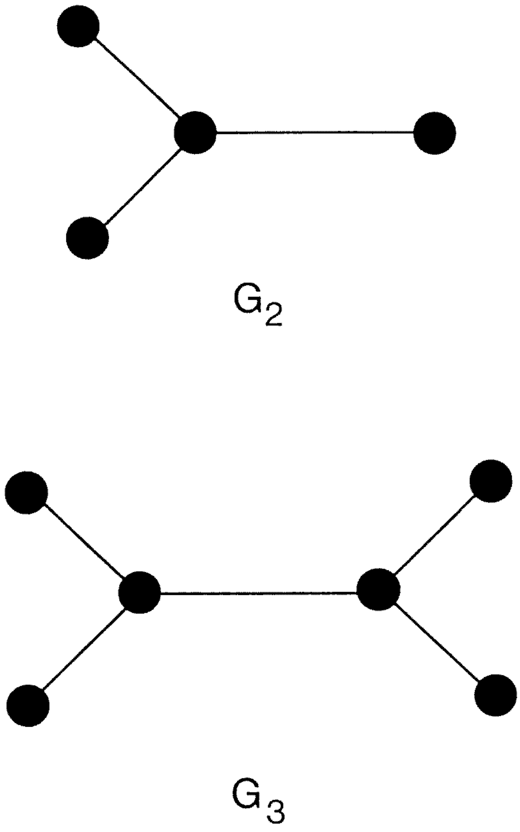

Fig. 5. G_2 and G_3 represent subgraphs corresponding to $^3\chi_c$ and $^5\chi_c$ terms, respectively, of connectivity indices.

6.4. THE COMBINED SET OF ALKANES, ALKYLBENZENES AND PAHs

It was of interest to see how far the PCs can predict boiling points of the diverse set (n = 140) consisting of 74 alkanes, 29 alkylbenzenes and 37 PAHs. Table 15 gives the summary of PCA, which shows that the first six PCs with eigenvalues greater than one explain 95.2% of the variance in the calculated TIs. Correlations of boiling points with PCs of this combined set are shown in table 16. It is clear from the results that the level of predictability of boiling points of the combined set is inferior to the corresponding levels of predictability of boiling points for alkanes (table 5), alkylbenzenes (table 9), or PAHs (table 13).

Table 15

Summary of principal components for the combined set of alkanes, alkylbenzenes, and polycyclic aromatic hydrocarbons

PC	Eigenvalue	Percent of variance	Cumulative percent
1	40.0	59.6	59.6
2	10.8	16.2	75.8
3	6.1	9.2	85.0
4	3.2	4.8	89.8
5	1.9	2.8	92.6
6	1.7	2.6	95.2

Table 16

Multiple regression analysis for prediction of boiling points of the combined set of alkanes, alkylbenzenes, and polycyclic aromatic hydrocarbons

Step	Variables	n	R^2	F	Standard error of estimate
1	PC_1	140	91.1	1411	48.2
2	PC_1, PC_2	140	94.2	1108	39.1
3	PC_1, PC_2, PC_3	140	96.4	1229	30.7
4	PC_1, PC_2, PC_3, PC_4	140	97.7	1407	25.0
5	$PC_1, PC_2, PC_3, PC_4, PC_5$	140	97.9	1230	23.9

7. Discussion

The major objectives of this paper were: (1) to investigate how far we are able to extract information encoded by TIs in a smaller number of parameters, viz. principal components (PCs), and (2) to study how far we can develop models to predict molecular properties from TIs or PCs derived from TIs.

That many of the 71 TIs calculated for the three sets of compounds are highly intercorrelated is clear from the eigenvalues of PCs for the three groups of compounds and from correlations of TIs with the derived PCs (tables 6, 10, and 14). This is in line with our earlier observations [43–45] and findings of others [1,88,89]. That the 71-dimensional space consisting of 71 calculated TIs lies on a subspace of lower dimension is clear from PCA of the four sets of compounds. For the three sets of compounds, viz. alkanes, alkylbenzenes, and polycyclic aromatic hydrocarbons, the first three PCs explain 92.9%, 89.9%, and 94.0% of variances of the original data, respectively (tables 4, 8, and 12). In fact, for the first three groups of relatively homogeneous structures, the first few PCs with eigenvalues greater than one explain most of the variance in the original data. The results presented here show that we can extract useful information from predictor parameters (TIs) in a relatively small number of orthogonal axes (PCs).

Correlations of TIs with PC_1 show that the following indices are most highly correlated with PC_1 (tables 6, 10, and 14): K_0–K_3, W, I_D^W, $\overline{I}_D^W$, ${}^1\chi$, ${}^2\chi$, ${}^1\chi^v$–${}^3\chi^v$. These TIs are related to the generalized shape and size of molecular graphs. Analogous findings were reported by us earlier for other large databases [43–45].

PC_2 is mainly correlated with the higher-order complexity indices, viz. *IC*, *SIC* and *CIC* parameters. These indices quantitate the degree of heterogeneity of atomic environments in chemical graphs. Our earlier studies with a diverse set of 3692 chemicals showed a similar pattern of correlation for PC_2 [44,45].

PC_3 shows some differences in the pattern of correlations with TIs. For alkanes (acyclic molecules) and alkylbenzenes, PC_3 is highly correlated with path-cluster terms of molecular connectivity indices. However, for PAHs, PC_3 is mainly correlated with cycle terms of connectivity parameters. This differential correlation of PC_3 seems to be dependent on the makeup of a particular database and captures a common and predominant feature present in a larger number of structures in that database. A similar conclusion is evident from correlations of cluster terms with PC_4 for alkylbenzenes (table 10) and cycle terms with PC_4 for PAHs (table 14).

Regarding the utility of PCs derived from TIs for SAR, it is clear from results presented in tables 5, 9, and 13 that for molecules with a reasonable range of structural variations, PCs are capable of predicting molecular properties. These results fall in the category of level 4 SAR or QSAR in our scheme of SAR (fig. 2). On the other hand, for the combined set of alkanes, alkylbenzenes, and PAHs, the boiling points could not be estimated as accurately (table 16) as with the individual structural classes (tables 5, 9, and 13) taken separately. The increased structural diversity of the combined set might be the reason for this observed difference of predictability of property. A similar observation was reported by us earlier for lipophilicity [90].

In conclusion, TIs encode useful structural information which can be conveniently extracted into orthogonal axes consisting of PCs. For structural classes of reasonable diversity, PCs performed very well in predicting boiling points, whereas the level of predictability decreased with increasing structural diversity of the chemicals under investigation.

Acknowledgements

The authors wish to express their appreciation to the US Environmental Protection Agency (Contract CR-815387) for support. Thanks to Anda Bellamy and Greg Grunwald for their assistance in the project. This paper is contribution number 57 of the Center for Water and the Environment.

References

[1] D.E. Needham, I.C. Wei and P.G. Seybold, J. Amer. Chem. Soc. 110(1988)4186.

[2] D.H. Rouvray and W. Tatong, Z. Naturforsch. 41a(1986)1238.

[3] Y. Gao and H. Hosoya, Bull. Chem. Soc. Japan 61(1988)3093.

[4] M. Randić, J. Amer. Chem. Soc. 97(1975)6609.

[5] O. Mekenyan, S. Dimitrov and D. Bonchev, Eur. Polymer J. 19(1983)1185.

[6] S.C. Basak, G.J. Niemi and G.D. Veith, in: *Computational Chemical Graph Theory*, ed. D.H. Rouvray (Nova Publ., New York, 1990), in press.

[7] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Ind. J. Chem. 20B(1981)894.

[8] S.K. Ray, S.C. Basak, C. Raychaudhury, A.B. Roy and J.J. Ghosh, Arzneim. Forsch. 32(1982)322.

[9] V.R. Magnuson, D.K. Harriss and S.C. Basak, in: *Chemical Applications of Topology and Graph Theory*, ed. R.B. King (Elsevier, Amsterdam, 1983), pp. 178.

[10] R.J. Baker, W.E. Acree and C.-C. Tsai, Quant. Struct.-Act. Relat. 3(1984)10.

[11] O. Mekenyan, D. Bonchev and N. Trinajstić, Int. J. Quant. Chem. 18(1980)369.

[12] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure–Activity Analysis* (research Studies Press, Letchworth, Hertfordshire, UK, 1986).

[13] N. Trinajstić, *Chemical Graph Theory*, Vols. 1 and 2 (CRC Press, Boca Raton, FL, 1983).

[14] J.W. Kennedy and L.V. Quintas, *Applications of Graphs in Chemistry and Physics* (North-Holland, Amsterdam, 1988).

[15] M. Randić, Int. J. Quant. Chem. Quant. Biol. Symp. 11(1984)137.

[16] A. Sabljic and N. Trinajstić, Acta Pharm. Yugosl. 31(1981)189.

[17] S.C. Basak, D.P. Gieschen, D.K. Harriss and V.R. Magnuson, J. Pharm. Sci. 72(1983)934.

[18] S.C. Basak, L.J. Monsrud, M.E. Rosen, C.M. Frane and V.R. Magnuson, Acta Pharm. Yugosl. 36(1986)81.

[19] S.C. Basak, Med. Sci. Res. 15(1987)605.

[20] S.C. Basak, Med. Sci. Res. 16(1988)281.

[21] N. Trinajstić, M. Randić and D.J. Klein, Acta Pharm. Yugosl. 36(1986)267.

[22] F. Harary, *Graph Theory* (Addison–Wesley, Reading, MA, 1969).

[23] F. Harary, J. Graph Theory 10(1986)iii.

[24] A.T. Balaban, J. Chem. Inf. Comput. Sci. 25(1985)334.

[25] L. Spialter, J. Amer. Chem. Soc. 85(1963)2012.

[26] L. Spialter, J. Chem. Doc. 4(1964)261.

[27] L. Spialter, J. Chem. Doc. 4(1964)269.

[28] D. Bonchev and N. Trinajstić, J. Chem. Phys. 67(1977)4517.

[29] S.C. Basak, A.B. Roy and J.J. Ghosh, in: *Proc. 2nd Int . Conf. on Mathematical Modelling*, ed. X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler (University of Missouri-Rolla, Rolla, MO, 1980), Vol. 2, p. 851.

[30] R. Sarkar, A.B. Roy and P.K. Sarkar, Math. Biosci. 39(1978)299.

[31] A.B. Roy, S.C. Basak, D.K. Harriss and V.R. Magnuson, in: *Mathematical Modelling in Science and Technology*, ed. X.J.R. Avula, R.E. Kalman, A.I. Liapis and E.Y. Rodin (Pergamon, New York, 1984), p. 745.

[32] G.A. Baker, J. Math. Phys. 7(1966)2238.

[33] A.T. Balaban and F. Harary, J. Chem. Doc. 11(1971)258.

[34] A.J. Schwenk, in: *New Directions in the Theory of Graphs*, ed. F. Harary (Academic Press, New York, 1973).

[35] M. Randić, MATCH 7(1979)5.

[36] L.V. Quintas and P.J. Slater, MATCH 12(1981)75.

[37] P.J. Slater, J. Graph Theory 6(1982)89.

[38] M. Randić, J. Chem. Inf. Comput. Sci. 24(1984)164.

[39] K. Szymanski, W.R. Muller, J.V. Knop and N. Trinajstić, J. Chem. Inf. Comput. Sci. 25(1985)413.

[40] D. Bonchev and N. Trinajstić, J. Comput. Chem. 2(1981)127.

[41] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy and S.C. Basak, J. Comput. Chem. 5(1984)581.

[42] A.T. Balaban, Chem. Phys. Lett. 89(1982)399.

[43] G.J. Niemi, R.R. Regal and G.D. Veith, in: *Environmental Applications of Chemometrics*, ed. J.J. Breen and P.E. Robinson (American Chemical Society, Washington, DC, 1985), p. 148.

[44] S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal and G.D. Veith, Math. Modelling 8(1987)300.

[45] S.C. Basak, V.R. Magnuson, G.J. Niemi and R.R. Regal, Discr. Appl. Math. 19(1988)17.

[46] S.C. Basak, in: *Practical Applications of Quantitative Structure–Activity Relationship (QSAR) Study in Environmental Chemistry and Toxicology*, ed. W. Karcher and J. DeVillers (Kluwer Academic, 1990), in press.

[47] M. Johnson, S.C. Basak and G. Maggiora, Math. Comput. Modelling 11(1988)630.

[48] S.C. Basak, G.J. Niemi and G.D. Veith, Math. Comput. Modelling, in press.

[49] C. Hansch, in: *Correlation Analysis in Chemistry*, ed. N.B. Chapman and J. Shorter (Plenum, New York, 1978), p. 397.

[50] J. Cairns, Jr. and D.I. Mount, Environ. Sci. Tech. 24(1990)154.

[51] R. Rosen, J. Theor. Biol. 18(1968)380.

[52] W. Elssasser, *The Physical Foundation of Biology* (Pergamon, Oxford, 1958).

[53] J. von Neuman, in: *Automata Studies*, ed. C.E. Shannon and J. McCarthy (Princeton University Press, Princeton, 1956).

[54] R. Rosen, Int. J. General Systems 3(1977)227.

[55] J.C. Arcos, Environ. Sci. Tech. 21(1987)743.

[56] G.D. Veith, D.J. Call and L.T. Brooke, Can. J. Fish Aquat. Sci. 40(1983)743.

[57] D.J. Klein, Int. J. Quant. Chem. Quant. Chem. Symp. 20(1986)153.

[58] N. Trinajstić, M. Randić and D.J. Klein, Acta. Pharm. Yugosl. 36(1986)267.

[59] M. Bunge, *Method, Model and Matter* (Reidel, Dordrecht/Boston, 1973).

[60] H. Primas, *Chemistry, Quantum Mechanics and Reductionism* (Springer, Berlin, 1981).

[61] S.J. Weininger, J. Chem. Educ. 61(1984)939.

[62] J.J. Sylvester, Amer. J. Math. 1(1878)64.

[63] C.L. Wilkins and M. Randić, Theor. Chim. Acta (Berl.) 58(1980)45.

[64] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).

[65] C.E. Shannon, Bell Syst. Tech. J. 27(1948)379.

[66] N. Wiener, *Cybernetics* (Wiley, New York, 1948).

[67] W. Ashby, *An Introduction to Cybernetics* (Wiley, New York, 1956).

[68] A.N. Kolmogorov, Probl. Peredachi. Inf. 5(1969)3.

[69] D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures* (Research Studies Press, Chichester, 1983).

[70] C.W. Marshall, *Applied Graph Theory* (Wiley–Interscience, New York, 1971).

[71] E. Trucco, Bull. Math. Biophys. 18(1956)129.

[72] A. Mowshowitz, Bull. Math. Biophys. 30(1968)175.

[73] A. Mowshowitz, Bull. Math. Biophys. 30(1968)225.

[74] N. Rashevsky, Bull. Math. Biophys. 17(1955)229.

[75] S.H. Bertz, Bull. Math. Biol. 45(1983)849.

[76] P.E. Long, *An Introduction to General Topology* (Charles E. Merrill Publ. Co., Columbus, OH, 1971).

[77] L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1956).

[78] S.C. Basak and V.R. Magnuson, Arzneim. Forsch. 33(1983)501.

[79] S.C. Basak, D.K. Harriss and V.R. Magnuson, POLLY: Copyright of the University of Minnesota (1988).

[80] E. Anderson, G.D. Veith and D. Weininger, Report No. EPA/600/M-87-021, Environmental Research Laboratory-Duluth, 1987.

[81] SAS/STAT user guide, release 6.03 edition (SAS Institute Inc., Cary, NC, 1988), p. 1028.

[82] S.S. Shapiro and M.B. Wilk, Biometrika 52(1965)591.

[83] R.D. Cook, Technometrics 19(1977)15.

[84] M.M. Tatsuoka, *Multivariate Analysis: Techniques for Educational and Psychological Research* (Wiley, New York, 1971).

[85] R. Gnanadesikan, *Methods for Statistical Analysis of Multivariate Observations* (Wiley, New York, 1977).

[86] M.J. Greenacre, *Theory and Applications of Correspondence Analysis* (Academic Press, New York, 1984).

[87] W. Karcher, *Spectral Atlas of Polycyclic Aromatic Compounds*, Vol. 2 (Kluwer Academic, Dordrecht/ Boston/London, 1988).

[88] I. Motoc, A.T. Balaban, O. Mekenyan and D. Bonchev, MATCH 13(1982)369.

[89] L.P. Burkhard, A.W. Andren and D.E. Armstrong, Chemosphere 12(1983)935.

[90] S.C. Basak, G.J. Niemi and G.D. Veith, J. Math. Chem. 4(1990)185.