Similarity methods in analog selection, property estimation and clustering of diverse chemicals

Subhash C. Basak,* Brian D. Gute, and Denise Mills

Natural Resources Research Institute, University of Minnesota Duluth,
5013 Miller Trunk Highway, Duluth, Minnesota 55811, USA
E-mail: sbasak@nrri.umn.edu

Abstract

This account summarizes Dr. Subhash Basak's work in the field of molecular similarity. In particular, it looks at the development and application of quantitative molecular similarity analysis (QMSA) techniques using physicochemical properties, topological indices, and atom pairs as descriptors for developing structure- or property-based similarity spaces and the use of a *k*-nearest neighbors (*k*NN) technique to estimate the properties or activities of chemicals within the space. Additionally, the account discusses the novel tailored similarity technique pioneered by Dr. Basak's research group and discusses the future of molecular similarity analysis techniques.

Keywords: Similarity, QMSA, topological indices, atom pair, tailoring, *k*NN technique

Contents

* Author to whom correspondence should be addressed. Tel: +218-720-4230; Fax: +218-720-4328; e-mail: sbasak@nrri.umn.edu

1. Introduction

The concepts of molecular similarity/dissimilarity have applications in various aspects of chemistry, pharmaceutical drug design, toxicology, and ecotoxicology. The similarity of objects is an intuitive notion, which has been used by human beings from time immemorial. In the realms of chemistry, biochemistry, and toxicology the notion of similarity is used in many ways. For example, a natural product chemist, in the aftermath of the chance discovery of a naturally occurring chemical possessing useful therapeutic effects, would want to know whether the chemical's analogs (similar molecules) also have similar therapeutic profiles. A positive answer will indicate that the molecule is not a lone example having the therapeutic activity, but belongs to a group of structures that possess the property of interest. This might lead to an understanding of the structural basis of the pharmacological action and help in the molecular design and development of more appealing chemical entities. To find analogs of the candidate chemical, one might search databases such as the Chemical Abstracts Service database[1] (27.5 million organic and inorganic substances) or other databases of natural products. At the current stage of scientific development, this cannot be done manually. One needs rapid, automated methods for measuring structural similarity. In the areas of toxicology and the hazard assessment of chemicals, similarity is used routinely. Most of the chemicals that are currently used for commercial and industrial purposes[2] have little or no test data, not even the simple physicochemical properties needed for their hazard estimation. In the face of this paucity of data, two approaches have been used: a) class-specific quantitative structure-activity relationship (QSAR) modeling, and b) modeling using structural analogs. The United States Environmental Protection Agency (USEPA) has class-specific QSAR models for more than 300 specific chemical classes. However, this covers only a small fraction of the Toxic Substances Control Act (TSCA) Inventory or the chemicals submitted in the Premanufacture Notification (PMN) process. If a chemical's toxicity is not amenable to prediction by class-specific QSARs, the default approach is to use its analogs in the estimation of the hazards posed by the chemical. Such analogs can be selected either subjectively by experts, based on their intuitive notions of structural similarity, or computationally, using some measures of intermolecular similarity based on specific structural aspects of the chemical under consideration.

The fundamental notions in the theory and practice of molecular similarity arise out of the structure-property similarity principal which states that *similar structures usually have similar properties*. The tacit assumption underlying this notion is that the relationships between the structures of molecules and their various physicochemical and biological properties are governed by smooth, although unknown, mathematical functions. In other words, the relationship between structure and property is such that small variations in molecular structure will lead to small perturbations in the magnitude of a given property, rather than to large, abrupt changes. Empirical observations ranging from the basic chemistry of the elements to the biochemical and toxicological effects of chemicals generally support such a notion. For example, in the periodic table of elements, two atoms belonging to the same elemental group are more similar to each

other with respect to physical properties and reactivity pattern than any two chosen from different groups. A useful concept in organic chemistry is the idea of the homologous series. The addition of one methyl group to a molecule generally increases the magnitude of properties such as boiling point and hydrophobicity by a discrete amount with a fair degree of regularity. In the realms of therapeutics and toxicology, chemicals having the same pharmacophore or toxicophore usually have similar pharmacological or toxicological profiles.[3] Similarity allows us to group together a set of selected objects, *e.g.*, molecules, that are judged by some measure to be not too different from each other. Often such exercises in similarity are conducted empirically or intuitively, and tend to work well.

There are also a large number of examples that run apparently counter to this simplistic notion of similarity. In comparing the toxicities of benzene and toluene, we see an interesting example of two apparently similar chemicals that have vastly different properties. Whereas benzene is a known carcinogen, toluene is not. The addition of a methyl group makes an extremely important difference in the metabolism and toxic action of these compounds. On the other hand, there is a large body of literature concerning bioisosteric molecules that have no apparent structural similarity but are recognized by cellular receptors as similar.[4] These are "rough spots" in the applications of molecular similarity.

Molecular similarity works well and gives useful results in the middle ground between these two extremes where, on the one hand, intuitively different molecules have very similar properties, and, on the other, apparently similar entities have different properties.

There are also practical problems with the computation and application of molecular similarity. We face the first serious quagmire when we attempt to quantify the notion of "similar" structures, because the concept of "structure" is not uniquely defined in chemistry. In fact, the structure of the same chemical entity can be represented by more than one mutually distinct and non-equivalent object (*vide infra*).

A perusal of the above indicates that the territory of molecular similarity is a messy area, being complicated by such factors as ambiguity in the representation of molecular structure, observation that in some cases small differences in structures may lead to large differences in biological function, and the phenomenon of bioisosterism where cellular targets can accommodate apparently widely different structures into the same site. Another factor in the computation of molecular similarity is that a large number of mathematical functions can be used to derive measures of similarity for a pair of molecules starting from the same set of structural descriptors.[5,6] In defense of the usefulness of similarity, one can say that this is a widely understood intuitive notion that works well to relate molecular structure to its function both qualitatively and quantitatively in many cases.

In this review, we will be concerned with various representations of molecular structure and the extraction of molecular descriptors from these representations. We will show how the use of different descriptors belonging to the same class of structural representation may lead to the selection of mutually different sets of analogs. We will also attempt to illuminate the basic nature of the similarity space. We will show how selected neighbors have been used in the

prediction of properties using similarity spaces derived from descriptors that maximally characterize the variance within the set of molecular structures. Finally, we will discuss our recently formulated idea of "tailored similarity" where the descriptor space for the computation of intermolecular similarity is property-driven instead of being intuitive or arbitrary.

2. The Molecular Structure Conundrum

Molecular descriptors are used for the characterization of structure and for the computation of intermolecular similarity. These descriptors are derived from different types of structural models for molecules.

The lack of a uniform approach in the characterization of structure is attested to by the multiplicity of methods by which chemical structure is modeled. At the most fundamental level, the structure of any model of an assembled entity (*e.g.*, a molecule) may be defined as the pattern of relationships among its constituent parts as distinct from the values associated with them.[7] In chemistry, one has to deal with structures of different levels of sophistication. In particular, there are a large number of models that seek to explain properties of molecules in terms of characteristics of their structure.[8] This is partly because we need models to predict diverse properties of molecules that might not originate from analogous molecular or sub-molecular phenomena. The term "molecule" means different things when it represents an assembly of identifiable atoms held together by fairly rigid bonds as compared to a collection of delocalized nuclei and electrons in which all identical particles are indistinguishable.[9] Consequently, the term molecular structure represents a group of non-equivalent conceptual entities. There is no reason to believe that when we discuss different topics (*e.g.*, organic synthesis, reaction rate theories, spectroscopic transitions, reaction mechanisms, or *ab initio* calculations) using the concept of molecular structure, the different meanings we attach to the term molecular structure ultimately flow from a single concept.[9,10] This fundamental problem has been stated by Woolley[10] as follows:

> ...there is no reason to suppose that the same basic idea can provide a basis for the discussion of *all* molecular experiments. This is understandable if one recognizes that every *physical and chemical concept is only defined with respect to a certain class of experiments*, so that it is perfectly reasonable for different sets of concepts, although mutually incompatible, to be applicable to different experiments.

The different methods of relating molecular structure to the properties of molecules differ not only in terms of choice of the critical set of entities used to represent molecular structure, but also in terms of techniques used to characterize molecular structure.[8,11–14]

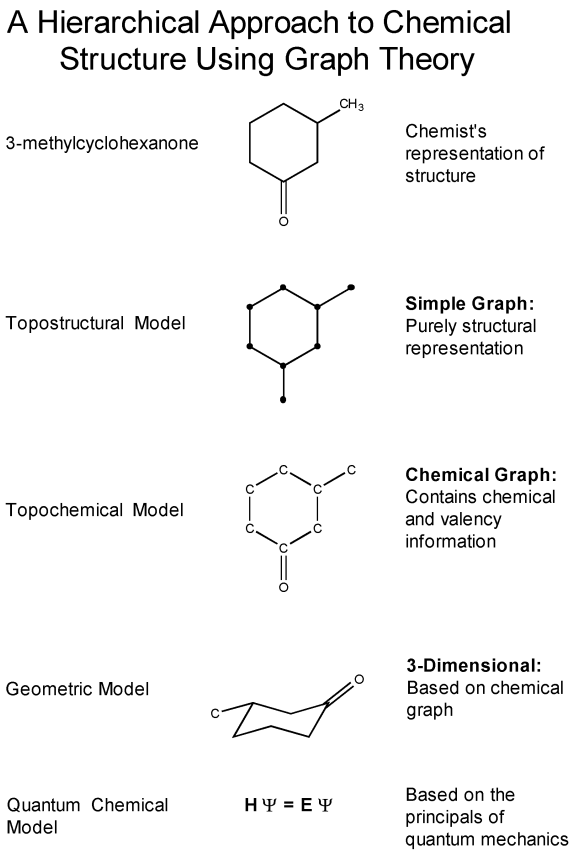


Figure 1. A hierarchical approach to the representation of chemical structure.

In the realm of structural chemistry, our group has recognized a hierarchy of four major representations; *viz.*, topostructural, topochemical, geometrical, and quantum chemical; of increasing complexity and difficulty of computation (Figure1).

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*[15] which can be empirically tested. For example, when it was suggested by Sylvester[16] 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 objects), 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}.[17] 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 *R* of real numbers and thereby orders the set empirically. Therefore, to predict the property from structure, it is

convenient if the structure can be mapped onto the set of real numbers, leading to molecular descriptors.[18,19] This is a more specific theoretical model based on the same model object (chemical graph) and can be accomplished by using specific graph invariants.

The predictive potential of a theoretical model depends primarily on: (1) the efficacy of the representation of the critical aspects of the structure related to the property of interest, and (2) the optimal use of the molecular descriptors.

This review examines the development and use of quantitative molecular similarity analysis (QMSA) methods that make use of easily calculated molecular descriptors derived from chemical graph theory, *viz.*, topostructural indices (TSI), topochemical indices (TCI), and a particular class of substructures known as atoms pairs (APs). Such descriptors can be calculated quickly and, therefore, can be applied to the analysis of medium-sized and large databases. Experimental physicochemical properties and quantum chemical parameters have also been used for similarity analysis,[5,20–23] though we will not review them in this chapter.

3. Calculation of Molecular Descriptors

A 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.[24] A structural model assigns to the points of G a realization in some applied field and each element of E indicates a pair of points which are in the finite nonempty irreflexive symmetric binary relation described by G. In a molecular graph (the conventional chemical structure), the set of atoms comprises the point set V and covalent chemicals bonds are elements of E. Such a graph retains the full topology of the molecule and represents molecular structure, where the word "structure" is used to denote a formal system of relations of certain logical types without emphasizing the entities which they relate. It is because of this general nature that graph-theoretic methods have been used for characterizing structure in such diverse areas as theoretical physics, chemistry, biological and social sciences, engineering, computer science and linguistics. Interestingly, it has been postulated that the logical homology of organization in apparently dissimilar systems is the reason why we find isomorphic laws in different fields of science.

In chemistry, two types of graphs, *viz.*, hydrogen-suppressed and hydrogen-filled graphs, are often used to model molecular structure. While in the former only the non-hydrogen atoms are represented by points, in the latter all atoms (including hydrogen atoms if present in the molecular formula) are represented by vertices. G_1 and G_2 are the hydrogen-suppressed and hydrogen-filled graphs, respectively, of ethyl acetate (*I*) (Figure 2).

Such a graph represents the "topology of a molecule" in the sense that it depicts the pattern of connectedness of atoms in the molecule, being, at the same time, independent of such metric aspects of molecular structure as equilibrium distance between nuclei, valence angles, etc.

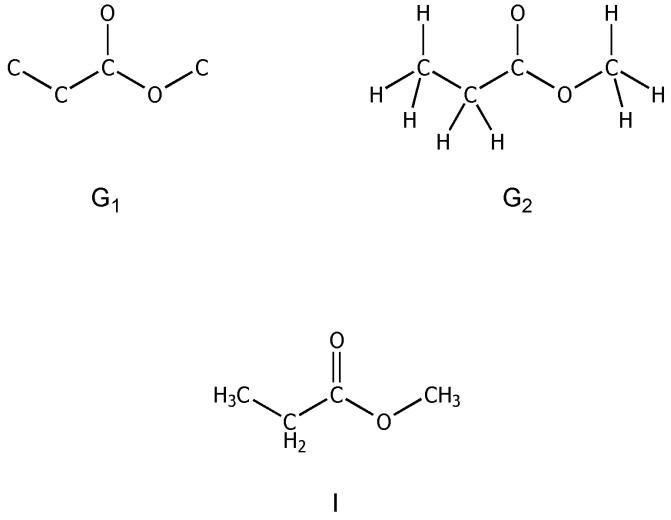
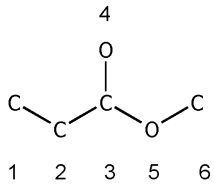


Figure 2. Ethyl acetate represented by its structural formula (*I*), the labeled hydrogen suppressed graph (G_1), and the labeled hydrogen-filled graph (G_2).

Molecular graphs can be represented by different types of matrices, *e.g.*, the adjacency matrix, distance matrix, and incidence matrix. Invariants derived from such matrices are used as numerical molecular descriptors.

The distance matrix $D(G)$ of a nondirected graph G with n vertices is a symmetric n x n matrix (d_{ij}), where d_{ij} is equal to the distance between vertices v_i and v_j in G. Each diagonal element d_{ij} of $D(G)$ is equal to zero. Since topological distance in a graph is not related to the weight attached to each edge (bond), $D(G)$ does not represent valence bond structures of molecules containing more than one covalent bond between adjacent atoms.

The adjacency matrix $A(G_2)$ and the distance matrix $D(G_2)$ for the ethyl acetate is illustrated in Figure 3.



G_1

$A(G_1)$ =

	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	0
3	0	1	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	0	1
6	0	0	0	0	1	0

$D(G_1)$ =

	1	2	3	4	5	6
1	0	1	2	3	3	4
2	1	0	1	2	2	3
3	2	1	0	1	1	2
4	3	2	1	0	2	3
5	3	2	1	2	0	1
6	4	3	2	3	1	0

Figure 3. The adjacency matrix (*A*) and distance matrix (*D*) for the hydrogen-suppressed graph (G_1) of ethyl acetate.

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

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

Vertex degree is integral to the calculation of the connectivity indices as formulated by Randic and Kier and Hall.[25]

The zero order connectivity index, ${}^0\chi$, is defined as:[25]

$${}^0\chi = \sum_i (\delta_i)^{-½} \tag{2}$$

and Randic's connectivity index, ${}^1\chi$, is defined as:[25]

$${}^1\chi = \sum_{all\ edges} (\delta_i \delta_j)^{-½} \tag{3}$$

Based on these two indices, Kier, Murray, Randic, and Hall developed a generalized connectivity index ${}^h\chi$ considering paths of type $v_0, v_1, \ldots v_h$ of length h in the molecular graph:[26]

$${}^h\chi = \sum (\delta_{v0}, \delta_{v1}, \ldots, \delta_{vh})^{-½} \tag{4}$$

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

Kier and Hall extended this method to include cluster (${}^h\chi_C$), path-cluster (${}^h\chi_{PC}$), and cyclic (${}^h\chi_{Ch}$) types of simple connectivity indices which encode information about branching patterns, paths extending from branch points, and the presence of ring fragments, respectively.[25] In order to encode information about bond order, bonding connectivity indices (${}^h\chi^b$) were developed using the tree method of Basak *et al*.[8]

A further extension of Kier and Hall's work introduced the valence connectivity indices that are based on vertex-weighted graphs where the weight, δ_i^v, of the i^{th} vertex is calculated as follows:[25]

$$\delta_i^v = Z_i^v - h_i \tag{5}$$

where Z_i^v is the number of valence electrons of the atom represented by the i^{th} vertex of the chemical graph and h_i is the number of hydrogen atoms attached to it. Valence connectivity indices, ${}^h\chi^v$, are calculated by replacing δ_i in equation 4 with δ_i^v. It is to be noted, however, that in the case of certain atoms; *e.g.*, chlorine, bromine, iodine, fluorine and sulfur; the δ^v values used are derived empirically through calibration with physicochemical properties.[25] However, the physical and/or graph-theoretic basis for these empirical adjustments remains far from clear.

The P_h (h = 0–10) parameters used in this study represent the number of occurrences of paths of length h in the hydrogen-suppressed molecular graph G. P_0 (paths of length = 0) is the number of vertices and P_1 (paths of length = 1) is the number of edges of G. Higher-order P_h terms can be calculated using graph-theoretic algorithms.

The Wiener index, W,[27] is the first topological index reported in the chemical literature. It is calculated from the distance matrix $D(G)$ of a hydrogen-suppressed graph G as the sum of entries in the upper triangular distance submatrix:[28]

$$W = \tfrac{1}{2}\sum_{ij} d_{ij} = \sum_h h \cdot g_h \tag{6}$$

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

Information-theoretic topological indices are calculated by the application of information theory to chemical graphs. An appropriate set A of n elements is derived from a molecular graph G depending upon certain structural characteristics. On the basis of an equivalence relation defined on A, the $\sum_i n_i = n$ set A is partitioned into disjoint subsets A_i of order n_i (i = 1, 2, …, h;). A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, ..., A_h$$

$$p_1, p_2, ..., p_h$$

where $p_i = n_i/n$ is the probability that a randomly selected element of A will occur in the i^{th} subset.

The mean information content of an element A is defined by Shannon's relation:[29]

$$IC = -\sum_{i=1}^{h} p_i \log_2 p_i \tag{7}$$

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

It is to be noted that information content of a graph G is not uniquely defined. It depends on the way the set A is derived from G as well as on the equivalence relation which partitions A into disjoint subsets A_i. For example, when A constitutes 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 orbits of

the automorphism group of G whereafter vertices belonging to the same orbit are considered equivalent.

Rashevsky[30] was the first to calculate information content of graphs where "topologically equivalent" vertices are placed in the same equivalence class. In Rashevsky's approach, two vertices u and v of a graph are said to be topologically equivalent if and only if for each neighboring vertex u_i (i = 1, 2, …, k) of the vertex u there is a distinct neighboring vertex v_i of the same degree for the vertex v. Subsequently, Trucco[31,32] defined topological information of graphs on the basis of graph orbits. In this method, vertices which belong to the same orbit of the automorphism group are considered topologically equivalent. While Rashevsky used linear graphs with indistinguishable vertices to symbolize molecular structure, weighted graphs or multigraphs are better models for conjugated or aromatic molecules because they more properly reflect the actual bonding patterns, *i.e.*, electron distribution.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar, Roy and Sarkar[33] calculated information content of chemical graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms or reaction centers are often modulated by physicochemical characteristics of distant neighbors, *i.e.*, neighbors of neighbors, it was deemed essential to extend this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. If r is any non-negative real number and v is a vertex of the graph G, then the open sphere $S(v, r)$ is defined as the set consisting of all vertices v_j in G such that $\partial(v, v_j) < r$. Obviously, $S(v, 0) = \emptyset$, $S(v, r) = v$ for $0 < r < 1$, and $S(v, r)$ is the set consisting of v and all vertices v_j of G situated at unit distance from v, if $1 < r < 2$.

One can construct such open spheres for higher integral values of r. For a particular value of r, the collection of all such open spheres $S(v, r)$, where v runs over the whole vertex set $V(G)$, forming a neighborhood system of the vertices of G. A suitably defined equivalence relation can then partition $V(G)$ into disjoint subsets consisting of topological neighborhoods of vertices up to r^{th} order neighbors. Such an approach has already been initiated and the information-theoretic indices calculated are called indices of neighborhood symmetry.[34]

We have symbolized chemical species by weighted linear graphs. Two vertices u_0 and v_0 of a molecular graph are said to be equivalent with respect to the r^{th} order neighborhood if and only if corresponding to each path $u_0, u_1, \ldots, u_r$ of length r there is a distinct path $v_0, v_1, \ldots, v_r$ of the same length such that the paths have similar edge weights, and both u_0 and v_0 are connected to the same number and type of atoms up to the r^{th} order bonded neighbors. The detailed equivalence relation is described in our earlier studies.[34,35]

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by equation 7. It is clear that the vertices of a graph belonging to the same equivalence class in terms of the above relation may be permuted without disturbing the relation already defined on the vertex set. Therefore, as pointed out by Mowshowitz,[36–39] measures of

molecular complexity give information content of structures in relation to a system of transformations leaving the structure invariant.

Basak, Roy and Ghosh[40] 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 from equation 7 and n is the total number of vertices of the graph.

Another information-theoretic invariant, complementary information content (CIC_r) is defined as:[41]

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

CIC_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r.

The information-theoretic index on graph distance, $I_D{}^W$, is calculated from the distance matrix $D(G)$ of a chemical graph G as follows:[42]

$$I_D^W = W \log_2 W - \sum g_h \cdot h \log_2 h \tag{10}$$

The mean information index, $\bar{I}_D{}^W$, is found by dividing the information index, $I_D{}^W$, by W.

Because the distance matrix of a graph does not account for the chemical nature of edge multiplicity of vertices, indices of neighborhood symmetry are capable of characterizing chemical structure more efficiently as compared to $I_D{}^W$ or $\bar{I}_D{}^W$.

Table 1 provides a list of calculated descriptors, along with brief descriptions and hierarchical classification. It is important to note, however, that different sets of descriptors were used in our various similarity studies.

4. Similarity Relation: The Tolerance Relation

Similarity or resemblance among elements of a set of objects is best analyzed in terms of the tolerance relation. For example, if we take the set of five-letter English words and call any two words similar if they differ at most by one letter, one can have the following sequence of *similar words*:

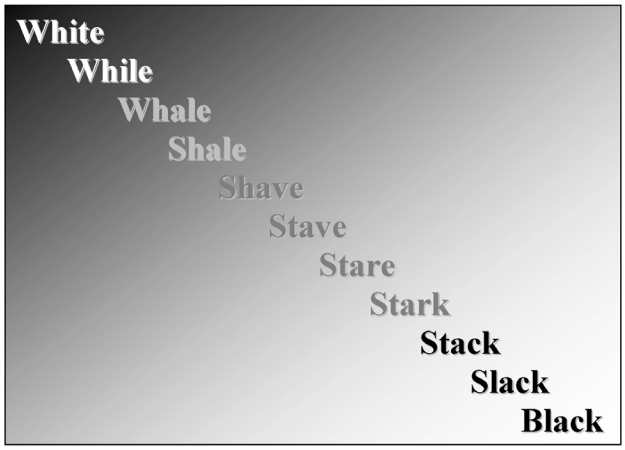


Figure 4. A sequence of similar words, leading from white to black.

Whereas any two consecutive words in the sequence are similar, it is obvious that distinct "similar" neighbors of any word are mutually dissimilar. This arises out of the fact that the tolerance relation is not transitive, although it is reflexive and symmetric:

Definition: The relation *A* on a set *M* is called *tolerance* (or *tolerance relation*) if it is reflexive and symmetric.[43]

In measuring structural similarity/dissimilarity of chemicals, we first describe the set of chemicals (objects) using a prescribed structure space, which consists of molecular descriptors characterizing different aspects of molecular structure. A chosen distance function or some association coefficient then quantifies the degree of similarity/dissimilarity of a pair of chemicals based on its magnitude. As in the case of language given above, one can start with one chemical, find its near neighbor, and find, in turn, an analog of the near neighbor based on some selected criteria. If one continues the process, it is not difficult to see that at some stage the chosen chemical will be structurally quite dissimilar to the one we started with. Then the selected analog would serve no purpose in characterizing the neighborhood of the molecule that was our query chemical. It is also difficult to say precisely at what point in the analog selection process that will happen.

Table 1. Theoretical molecular descriptors and brief definitions

	Topostructural (TS)
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance *h*
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
$^h\chi$	Path connectivity index of order *h* = 0–10
$^h\chi_C$	Cluster connectivity index of order *h* = 3–6
$^h\chi_{PC}$	Path-cluster connectivity index of order *h* = 4–6
$^h\chi_{Ch}$	Chain connectivity index of order *h* = 3–10
P_h	Number of paths of length *h* = 0–10
J	Balaban's J index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order (# of non-H atoms), and distance sum; operation *y* = 1–5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1–5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation *y* = 1–5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation *y* = 1–5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation *y* = 1–5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation *y* = 1–5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation *y* = 1–5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation *y* = 1–5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation *y* = 1–5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation *y* = 1–5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation *y* = 1–5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation *y* = 1–5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation *y* = 1–5
	Topochemical (TC)
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph

Table 1. Continued—Topochemical (TC)

I_{orb}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi^b$	Bond path connectivity index of order h = 0–6
${}^h\chi^b_C$	Bond cluster connectivity index of order h = 3–6
${}^h\chi^b_{Ch}$	Bond chain connectivity index of order h = 3–6
${}^h\chi^b_{PC}$	Bond path-cluster connectivity index of order h = 4–6
${}^h\chi^v$	Valence path connectivity index of order h = 0–10
${}^h\chi^v_C$	Valence cluster connectivity index of order h = 3–6
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order h = 3–10
${}^h\chi^v_{PC}$	Valence path-cluster connectivity index of order h = 4–6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
k_0	Kappa zero
k_h	Kappa simple indices (h = 1–3)
ka_1-ka_3	Kappa alpha indices
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1–5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1–5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1–5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1–5
ANZ_y	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1–5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1–5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; operation y = 1–5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
${}^h\chi^v$	Valence path connectivity index of order h = 7–10
${}^h\chi^v_{Ch}$	Valence chain connectivity index of order h = 7–10
si	Shannon information index
totop	Total Topological Index t
sumI	Sum of the intrinsic state values I
sumdelI	Sum of ΔI values

Table 1. Continued—Topochemical (TC)

tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (k_1* k_2/nvx)
$\overline{I_D^C}$	Bonchev-Trinajstić information index
I_D^C	Bonchev-Trinajstić information index
Wp	Wiener p
Pf	Platt f
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
numHBd	Number of hydrogen bond donors
numwHBd	Number of weak hydrogen bond donors
numHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =CH–
SHtvin	E-State of C atoms in the terminal vinyl group, $=CH_2$
SHavin	E-State of C atoms in the vinyl group, =CH–, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH, –NH2, –NH–, –SH, and #CH
SHwHBd	Weak hydrogen bond donor index, sum of C–H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH, –NH2, –NH–, >N–, –O–, –S–, along with –F and –Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (y = 2–10)
$SHBint_y$	E-State descriptors of potential internal hydrogen bond strength (y =2–10)
	Electrotopological State index values for atoms types: SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss,Bem, SssBH ,SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC,StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SssssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SsssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SssssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb

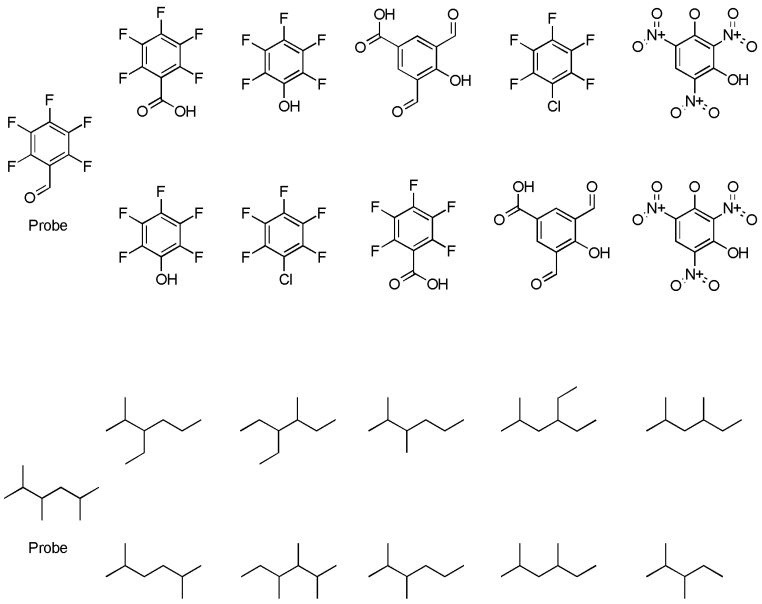


Figure 5. Five nearest neighbors (analogs), from left to right, for the probe chemicals pentafluoro-methoxybenzene and 2,3,5-trimethylhexane selected using Euclidean distance and atom pair based similarity spaces. The row above the probe shows analogs selected using the Euclidean distance method, while the row below the probe shows analogs selected using the atom pair method.

Examination of the analogs of pentafluoro-methoxybenzene and 2,3,5-trimethylhexane presented in Figure 5 attests to the idea that if one looks at the structure of the probe chemical and its selected neighbors, one can see a certain amount of intuitive similarity between the chemical structures. It is reassuring to see that our intuitive notions of structural similarity are upheld by a quantitative method for determining structural similarity.

5. Quantitative Molecular Similarity Analysis (QMSA)

Our group has been interested in quantifying the molecular similarity/dissimilarity of chemicals using topological descriptors. The impetus for this research comes principally from two directions. First, molecular similarity methods provide a way of combining many structural features in the ordering of a set of molecules. Second, quantitative molecular similarity methods can be used in the quantitative selection of molecular analogs. Molecular analogs are often used to estimate the therapeutic/toxic potential of a chemical of interest for the evaluation of chemicals for drug discovery and hazard assessment. Often, the features used for such comparisons are intuitively selected. We have been interested in developing similarity methods that are objective, fast and generally applicable to any chemical species, real or hypothetical.

Similarity methods based on topological indices and substructures fall into this category since these parameters can be calculated for any arbitrary molecular structure. Similarity methods based on experimental data will have limited applicability because in most real world situations even the simplest experimental data, *e.g.*, boiling point, melting point, and vapor pressure, are not available for the majority of chemicals.[44]

In particular we have used Euclidean distance in a principal component (PC) space derived from a large number of indices to quantify intermolecular similarity. Another approach, based on a particular type of molecular substructure, called atom pairs (APs), uses a Tanimoto-type association coefficient to measure the similarity/dissimilarity of molecules. These techniques have been used in the following ways: 1) to define a variety of structure spaces to quantify molecular similarity, 2) in the selection of analogs, 3) to carry out comparative studies of spaces derived from measured physicochemical properties *vis-à-vis* topological descriptors, 4) to select a number of neighbors of a chemical in various structure spaces to estimate properties of the target chemical, and 5) to estimate toxic modes of action (MOA) of chemicals from the MOA of neighboring chemicals in different structure spaces. More recently we have developed a new approach to the formulation of similarity spaces. This new approach, the tailored similarity method, creates similarity spaces optimized for the molecular or biological property of interest. The following is a summary of the research already carried out by our group in these areas.

5.1 Databases

We have used a wide variety of data sets and databases over the past fourteen years. Here we summarize the data that has been analyzed using a variety of similarity measures. Later, in discussing specific studies, we will refer to these databases by number.

5.1.1 TSCA data from ASTER

This data comes from the ASTER (Assessment Tools for the Evaluation of Risk) database[45] developed by the USEPA from data available in the Toxic Substances Control Act (TSCA) Inventory.

5.1.1.1 Normal boiling point (bp)

Experimental data for the normal boiling points (bp) of 2,926 chemicals were extracted from the ASTER system. The full set of 2,926 chemicals was used in the QMSA-based estimation of normal boiling point.

5.1.1.2 Normal vapor pressure ($\log_{10}P_{vap}$)

Experimental data for normal vapor pressure (P_{vap}) was used in modeling 469 chemicals. The subset used in this study (469 chemicals) represents a chemically diverse set with experimental values for P_{vap} ranging from 3 to 10,000 mmHg.

5.1.2 CFC normal boiling point data

Another set of normal boiling point data was collected from four standard reference texts; *Beilstein's Handbuch der Organischen Chemie*, the *CRC Handbook of Chemistry and Physics*, Heilbron's *Dictionary of Organic Compounds*, and Smith and Srivastava's *Thermodynamic Data for Pure Compounds, Part B*; for use in several studies by Balaban *et al*.[46,47] This set consisted of boiling point data for 276 chlorofluorocarbons (CFCs) with carbon skeletons containing one to four atoms. Nine of the compounds included in the studies by Balaban *et al.* were removed as outliers, leaving a total of 267 CFCs. The compounds were judged to be outliers since their boiling points were greater than two standard deviations from the mean boiling point for the entire set.

5.1.3 Hydrocarbon normal boiling point data

The hydrocarbon data set used in several of our similarity studies includes 73 alkanes taken from Needham *et al*.,[48] 29 alkylbenzenes taken from the collection of Mekenyan *et al*.,[49] and 37 polycyclic aromatic hydrocarbons taken from the collection of Karcher.[50] Normal boiling point data was taken from the referenced literature for these 139 chemicals.

5.1.4 Octanol/water partition coefficient (logK_{OW})

Experimental data for 213 chemicals was used to model octanol/water partition coefficient (logK_{OW}). This data represents a small subset of the data comprising the STARLIST data set used as the basis for the CLOGP program.[51] This subset represents the group of chemicals that have no known duplicates within STARLIST, no known errors, experimental logK_{OW} values in the range between –2 and 5.5, and zero explicit hydrogen bonding centers as calculated by the program HB_1.[52]

This subset was chosen to examine the effectiveness of models based on topological indices in the prediction of octanol/water partition coefficient for compounds which do not have explicit hydrogen-bonding centers and that fall within a well-defined range for experimental logK_{OW} values. By well-defined we refer to the findings of Brooke *et al*.[53] and De Brujin *et al*.[54] demonstrating that the experimental determination of logK_{OW} values greater than 5.5 is problematic.

5.1.5 Industrial pollutants

One interesting data set that has been analyzed in a number of different studies is a set of 76 industrial chemicals. This set, first analyzed by Basak and Grunwald,[55] represents the intersection of a subset of the TSCA inventory (~10,000 chemicals), melting point, boiling point, and logP data available from the ASTER system,[45] and a database of solvatochromic parameters consisting of α (hydrogen bond donor acidity), β (hydrogen bond acceptor basicity), V/100 (molar volume), and π (polarizability). The solvatochromic parameters were provided by Kamlet.[56]

5.1.6 JP-8 mixture database

This data set represents a subset of the identified constituents of jet propellant formulation number eight (JP-8),[57] a set of 166 hydrocarbons. At the time of this study 228 hydrocarbons had been identified as constitiuents of the jet propellant used by the United States Armed Forces, not including a wide variety of additives. This subset represents the graph distinct molecules for which boiling point, vapor pressure, heat of vaporization, water solubility, adsorption coefficient, and logP were all available from the ASTER system.[45] However, even for the reduced set of 166 compounds, most of the data values available from ASTER were calculated, not experimental values.

5.1.7 Inhibition of complement

The benzamidines used in this study consist of a group of 107 compounds taken from the collection of Hansch and Yoshimoto.[58] The data was compiled from the original studies by Baker *et al*.[59–62] in which Baker and his students experimentally determined the inhibition of guinea pig complement by benzamidines. The data supplied by Hansch and Yoshimoto consisted of the measured log(1/C) values, where C is the micromolar concentration for 50 percent inhibition of the complement system (IC_{50}).

5.1.8 Inhibition of microsomal *p*-hydroxylation of aniline

Experimental data for the inhibition of the microsomal *p*-hydroxylation of aniline (pIC_{50}) for a set of nineteen alcohols was taken from the work of Cohen and Mannering.[63]

5.1.9 Acute aquatic toxicity

Data on acute aquatic toxicity $-\log(LC_{50})$ in fathead minnow (*Pimephales promelas*) was taken from the work of Hall, Kier and Phipps.[64] Their data was compiled from eight other sources, as well as some original work which was conducted at the U.S. Environmental Protection Agency (USEPA) Environmental Research Laboratory in Duluth, Minnesota. The set includes data for 69 benzene derivatives. According to the authors, the set of benzene derivatives was tested using methodologies which were comparable to their 96-hour fathead minnow toxicity test system. The derivatives chosen for this study have seven different substituent groups that are all present in at least six of the molecules. These groups consist of chloro, bromo, nitro, methyl, methoxyl, hydroxyl, and amino substituents.

5.1.10 CRC identified carcinogens/noncarcinogens

A set of 520 compounds was taken from the *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database*.[65] This set consists of 260 mutagens and 260 non-mutagens based on qualitative assessments of the Ames' mutagenicity assay. Please see the original manuscript for details.[66]

5.1.11 Aromatic amine mutagenicity

A set of 127 aromatic and heteroaromatic amines was previously collected from the literature by Debnath *et al*.[67]

5.1.11.1 Aromatic amine mutagenicity (TA98 + S9)

Benigni later reanalyzed the work of Debnath *et al.*[67] and assigned positive or negative values for the compounds based on their mutagenicity results from the *S. typhimurium* TA98 + S9 microsomal preparation.[68] This discriminant data was used in modeling mutagenicity.

5.1.11.2 Aromatic amine mutagenicity subset (TA98 + S9)

From the previously mentioned set of 127 aromatic and heteroaromatic amines, a subset of 95 compounds was taken. These 95 compounds all have mutagenic potency values greater than zero (Rev/nmol > 0) as measured in the *S. typhimurium* TA 98 + S9 microsomal preparation.[67] In this study, the quantitative mutagenic potency was modeled.

5.1.11.3 Aromatic amine mutagenicity (TA100)

Data for this set of compounds tested in *S. typhimurium* TA100 were also available from the work of Debnath *et al*.[67]

5.1.12 Nitrosamine mutagenicity

Mutagenicity data for a set of 15 nitrosamine compounds was taken from the work of McCann *et al*.[69] The data available for these 15 compounds was recorded as mutagenic potency (Rev/nmol) measured in the Ames assay.

5.1.13 Diverse mutagens and carcinogens

This data set is a set of 277 chemicals presented by Yamaguchi *et al*.[70]

5.1.13.1 Yamaguchi diverse mutagens

This subset of Yamaguchi's data is comprised of all of the 277 chemicals that had reported results for mutagenicity in the Ames test, mutagenicity in the medium-term liver carcinogenesis bioassay, and carcinogenicity in the two-year rodent bioassay in rat and/or mouse. This subsetting resulted in a set of 113 chemicals, 68 of which are classified as non-mutagens and 45 of which are classified as mutagens in the Ames test.

5.1.13.2 Yamaguchi diverse carcinogens

The Yamaguchi set (mentioned above) has also been used to model carcinogenicity in the two-year rodent bioassay in rat and/or mouse. The set used for this study is the same 113 chemical subset as was used in modeling mutagenicity. However, while the split for mutagenicity was 68 non-mutagens to 45 mutagens, 34 of the chemicals were classified as non-carcinogens and 79 were classified as carcinogens in the two-year rodent bioassay.

5.1.14 Toxic mode of action

This database is composed of 283 chemicals selected from the 617-chemical Mid-Continent Ecology Division-Duluth fathead minnow database. The entire database has been evaluated for toxic mode of action. The compounds selected represent eight modes of action for which higher confidence was associated with the final mode classification. See the original study by Basak *et al*.[71] and the report on the full database by Russom *et al*.[45] for more specific details on the database.

6. Arbitrary Similarity Methods

The bulk of our previous work in the field of molecular similarity falls under the heading of arbitrary similarity methods. These methods use either molecular substructures or fragments, as in the atom pair (AP) method, or topological indices, as in the Euclidean distance method, to characterize the overall population of molecules within the data set. In this way, the set itself is what we are trying to best characterize with no interest at all in how the property of interest relates to the structural data that best characterizes the diversity within the set of molecules.

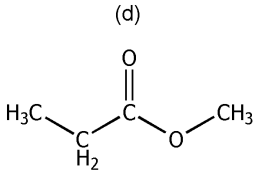
6.1 Chemical descriptors for QMSA

6.1.1 Atom pairs

The atom pair approach examines the structure of each chemical and compares chemicals based on the presence of identical substructures within each molecule. This approach has proven useful in the creation of structure spaces.[66,71–83] An atom pair represents any two atoms in the molecule and includes information about their path-wise interatomic separation and also encodes the presence or absence of π-orbitals. The method of Carhart *et al*.[84] was employed in the calculation of atom pairs (APs). This method defines an atom pair as a substructure consisting of two non-hydrogen atoms *i* and *j* and their interatomic separation:

<atom descriptor i> – <separation> – <atom descriptor j>

where <atom descriptor> contains information regarding atom type, number of non-hydrogen neighbors and the number of π electrons. The interatomic separation is defined as the number of atoms traversed in the shortest bond-by-bond path containing both atoms. *APProbe*[85] was used to calculate the atom pairs for each molecule and generate the intermolecular similarity scores for both sets of compounds.



Ethyl acetate

	Atom Pair	Path	Frequency of Occurrence
1	CX_1-2-CX_2	*ab*	1
2	CX_2-2-C.X_3	*bc*	1
3	C.X_3-2-O.X_1	*cd*	1
4	C.X_3-2-OX_2	*ce*	1
5	OX_2-2-CX_1	*ef*	1
6	CX_1-3-C.X_3	*abc, cef*	2
7	CX_2-3-O.X_1	*bcd*	1
8	CX_2-3-OX_2	*bce*	1
9	CX_1-4-O.X_1	*abcd, dcef*	2
10	CX_1-4-OX_2	*abce*	1
11	CX_2-4-CX_1	*bcef*	1
12	CX_1-5-CX_1	*abcef*	1

Figure 6. Calculation of atom pairs for ethyl acetate.

Figure 6 demonstrates the calculation of atom pairs for ethyl acetate. Ethyl acetate has fourteen total atom pairs, twelve of which are unique. In figure 6, X_n represents the number of non-hydrogen neighbors, "." represents the one π-electron and C and O are the atomic symbols for the atoms in each pairing. The "-*k*-", where k = 1–5, are the atomic separation values including the starting atom and the ending atom in each pairing.

As was mentioned earlier, the atom pair method utilizes a Tanimoto-type coefficient for measuring molecular similarity. This associative measure described by Carhart *et al*.[84] is based on atom pair descriptors. The measurement is the ratio of the number of shared atom pairs between two molecules over the total number of atom pairs present in the two molecules. Similarity (*S*) between molecules *i* and *j* is defined as:

$$S_{ij} = 2C/(T_i + T_j) \quad (11)$$

where C is the number of atom pairs common to molecule *i* and *j*. T_i and T_j are the total number of atom pairs in molecules *i* and *j*, respectively. The numerator is multiplied by a factor of 2 to reflect the presence of shared atom pairs in both compounds. These similarity scores were also calculated using *APProbe*.[85]

6.1.2 Topological indices

In constructing structure spaces from topological indices there are many different potential approaches, especially given the large number of topological indices currently available. In our studies we have used five major approaches to the development of molecular structure spaces. The first, and most obvious approach, is to use individual TIs as the dimensions within the structure space. However, as noted earlier, with the large number of indices currently available, this quickly becomes a problem. We do not really want to use 50- or 100-dimensional structure spaces. The fewer indices used in creating the space, the better understanding we will have of the aspects of structure being represented in the space. Also, our studies have shown that many of the TIs are highly intercorrelated and are therefore not as useful as unique dimensions within a structure space. To solve this problem we have used variable clustering[86] to form disjoint clusters of TIs and have extracted individual TIs from each of the clusters. This approach has been used on both large and small data sets[76,80,81] and can be further augmented by rescaling the indices so that each index has variance equal to one and a mean of zero.[76,78,80]

Another approach to reducing the dimensionality of the TI problem and removing the problem of index intercorrelation is to create orthogonal descriptors based on linear combinations of the original indices. This has been accomplished using two statistical approaches. First, the disjoint clusters resulting from the variable clustering procedure can be combined into new, essentially orthogonal variable clusters (VCs).[76,78,80] Secondly, a principal components analysis (PCA)[86] can be used to create linear combinations of indices which form truly orthogonal principal components (PCs).[8,66,71–83] Since a number of new PCs equal to the original number of descriptors is generated, it is necessary to have a standardized approach to select a reduced subset of PCs. The standard approach to this problem is to select the PCs with eigenvalues greater than or equal to one. This generally retains roughly the first ten percent of the PCs. Either approach results in greatly reducing the dimensionality of the structure space and avoids the problems caused by highly intercorrelated indices. Both methods have been used in a number of our similarity studies.

Whether we use TIs, PCs or VCs, we need some means of calculating the intermolecular similarity between the chemicals within the data sets. For this approach we have used a standard measure of Euclidean distance (*ED*) within an *n*-dimensional space. The *ED* between molecules *i* and *j* is defined as:

$$ED_{ij} = \left[\sum_{k=1}^{n} \left(D_{ik} - D_{jk} \right)^2 \right]^{1/2} \tag{12}$$

where *n* equals the number of dimensions and D_{ik} equals the data value of the k^{th} dimension for compound *i*.

6.1.3 Physicochemical properties *vis-à-vis* topological descriptors

The reasoning and chemical intuition developed by most medicinal chemists and toxicologists has far more to do with familiarity with physicochemical properties and chemical structures than theoretical descriptors. So, it was of interest to see what types of analogs are selected from databases based on physicochemical properties *vis-à-vis* topological descriptors (either selected TIs, PCs derived from TIs, or APs). This could only be done on restricted sets of molecules for which a suite of experimental property data was available. The primary set that we have employed in these studies contains only seventy-six chemicals for which six experimental properties, *viz.*, lipophilicity (logK_{OW}), boiling point, melting point, molar volume (V/100), hydrogen bond donor acidity (α), hydrogen bond acceptor basicity (β), and polarizability (π^*), were available.[55,72] However, recently we have been able to develop a second database of 166 hydrocarbons (constituents of JP-8) for which we were able to obtain calculated values for boiling point, vapor pressure, heat of vaporization, water solubility, adsorption coefficient, and log$P$.[83] These calculated values were all obtained from the USEPA's ASTER (ASsessment Tools for the Evaluation of Risk) system.[45] Unfortunately both of these sets are very limited, but this serves to illustrate just how difficult it can be to find experimental data for chemicals of toxicological interest.

Qualitatively, our results on the selection of analogs using physicochemical and topological spaces show that for any particular query chemical the selected set of neighbors is essentially the same with some minor variation, though the order of neighbor selection differs.[55] So, the physicochemical properties are not expected to give analogs much different from those picked by topological methods. However, when we look at the issue quantitatively we see that the overlap between theoretical spaces and spaces constructed from physicochemical properties is half that shown between similar theoretical models (see Table 2). For instance, when comparing the ten nearest neighbors for the industrial pollutant database (see section 5.1.5), the PC space derived from TIs and the AP space share an average of five neighbors in common. However, the TI and property spaces have an average of 3.5 neighbors in common and the AP and property spaces share an average of 4.1 neighbors in common. For further comparison of the overlap of theoretical and property spaces see the summary statistics in Table 2. The spaces described in the table are a standard atom pair space (AP), a 10-dimensional PC space based on topological indices (TI), and a property space based on the seven physicochemical properties mentioned above (Prop). The largest variation occurs between the TI space and the Prop space, though there are a significant number of unique neighbors selected by all three methods (on average more than 50%). Of course, the distinct advantage of topologically-based methods is that they can be applied to all chemicals, even those that have not yet been synthesized or tested.

Table 2. Average number of identical neighbors selected from three similarity spaces at varying levels of *K*

K	TI vs. AP	AP vs. Prop	TI vs. Prop
5	2.3	1.9	1.6
10	5.0	4.1	3.5
15	8.1	6.3	5.7
20	11.0	8.9	8.6
25	14.3	12.1	11.9
30	17.4	15.7	15.8
35	21.1	20.4	20.0
40	25.5	24.6	25.0

More recently, we have compared the three similarity spaces discussed above in their ability to calculate lipophilicity (log*P*).[83] This was done with both the set of 76 industrial pollutants and the 166 JP-8 constituents (see sections 5.1.5 and 5.1.6). Without question, the similarity spaces derived from physicochemical property data (experimental or calculated) were superior in their ability to estimate log*P*. The results clearly show that estimation based on spaces derived from physicochemical property data is preferable; however, in many cases it is impractical. As such, it is reassuring to know that theoretical structure spaces provide reasonable, if not ideal, estimates.

6.1.4 Selection of mutually different molecular similarity methods

Optimal characterization of molecular structure is prerequisite to the creation of useful similarity spaces and the prediction of the physicochemical properties, biological activities, or toxicities of chemicals for which very little experimental data is available. To this end, we have been exploring the creation of a variety of molecular similarity spaces and comparing them to determine the extent to which they are similar or different.

An early study[66] examined the degree of neighborhood overlap between the PC-based Euclidean distance method and the AP-based similarity method. This study examined the set of 139 hydrocarbons (see section 5.1.3), and determined that the two spaces had an approximately 40% overlap. This research was continued later and in greater depth[72] when we looked at the degree of overlap between four structure spaces and one physicochemical property space using the 76 industrial pollutants (see section 5.1.5). The four structure spaces included principal component spaces derived from topostructural indices, topochemical indices, the combined set of topostructural and topochemical indices, and an AP-based structure space. The property space was a PC space derived from lipophilicity (logK_{OW}), boiling point, melting point, molar volume *V*/100), hydrogen bond donor acidity (α), hydrogen bond acceptor basicity (β), and polarizability (π^*). This study showed the highest degree of overlap between the topochemical space and the space using the combined set of topological indices (Table 2). However, a high degree of overlap existed between all four structure spaces, indicating that there is a high degree of similarity

between the spaces. The physicochemical property space showed the least overlap with the structure spaces.

This study resulted in the discovery that for this particular set of compounds, while the degree of overlap between the groups of analogs selected by theoretical descriptor spaces is relatively high, the similarity space constructed from physicochemical property data provided relatively unique groups of analogs with regards to those selected from the theoretically derived similarity spaces. This shows that if one is attempting to determine the optimal characterization for a similarity space it is best to employ two or three distinct methods, *e.g.*, one theoretical space and one property space, rather than two theoretical spaces that may have a high degree of overlap.

6.2 Selection of analogs using molecular similarity methods

Due to the constant problem of limited availability of toxicity and other relevant data, risk assessment must often be carried out with limited or no experimental data.[87] One approach that is used in this situation is the selection of chemical analogs to estimate the property of the chemical of interest. This use of analogs is based on the tacit assumption that similar chemical structures have similar biological activity profiles. Analog selection is generally carried out by an expert and as a result is a subjective process, but our research has shown that molecular similarity methods can be used to provide a quantitative approach to analog selection.[66,73,75,76,88] Additionally, these methods may quantify aspects of structure that are not perceived by the expert and therefore will result in the selection of analogs that would not have been selected by the expert. Additionally, these methods provide a rapid means of selecting analogs for a large set of chemicals.

In our initial studies, we focused mainly on examining the analogs selected by a single similarity method (Euclidean distance within an *n*-dimensional PC-space) to ensure that "reasonable" analogs were being selected. To this end, we commonly examined the ten nearest neighbors of a selected subset of chemicals to determine how closely related they were to our structure of interest (often referred to as a probe chemical).[8,75] However, it was of interest to see how other methods, AP-based and property-based spaces, performed in analog selection. Also, it was necessary to look at analog selection in a variety of similarity spaces to better understand the degree of similarity/dissimilarity between the spaces. If a space selects different analogs, it is significantly different from another space. See Figures 5 and 7 for examples of analog selection.

6.3 Estimation of properties of chemicals using *k*NN method

The basic assumption of QSAR/QSPR studies is that similar structures usually have similar properties.[19] It follows from this idea that we should be able to estimate physicochemical, biomedicinal, and toxicological properties of chemicals from the known properties of their nearest neighbors in various structure spaces. We have attempted to test this idea using *k*,

ranging between 1 and 25, nearest neighbors of chemicals in structure spaces generated using TIs and APs. Here we summarize some results of our research in this area.

Similarity spaces derived from topological indices and atom pairs have been used in the estimation of a variety of properties: normal boiling point,[76,78,80,88] mutagenic potency (ln*R*),[76,78–80] lipophilicity (log*P*),[78,83] classification of mutagens,[66,74] inhibition of microsomal *p*-hydroxylation of aniline,[82] acute aquatic toxicity,[75] and vapor pressure (*vide infra*). In all cases, we have found that molecular similarity methods can provide reasonable estimates for many properties. However, these methods also have weaknesses. As we mentioned earlier, the similarity relationship is a tolerance relationship; therefore, if we go too far from our probe chemical, the ability to acquire reasonable estimates decreases. Similarly, if the similarity space is not dense enough, we will be using neighbors that are not very similar to our probe and we will have poor estimates. This concept has been shown time and again in our studies. In most cases we have found that using four to eight nearest neighbors is ideal for property estimation. However, in smaller sets and relatively small, diverse sets, a smaller number of neighbors seems to be more desirable.[75,78,82] This makes sense, since this represents the cases illustrated above. In small, congeneric sets of chemicals, it is likely that minor feature changes will have a fairly significant impact on the property of interest and the lack of structural diversity will also impact the quality of the estimates. In larger sets, we are more likely to discover sparse pockets that must rely on distant neighbors for estimates. Ideally, our similarity space would have a uniform distribution both structurally and for the measured values of our property of interest. Unfortunately, we have yet to find such a data set. In the meanwhile, it is certainly best to explore the neighborhoods within our similarity spaces to find the range of *k* that best suits the space and the property under examination.

Another concern with the *k*NN method is that it results in property smoothing, or loss of data variance. As was shown in one of our more recent studies,[89] data variance drops off dramatically with an increase in *k*. Therefore, we need to find similarity spaces that best model the property of interest with the fewest possible neighbors. Again, this data degredation is dependent on the density of the property space. If the property of interest is well distributed within the structure space, then variance will degrade more slowly. However, in a structure space where the property is poorly distributed, data variance will be lost more quickly.

6.4 Estimation of toxic modes of action (MOA) from neighboring chemicals

Following the structure-property similarity principle, one may argue that similar chemicals should have analogous biochemical modes of action (MOA). To test this idea we selected a large and structurally diverse set of 283 chemicals (see section 5.1.14) for which MOA of aquatic toxicity was known with confidence and estimated their MOA from the MOA of their five nearest neighbors.[71] The MOA data was classified as follows: narcosis I (baseline narcosis), narcosis II (polar narcosis), mixed narcosis I/II, oxidative phosphorylation uncoupling, acetylcholinesterase (AchE) inhibition, electrophile/proelectrophile reactivity, and central

nervous system effecting chemicals (neurotoxicants and respiratory blockers). *k*NN estimation, as well as neural network and discriminant analysis techniques, were used to attempt to correctly classify these chemicals. A tiered analysis was conducted in which narcotic and electrophile reactivity classes (narcosis I, narcosis II, mixed narcosis I/II, and electrophile/proelectrophile reactivity) were grouped together in the first tier and then classified separately in the second tier. The results showed that we could correctly predict MOA for about 90% of the chemicals in tier I and 75% of the chemicals in tier II using this method. The results were similar for the other two classification methods.

7. Tailored Similarity

7.1 Background

The QMSA methods mentioned previously are arbitrary similarity methods because they are based solely on structure spaces, the elements of which are selected without considering the property (or activity) of interest. Designers of QMSA methods seem to fall into two major categories. Some have an intuitive notion of the structural attributes that will be useful for the property of interest, while others take a wide variety of structural attributes representing important and diverse aspects of molecular structure. In the latter case, because these arbitrary structural descriptors encompass such a broad range of important structural features, they are expected to be useful in estimating the property of interest through the selection of proper analogs.

To correct for the arbitrary nature of the QMSA methodology, we have recently developed the idea of tailored similarity.[90–93] In this scheme we begin by selecting the elements of the structure space based on a specific property. The structure space thus created is then used for the selection of analogs and estimation of properties of chemicals from their analogs. It has been found that both for physicochemical and toxicological properties (using data sets detailed in sections 5.1.1.2, 5.1.4, 5.1.5, 5.1.6, and 5.1.11.2), tailored similarity spaces outperform arbitrary QMSA models.

7.2 Tailored QMSA methods

Currently the tailored similarity method has only been used with topological indices. The TIs can be used individually or combined into orthogonal descriptors using PCA as described previously. The main difference between tailored and arbitrary similarity comes with the selection of descriptors. While arbitrary methods use techniques that select descriptors that best characterize the variance in the descriptor set, thus optimally characterizing the available structure space, the tailored method uses ridge regression (RR) to choose a subset of descriptors optimal for the property of interest. The ridge regression method is useful in cases where the descriptors are

highly multicollinear and where the number of descriptors is substantially larger than the number of observations.[94] Conceptually, RR can be thought of as recasting the regression as one using the principal components of the predictor variables as new predictors. It differs in that in principal component regression the leading components are retained and used just as in ordinary least squares regression while the trailing components are dropped. RR retains all components, but downweights each of them in accordance with the component's eigenvalue and the "ridging constant" *k*.

The ridge parameter *k* controls the amount of smoothing in ridge regression. If *k* is large, then all regression coefficients are "shrunk" towards zero. Smaller *k* values shrink the directions of small eigenvalues substantially, but the directions with large eigenvalues less so. A suitable value for *k* needs to be found when performing ridge regression. In the current study, the *k* value was chosen to minimize the prediction sum of squares (PRESS), a cross-validation measure. As in jack-knifing, each compound in turn is temporarily omitted from the data set and the RR fitted to the remaining compounds. The resulting model is used to predict the compound that was held back. PRESS is the sum of squares of the differences between the actual values and these holdout predictions. The cross-validated R^2 is defined in terms of PRESS and provides an honest measure of the predictive power of the modeling approach.

Once a set of descriptors is chosen based on the ridge parameter *k*, the Euclidean distance method is used to determine the degree of similarity between all pairs of chemicals within the *n*-dimensional space.

Table 3. Summary of the first twelve principal components derived from a set of 222 topological indices calculated for a set of 166 JP-8 constituents

PC	Eigenvalue	Proportion of Variance Explained	Cumulative Variance Explained	First Most Correlated TI		Second Most Correlated TI	
1	107.93	0.417	0.417	AZN_3	0.99392	P_0	0.9937
2	25.87	0.100	0.517	ASN_2	0.86895	DSN_5	0.86148
3	18.85	0.073	0.589	$\Delta^0\chi$	0.91315	ncirc	0.85846
4	12.55	0.049	0.638	sumDELI	0.74373	IC_1	0.69717
5	8.76	0.034	0.672	ANZ_1	0.62465	Fw	0.56555
6	7.79	0.030	0.702	$^7\chi$	0.51019	$^8\chi$	0.48020
7	6.08	0.024	0.725	dxp8	0.53116	$^8\chi$	0.52996
8	5.61	0.022	0.747	SHHBd	0.49222	numHBd	0.48906
9	5.18	0.020	0.767	SHHBd	0.55899	numHBd	0.54654
10	4.11	0.016	0.783	$^4\chi_{Ch}$	0.77510	$^3\chi_{Ch}$	0.76064
11	3.86	0.015	0.798	$^5\chi^v$	0.42878	$\Delta^7\chi$	0.37443
12	3.55	0.014	0.812	SHwHBd	0.52762	SHCHnX	0.52237

The example of the tailored similarity method presented here uses the TSCA normal vapor pressure database (see section 5.1.1.2). In the construction of the similarity spaces to model normal vapor pressure, our standard method has been altered slightly. Normally we would keep all PCs with eigenvalues greater than or equal to one. However, in this instance the number of PCs meeting this requirement was rather high due to the diversity of descriptors and the diversity of the database. If we kept all PCs with eigenvalues greater than or equal to one, we would have a 32-dimensional similarity space. Since this seems a bit high, we have decided to use the PCs with eigenvalues greater than or equal to three. This leaves a set of twelve PCs, derived from a set of 277 molecular descriptors, that explain approximately 81% of the variance in the descriptor set (see Table 3). Since we have retained the first twelve PCs, we will also retain the twelve highest ranked TIs from the RR procedure (Table 4). In each instance, we have then proceeded to create twelve-dimensional structural similarity spaces using the Euclidean distance method described earlier.

Table 4. Twelve TIs selected by RR for the 467 TSCA vapor pressure set. Indices common to both RR and PCs are indicated in bold.

	TIs from RR (t-value)	
1	SsF	(9.55)
2	SssO	(7.17)
3	NHBint3	(-6.89)
4	**P_0**	(-6.59)
5	SsNH2	(6.59)
6	IC_0	(-6.55)
7	SHHBa	(6.55)
8	$SHBint_3$	(6.41)
9	nelem	(-5.98)
10	DN^2N_3	(-5.96)
11	DN^21_4	(-5.84)
12	DN^2N_4	(-5.84)

7.3 Analog selection by tailored versus arbitrary QMSA methods

Any method that develops a unique similarity space should also result in the selection of a unique set of chemical analogs for any probe (query) chemical. Unless we hope to find unique, non-intuitive analogs, it should also be easy to see some structural resemblance between the probe chemical and its analogs. As a further example of this, we present the analogs of 1,1-dichloroacetylmethoxide as selected by the standard arbitrary similarity method using Euclidean distance and the tailored similarity method using Euclidean distance. As can be seen in Figure 7, there is minimal overlap between the structures selected by the arbitrary and tailored similarity

methods, however, most of the analogs show noticeable similarity to the probe compound (1,1-dichloroacetylmethoxide).

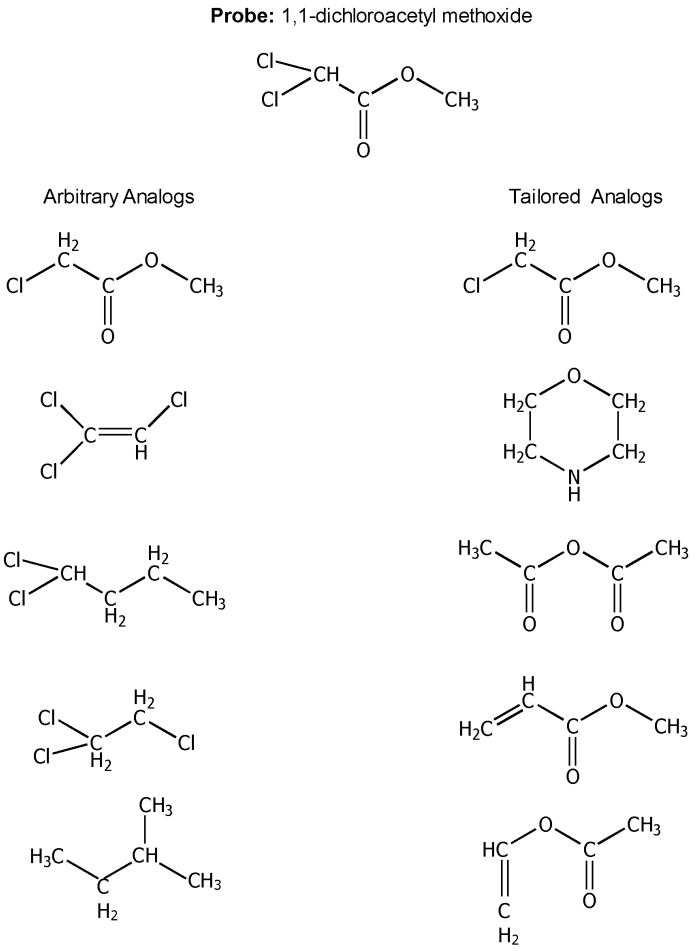


Figure 7. Chemical analogs of 1,1-dichloroacetyl methoxide as selected from the TSCA Vapor Pressure data set by arbitrary and tailored similarity methods. The analogs selected using the arbitrary method are in the left column and those selected using the tailored similarity method are on the right.

7.4 *k*NN property estimation: Tailored versus arbitrary QMSA methods

It is also important that our quantitative molecular similarity method show some degree of utility in property estimation. To that end we have employed the *k*NN method to estimate normal vapor pressure within our arbitrary and tailored similarity spaces. In both cases we have used sets of analogs (or neighbors) with *k* varying from one to ten and at the intervals fifteen, twenty and twenty-five. While the use of fifteen, twenty or twenty-five neighbors is not useful due to the high loss of data variance (as mentioned earlier), it is still informative to look at these sets of

neighbors as they give us some indication of the density of the property distribution within our structure space. If the correlation coefficient drops off rapidly, it is necessarily the result of selecting neighbors whose measured property values are significantly different from those of the probe chemical. In this case either our chemical space is too sparse or it is not a good space for modeling the property of interest.

The results for the two similarity spaces are presented in Figures 8 and 9. It is interesting to note that the curves for the arbitrary and tailored similarity spaces appear almost identical, while the correlation coefficient (*R*) for the tailored similarity space remains about three-one-hundredths above the value for the arbitrary space. An identical trend is found on the plot of standard error (*s.e.*) versus *k* (Figure 9). The best model using tailored similarity uses the four nearest neighbors. This model has the highest correlation coefficient (*R* = 0.884) and the lowest standard error (*s.e.* = 0.346) of all of the models. Figure 10 presents a scatterplot of estimated versus experimental values for normal vapor pressure for the set of 469 chemicals.

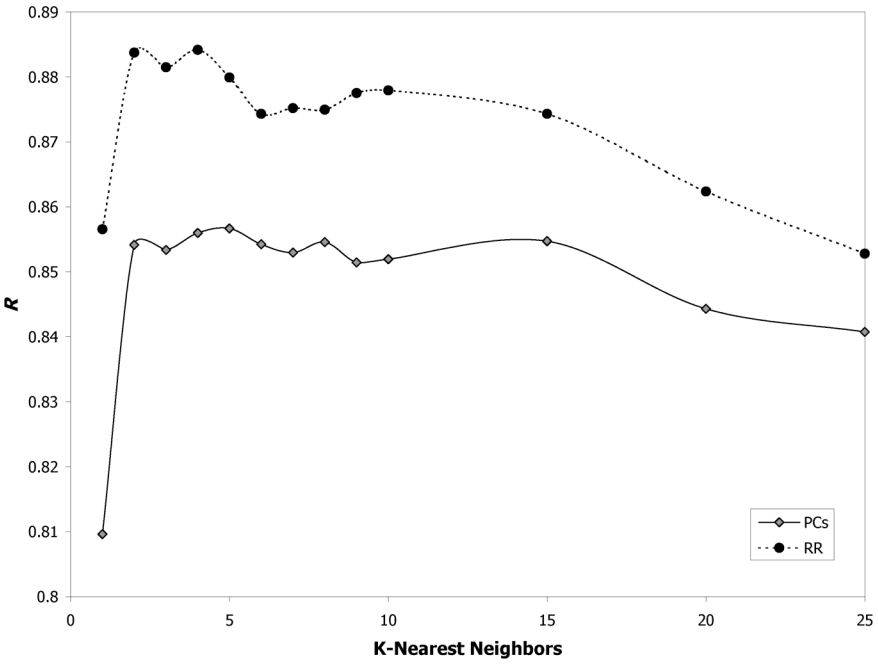


Figure 8. Plot of correlation coefficient (*R*) versus *k*-nearest neighbors (*k* = 1–10, 15, 20, 25) for modeling normal vapor pressure using arbitrary and tailored similarity methods.

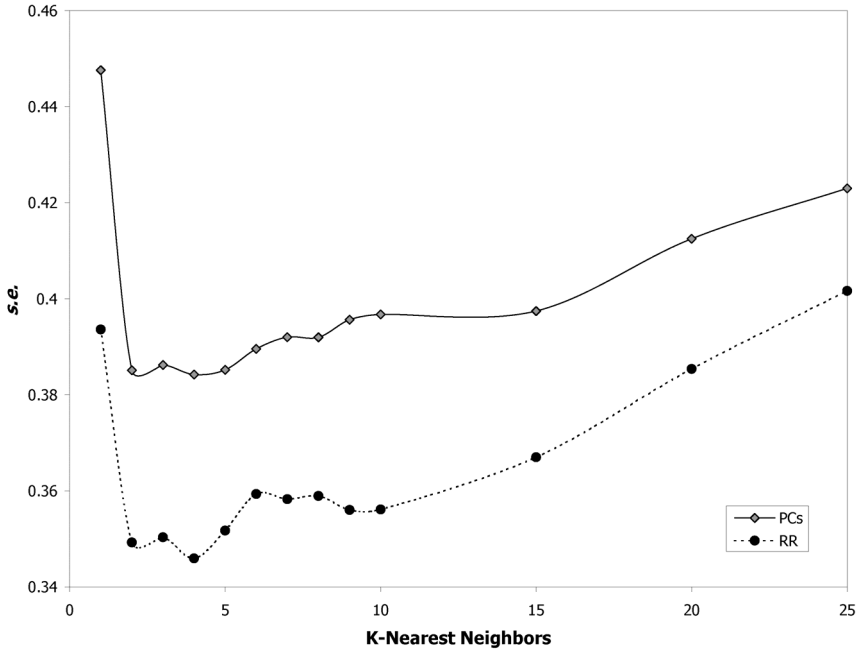


Figure 9. Plot of standard error (*s.e.*) versus *k*-nearest neighbors (*k* = 1–10, 15, 20, 25) for modeling normal vapor pressure using arbitrary and tailored similarity methods.

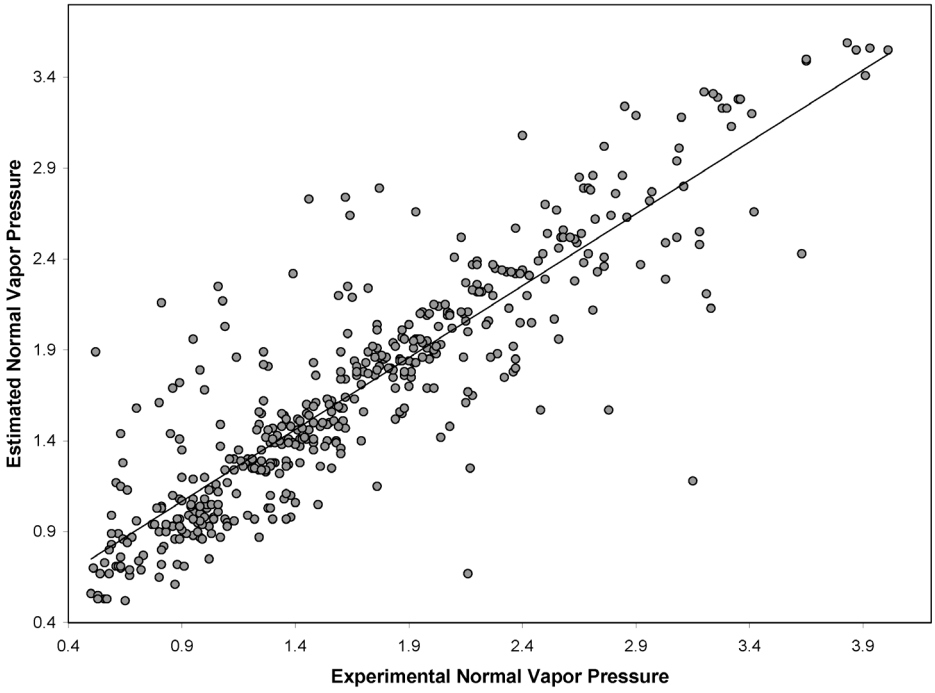


Figure 10. Plot of experimental versus estimated normal vapor pressure for 469 TSCA chemicals using the 4-nearest neighbors selected using the tailored similarity method.

8. Molecular Dissimilarity in Clustering of Databases

8.1 Background

In addition to the selection of analogs and property estimation, similarity spaces can be used in cluster-analysis. This technique assesses the molecular similarity and examines the distances between molecules within the similarity space to form clusters of related compounds. The clusters are formed around a central point, the centroid, and have a set radius based on the molecular density around the centroid. The distance from the cluster centroid to any compound within that cluster can be measured, telling us which compounds are nearest the centroid and which compounds are furthest away. This type of study is useful in scanning large real or virtual chemical libraries in looking for new pharmaceutical leads or for other testing problems in which the number of compounds is simply too large, and therefore it is too expensive, to subject the entire set to proper toxicological screening. In this situation, representatives from each of the clusters can be tested on the assumption that since the compounds within each cluster are similar, their properties should also be similar.

In the field of toxicology, there are a great many situations where there is a need to respond to the toxicity profiles of complex mixtures when the corresponding toxicity data for the individual components are not available. In this case, one can use clustering to find related groups of chemicals and to identify the major component types in the mixture. In this way, if our assumption that similar chemicals are clustered together and since they are structurally similar they should also exhibit similar properties, then our clusters should be fairly homogeneous in terms of their activity and toxicity profiles. In essence, if the clusters are small enough, we could consider the chemicals within the cluster to be the same. In this way, we can narrow down the number of chemicals which need to tested to determine the true culprits in mixture toxicity. In other words, if we have twelve well-characterized clusters we can select the chemical closest to the centroid as being representative of the cluster and test 12! mixtures rather than 120! or more. For example, there are more than 228 identified components of JP-8, the jet fuel mixture in widespread use by the United States Armed Forces. This complex fuel mixture has been shown to cause skin irritation, immunosuppression, and systemic toxicity.[95] If we were to test all possible mixtures of known JP-8 chemicals, then we would have to test 228! mixtures. However, if we can successfully cluster the JP-8 constituents into 15 or 20 well-defined clusters, then we have greatly reduced the complexity and the cost, in both time and resources, to evaluate the toxicity of JP-8.

Another important application of clustering is in pharmaceutical drug design. During the discovery portion of pharmaceutical research and development, a prohibitively large set of chemicals is generally available as candidates for screening.[96] In such a situation, testing all compounds exhaustively is not practical. Dissimilarity methods (clustering methods based on TIs or APs) provide a useful means to select a limited subset of "structurally dissimilar" chemicals from the library of drug-like molecules for testing. Lajiness used such method based on PCs

derived from *POLLY* with tremendous success in drug design.[96] Whereas combinatorial chemistry leads to an "explosion" in the number of molecular species, clustering can be used to reduce the size of the problem to a manageable level.

8.2 Clustering of JP-8 chemicals

JP-8 is a complex fuel mixture of many chemicals. One report claims to have characterized 228 distinct chemicals in this mixture, while others claim to have characterized over 1,000 distinct chemicals. And, as was stated earlier, this complex fuel mixture has been shown to have toxic effects on the skin, immune system, and the whole organism. To fully understand the toxicology of JP-8, one would need to carry out experiments on all binary, ternary, quaternary, …, *n*-nary mixtures of the constituents forming the "chemical soup" that is JP-8. This is a combinatorial explosion of 228! (or, in the worst case, 1000!) possible candidate mixtures to be tested experimentally.

In examining the JP-8 database, we have used a reduced subset consisting of the 194 topologically distinct chemicals present within the larger set of 228 JP-8 constituents. Even in this reduced set, in order to ascertain which particular component or which mixture of components is responsible for the toxic effects, one needs to carry out toxicological evaluations of the 194! possible mixtures of JP-8 constituents. This is impractical and would be terribly costly. The goal of our approach of clustering the chemicals is to decrease the amount of laboratory testing necessary for JP-8 chemicals. The idea is that properly selected clusters will have chemically (and hopefully toxicologically) similar chemicals in the each cluster.

The problem that rears its ugly head is the choice of properties on which to base the clustering. A search of the ASTER database of USEPA shows that very few of the JP-8 constituents have measured property values for even common properties such as boiling point, vapor pressure, melting point, or lipophilicity. So, clustering based on experimental data is not feasible at this time. Given this situation we decided to use theoretical molecular descriptors (TIs) computed by *POLLY* to cluster the JP-8 chemicals.

We clustered the JP-8 constituents into a set of seven clusters derived from a set of 94 TIs calculated by *POLLY v2.3*. The name and structure of the chemical closest to the centroid of each of the seven clusters is presented in Figure 11. One chemical, or a limited number of properly chosen chemicals, from each cluster can make the experimental situation much more manageable to the laboratory toxicologist. Table 5 presents the names of the 194 JP-8 constituents grouped into the 7 clusters and sorted by their abundance in the JP-8 mixture. This clustering has also been done using a physicochemical property space and an AP-based structure space. Due to the lack of toxicological data, we have not been able to investigate whether molecules in the same cluster also have similar toxicity profiles. Such studies can be attempted as more experimental data on toxicity and toxic mode of action (MOA) become available.

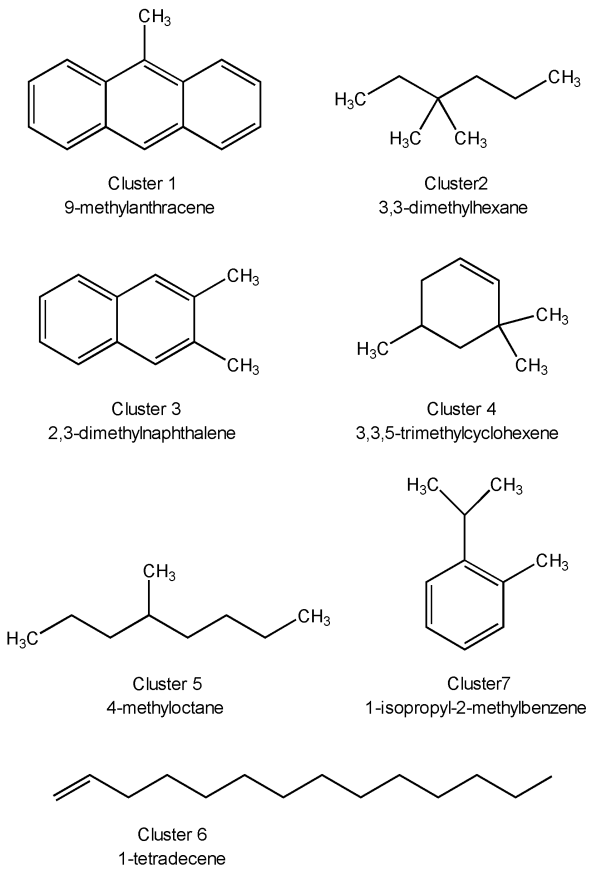


Figure 11. Chemical structures of the JP-8 constituents closest to the center of each cluster for the seven clusters listed in Table 5.

The objective of the JP-8 clustering study was to see whether we could decompose the large set of 194 chemicals into structurally (and hopefully) toxicologically distinct groups. As was stated earlier, the functional homogeneity of the chemicals belonging to the same cluster cannot be verified at this stage owing to the lack of toxicological test data. A perusal of the structures in the various clusters would indicate that chemicals in the same cluster are more similar to one another structurally as compared to those chosen from different clusters. For example, Cluster 1 contains fluorene, pyrene, anthracene, and methanthracene, all large, fused-ring systems; Cluster 2 contains four to eight carbon alkanes that show a fairly high-degree of branching; Cluster 3 consists of bicyclic compounds, primarily napthalene derivatives; and Cluster 4 consists of single ring compounds. Clusters 5 and 6 are a little harder to classify as the mixture of alkanes and alkenes becomes a little less intuitive, but we see that Cluster 7 is also fairly well-characterized as a group of substituted benzene derivatives. While, as was mentioned before, we cannot verify the toxicological relevance of this clustering, the structural aspect of the clustering does seem to be well-defined.

Table 5. 194 JP-8 constituents grouped into seven clusters. The chemicals are ranked within each cluster by their average concentration within the fuel mixture and the distance from each compound to the cluster centroid is noted

Cluster	Chemical name	Avg. Conc.	Distance
1	fluorene	27.5	1.229
1	pyrene	0.1	1.909
1	9-methylanthracene	0.1	0.645
1	ISTD (d10-anthracene)	0.0	1.011
2	3,3- or 2,5-dimethylheptane	1570.0	1.840
2	2,3,3,4-tetramethylpentane	1370.0	2.274
2	3,3-diethylpentane	703.0	1.714
2	2,2,4,4-tetramethylpentane	353.0	2.858
2	2,4-dimethyl-3-ethylpentane	147.0	1.611
2	4,4-dimethylheptane	141.0	1.520
2	2,4-dimethylhexane	126.0	1.594
2	2,3,4-trimethylhexane	114.0	1.541
2	2,3,3-trimethyl-1-butene	96.0	2.640
2	3,3-dimethylhexane	51.3	0.818
2	3,4-dimethylhexane	46.5	1.084
2	2,2,3-trimethylhexane	42.5	1.508
2	2,4,4-trimethylhexane	36.8	1.531
2	3,3-dimethylpentane	21.4	1.714
2	2,3,4-trimethylpentane	20.5	0.976
2	3-ethylpentane	19.0	2.325
2	2,2,3,3-tetramethylpentane	12.6	2.750
2	2,2,3-trimethylbutane	12.4	2.689
2	2,2-dimethylhexane	11.7	1.596
2	iso-octane	9.8	1.461
2	2,3,3-trimethylpentane	9.2	1.280
2	2,4,4-trimethyl-2-pentene	6.9	1.539
2	2,2,4-trimethylhexane	5.5	1.727
2	2,3,3-trimethyl-1,4-pentadiene	4.2	1.529
2	t-3,4,4-trimethyl-2-pentene	3.9	0.837
3	1-methylnaphthalene	3590.0	1.401
3	2-methylnaphthalene	2980.0	1.535
3	naphthalene	2140.0	2.698
3	2,6-dimethylnaphthalene	923.0	0.762
3	1,1'-biphenyl	542.0	1.872
3	2-ethylnaphthalene	455.0	1.363

Table 5. Continued

Cluster	Chemical name	Avg. Conc.	Distance
3	cyclohexylbenzene	387.0	1.403
3	1,2-dimethylnaphthalene	204.0	1.031
3	1,5-dimethylnaphthalene	194.0	0.791
3	1,4-dimethylnaphthalene	191.0	0.814
3	2,3-dimethylnaphthalene	145.0	0.634
3	1,1,6-trimethyltetralin	122.0	2.303
3	1-ethylnaphthalene	119.0	1.267
3	dicyclopentadiene	4.3	2.669
3	1,8-dimethylnaphthalene	2.2	0.911
3	1,3,5-triisopropylbenzene	0.4	3.850
3	hexaethylbenzene	0.2	4.270
4	1,2,4-trimethylbenzene	12300.0	1.830
4	1,2,3-trimethylbenzene	5660.0	1.677
4	1,3,5-trimethylbenzene	4270.0	1.838
4	3-ethyl-2-methylheptane	3260.0	1.733
4	1,1,3-trimethylcyclopentane	2210.0	2.057
4	c-1,2,3-trimethylcyclohexane	2050.0	1.303
4	1,2,4,5-tetramethylbenzene	1950.0	1.984
4	1,2,4-trimethylcyclohexane	1860.0	1.357
4	c-1,3-dimethylcyclohexane	1630.0	1.417
4	1-ethyl-1-methylcyclohexane	840.0	1.194
4	2,2,5,5-tetramethylhexane	474.0	3.034
4	c-1,2-dimethylcyclohexane	440.0	1.386
4	c-1,4-dimethylcyclohexane	327.0	1.658
4	3,4-diethylhexane	302.0	2.387
4	c,t,c-1,2,3-trimethylcyclopentane	192.0	1.287
4	c,c,t-1,2,3-trimethylcyclohexane	175.0	1.303
4	1,3,5-trimethylcyclohexane	115.0	1.253
4	c,c,c-1,3,5-trimethylcyclohexane	113.0	1.253
4	c,c,t-1,3,5-trimethylcyclohexane	91.8	1.253
4	3,5,5-trimethylcyclohexene	65.8	1.166
4	t-1,1,3,4-tetramethylcyclopentane	35.3	1.819
4	1-ethyl-1-methylcyclopentane	20.6	1.315
4	t-1,1,3,5-tetramethylcyclohexane	15.8	2.066
4	2,2,4,6,6-pentamethylheptane	12.1	2.872
4	3,3,5-trimethylcyclohexene	9.0	1.150
4	3,3,5-trimethylheptane	4.9	1.757
4	1,1,3,3-tetramethylcyclopentane	3.8	2.698

Table 5. Continued

Cluster	Chemical name	Avg. Conc.	Distance
4	3,4,5-trimethylheptane	3.6	1.802
4	1,3,5,5-tetramethyl-1,3-cyclohexadiene	1.3	1.757
5	n-decane	33700.0	2.308
5	n-nonane	18400.0	1.618
5	4-methylnonane	6410.0	1.575
5	m & p-xylenes (co-eluting)	6050.0	1.805
5	3-methylnonane	6050.0	1.695
5	2,6-dimethyloctane	5960.0	1.632
5	n-octane	5050.0	1.461
5	3-methyloctane	4160.0	0.824
5	ethylcyclohexane	4150.0	1.611
5	o-xylene	3450.0	1.764
5	3,6-dimethyloctane	2910.0	1.626
5	3,5-dimethyloctane	2900.0	1.734
5	4-methyloctane	1980.0	0.740
5	5-methylnonane	1940.0	1.489
5	2,7-dimethyloctane	1840.0	1.982
5	toluene	1750.0	1.966
5	n-heptane	1700.0	2.113
5	2,3-dimethylheptane	1570.0	1.323
5	2-methylheptane	1490.0	1.017
5	2,6-dimethylheptane	1330.0	1.492
5	3-ethyloctane	1270.0	1.434
5	ethylbenzene	1270.0	1.667
5	2,3-dimethyloctane	1220.0	1.745
5	4-ethyloctane	1180.0	1.439
5	3-heptene	1170.0	2.249
5	4-methylheptane	1050.0	1.176
5	1-nonene	712.0	1.587
5	3-ethylhexane	620.0	1.596
5	3,4-dimethylheptane	536.0	1.727
5	3-methylheptane	502.0	1.086
5	2,4-dimethylheptane	356.0	1.338
5	2-methylhexane	285.0	2.086
5	benzene	270.0	4.275
5	4-ethylheptane	261.0	1.069
5	1-decene	259.0	2.225
5	propylcyclopentane	118.0	1.412

Table 5. Continued

Cluster	Chemical name	Avg. Conc.	Distance
5	1-heptene	115.0	2.239
5	4-nonene	56.4	1.589
5	t-1,3-dimethylcyclopentane	44.6	2.247
5	2-ethyl-1-hexene	29.7	1.209
5	2-octene	27.5	1.509
5	4-methylcyclohexene	6.7	1.827
5	2,5-dimethylhexane	6.6	2.087
5	3,3- or 2,5-dimethylheptane	0.0	1.109
5	m- & p-xylenes (co-eluting)	0.0	1.995
6	n-undecane	28800.0	2.672
6	n-dodecane	27200.0	1.943
6	n-tridecane	26000.0	1.450
6	n-tetradecane	16800.0	1.224
6	2,6-dimethylundecane	10700.0	1.678
6	n-pentadecane	7170.0	1.317
6	2,6,10-trimethyldodecane	5020.0	1.816
6	n-hexadecane	2390.0	1.578
6	2,6,11-trimethyldodecane	1860.0	1.662
6	heptylcyclohexane	1660.0	1.348
6	n-pentylbenzene	621.0	2.196
6	n-hexylbenzene	550.0	1.638
6	1-undecene	490.0	2.755
6	1-dodecene	473.0	1.978
6	1-tridecene	303.0	1.400
6	n-heptylbenzene	273.0	1.361
6	2,6,10,14-tetramethylpentadecane(pristan)	180.0	3.245
6	n-octylbenzene	109.0	1.396
6	cyclododecane	39.5	4.016
6	1-tetradecene	36.7	1.053
6	n-nonylbenzene	25.7	1.592
6	1-hexadecene	25.0	1.322
6	2,6,11,15-tetramethylhexadecane	8.8	3.497
6	n-decylbenzene	5.3	1.863
6	1-phenyltridecane	1.0	2.882
7	2-methyldecane	9020.0	2.679
7	3-methyldecane	8470.0	2.395
7	butylcyclohexane	6880.0	1.825
7	4-methyldecane	6100.0	2.222

Table 5. Continued

Cluster	Chemical name	Avg. Conc.	Distance
7	indane (2,3-dihydro-1H-indene)	4050.0	2.281
7	n-propylcyclohexane	4030.0	1.643
7	1-ethyl-3-methylbenzene	3590.0	1.844
7	1-ethyl-4-methylbenzene	3150.0	1.833
7	1-ethyl-2-methylbenzene	2750.0	1.571
7	isopropylcyclohexane	2740.0	1.579
7	1-ethyl-3,5-dimethylbenzene	2730.0	1.114
7	1-ethyl-2,4-dimethylbenzene	2680.0	0.958
7	2,2,3-trimethyldecane	2670.0	2.939
7	1-propyl-2-methylbenzene	2570.0	1.182
7	1-ethyl-3,4-dimethylbenzene	2140.0	1.178
7	1-isopropyl-4-methylbenzene	2100.0	1.109
7	1,2,3,5-tetramethylbenzene	2040.0	1.639
7	propylbenzene	1950.0	1.933
7	1,2,3,4-tetramethylbenzene	1810.0	1.741
7	1-ethyl-2,5-dimethylbenzene	1720.0	0.874
7	t-butylbenzene	1520.0	1.570
7	isopropylbenzene	1450.0	1.644
7	1,3-diethylbenzene	1430.0	0.879
7	1,2-dimethyl-3-ethylbenzene	1400.0	0.947
7	1-isopropyl-3-methylbenzene	1250.0	1.057
7	1-propyl-4-methylbenzene	1200.0	1.012
7	butylbenzene	1080.0	1.915
7	sec-butylbenzene	832.0	0.841
7	1,4-diethylbenzene	648.0	1.538
7	1,2-diethylbenzene	576.0	0.951
7	1,3,5-triethylbenzene	541.0	1.926
7	2-ethyl-1,3-dimethylbenzene	505.0	0.954
7	isobutylbenzene	458.0	1.215
7	(1-ethylpropyl-)-benzene	399.0	1.046
7	1-ethyl-3-isopropylbenzene	352.0	0.891
7	1-isopropyl-2-methylbenzene	261.0	0.810
7	(1,2-dimethylpropyl-)-benzene	260.0	1.086
7	sec-pentylbenzene	253.0	1.373
7	(3-methylbutyl-)-benzene	242.0	2.018
7	(2-methylbutyl-)-benzene	131.0	1.631
7	1-isopropyl-4-methylcyclohexane	89.4	1.633
7	1-t-butyl-3-methylbenzene	39.2	1.812

Table 5. Continued

Cluster	Chemical name	Avg. Conc.	Distance
7	1,2-diisopropylbenzene	22.2	2.448
7	(1,1-diethylpropyl-)-benzene	22.1	2.884
7	1-t-butyl-3,4,5-trimethylbenzene	20.2	3.304
7	1-t-butyl-3,5-dimethylbenzene	17.2	2.544
7	1,4-diisopropylbenzene	14.4	2.242
7	neopentylbenzene	14.1	1.734
7	3,7,7-trimethylbicyclo(4.1.0)-3-heptene	10.7	3.221

Clustering of JP-8 chemicals mentioned above was carried out using arbitrary similarity spaces. A tailored space could not be derived because of the lack of test data. Tailored spaces based on toxicity data and toxicologically relevant properties might lead to clustering relevant to the hazard assessment of JP-8.

8.3 Psoralen studies

More recently, we developed a large virtual library of nearly 250,000 psoralen derivatives[97] to show the utility of the combinatorial-cum-clustering method in practical drug design. We created a virtual library of 248,832 psoralen derivatives and clustered them. Figure 12 shows the psoralen skeleton, the substitution points, and the twelve substituent groups used in the creation of this virtual library. These structures were grouped into a limited number of clusters (15, 20, and 25) and the clusters were examined. Table 6 lists the 25 clusters and presents the number of compounds in each cluster along with the minimum and maximum distance from the centroid within each of the clusters. The compounds closest to the centroid of each cluster are summarized in Table 7. This technique allows for the selection of a small number of compounds, representative of each of the clusters, to form a manageable subset for costly laboratory testing. One can choose either the chemical closest to the centroid of each cluster or some small subset from each cluster as representative of the cluster. In this way, one can test a smaller number of chemicals as compared to the large set of nearly 250,000 psoralen derivatives and examine the potential of the large set of structures. Whereas combinatorial chemistry leads to an "explosion" in the number of candidate chemicals, clustering can be used to bring the number of candidate chemicals down to a manageable level.

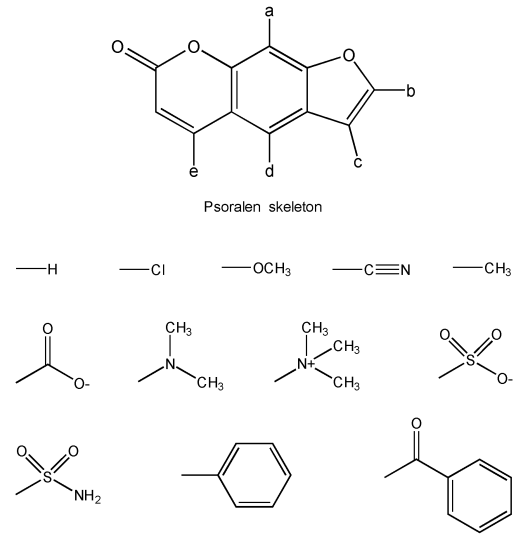


Figure 12. Structural skeleton of psoralen with five possible substitution points identified and the 12 substituents used to generate a virtual library of 248,832 psoralen derivatives.

Table 6. Minimum distance and maximum distance to cluster center and cluster size

Cluster	Minimum Distance	Maximum Distance	N
1	1.83	7.57	47
2	1.31	8.43	9484
3	1.25	11.69	1085
4	1.99	8.34	2141
5	0.95	8.44	1303
6	1.15	8.45	31756
7	3.13	8.32	393
8	1.72	7.48	8072
9	2.57	9.94	249
10	0.93	9.54	6297
11	1.73	9.65	3497
12	0.92	9.57	8780
13	0.43	9.28	38286
14	2.77	7.70	950
15	0.69	10.38	31526
16	1.87	7.58	720
17	1.07	9.03	20907
18	0.70	11.90	7186
19	0.86	10.20	7391
20	1.95	9.85	11320
21	1.47	10.53	5346
22	1.09	9.22	36661
23	1.19	8.61	7906
24	2.09	8.47	1605
25	0.97	10.19	5924

Table 7. Substituents at each position for the chemical closest to cluster center

Cluster	a	b	c	d	e
1	–H	–H	–H	–H	–Cl
2	$–N^+(CH_3)_3$	$–OCH_3$	$–CHOC_6H_6$	$–N^+(CH_3)_3$	–CN
3	$–SO_3–$	$–N^+(CH_3)_3$	$–N^+(CH_3)_3$	$–N^+(CH_3)_3$	$–C_6H_6$
4	$–C_6H_6$	–H	–Cl	–H	$–CH_3$
5	–H	–H	–CN	$–N^+(CH_3)_3$	$–N^+(CH_3)_3$
6	$–N^+(CH_3)_3$	$–SO_3–$	$–OCH_3$	–CN	$–OCH_3$
7	–H	–H	–H	–CN	$–N^+(CH_3)_3$
8	$–CH_3$	$–CO_2–$	$–CO_2–$	–Cl	–H
9	$–N^+(CH_3)_3$	–H	–H	$–N^+(CH_3)_3$	$–N(CH_3)_2$
10	$–CHOC_6H_6$	$–CH_3$	–H	$–CHOC_6H_6$	$–CO_2–$
11	$–C_6H_6$	$–CHOC_6H_6$	$–C_6H_6$	$–CHOC_6H_6$	$–SO_3–$
12	–Cl	–H	$–CH_3$	$–CHOC_6H_6$	$–OCH_3$
13	$–SO_2NH_2$	$–SO_3–$	$–C_6H_6$	$–CH_3$	$–N(CH_3)_2$
14	–CN	$–OCH_3$	–H	–Cl	–H
15	$–N(CH_3)_2$	$–CHOC_6H_6$	$–CHOC_6H_6$	–H	$–SO_3–$
16	–CN	$–CH_3$	–H	–H	$–CH_3$
17	$–OCH_3$	$–CHOC_6H_6$	–Cl	$–OCH_3$	$–OCH_3$
18	$–N^+(CH_3)_3$	$–SO_2NH_2$	$–N(CH_3)_2$	$–N^+(CH_3)_3$	–H
19	$–N^+(CH_3)_3$	$–SO_2NH_2$	–H	$–C_6H_6$	–CN
20	$–SO_2NH_2$	$–SO_2NH_2$	$–N(CH_3)_2$	$–N^+(CH_3)_3$	$–SO_2NH_2$
21	$–CO_2–$	–H	$–CH_3$	–Cl	$–SO_2NH_2$
22	$–CO_2–$	$–OCH_3$	$–CH_3$	$–CO_2–$	$–SO_3–$
23	$–CH_3$	$–CO_2–$	$–CH_3$	–H	$–N(CH_3)_2$
24	$–N(CH_3)_2$	$–N(CH_3)_2$	$–N(CH_3)_2$	$–CO_2–$	$–N(CH_3)_2$
25	–CN	–CN	–H	$–N^+(CH_3)_3$	–CN

9. Similarity in the Post-Genomic Era

We usually compute similarity of chemicals in terms of calculated descriptors, the main reason being that they are available for any arbitrary chemical species. But physical properties and biological test data can also be used for this purpose, if available. In this post-genomic era, large sets of data on effects of drugs and toxicants on the genome and proteome are gradually becoming available. Such data can add context-specific information in computing the similarity/dissimilarity of chemicals for certain purposes. In particular, our Natural Resources Research Institute team and collaborators have been involved in developing compact descriptors for DNA sequences and proteomics patterns which, in analogy with calculated chemodescriptors, describe biological molecules or systems such as DNA sequences or the patterns of distribution

of proteins in the 3-D gels. It is tempting to speculate that such "biodescriptors" will find application in SAR studies and similarity/dissimilarity analyses of molecular systems and provide new insight about their biological action.

10. Discussion and Conclusions

In this paper we have reviewed results of QMSA studies carried out by our research team during the past fifteen years. Most of the similarity measures are based on algorithmically-derived parameters. We have carried out some limited number of studies in which property-based and structure-based QMSA methods have been compared. The arbitrary property-based methods seem to select "better" analogs and give superior activity estimates as compared to arbitrary structure-based QMSA methods. Further studies are needed in this area with large data sets and more diverse properties before any definite conclusions can be made in this matter.

As has been mentioned previously, our research has shown repeatedly that a small set of nearest neighbors (generally 3 to 8 structural analogs) provides the best property estimates for a given probe chemical.[66,76–80,88–90] In addition, in a recent study we have shown the danger of using too many neighbors, especially in small (fewer than 200 chemical) data sets, though our study showed similar problems for a database of 469 chemicals.[89] Using too many neighbors decreases the over-all variance in the data, with a potential for marginalizing the predictions and under-predicting highly active chemicals (or over-predicting chemicals of low activity). So again, it is best to stay with relatively small sets of neighbors, somewhere between 3–8. See our 2001 study in the *Journal of Molecular Graphics and Modelling* for more details.[89]

A new addition in the QMSA arena is tailored similarity, a novel approach where context or property-specific descriptors are used to create a structure space for the computation of intermolecular similarity. Our limited experience in this area shows that this method holds great promise for application to chemical domains relevant to drug design and the hazard assessment of chemicals.[90–93] However, much work remains to be done in validating this methodology. Also, techinques are needed to identify the domain of applicability of the similarity space. As with all similarity spaces, the user must be cautious when adding new, unknown chemicals to the space. Those that fall beyond the bounds of the chemicals currently within the structure space are easily identified by descriptor values outside the ranges for those descriptors within the data set analyzed. However, the user must also remember that the outer bounds of the structure domain are likely to be sparsely populated, and predictions for chemicals along these outer "edges" of the space will be biased, since neighbors will not be evenly distributed around the probe compound. If we consider the prediction space to be a perfect sphere, with hundreds of points evenly distributed within that sphere, a point that lies well within the area of the sphere will be surrounded by evenly distributed points to a certain distance less than the radius of the sphere. Those points that lie near or along the outer surface of the sphere will have an uneven distribution of neighboring points. There will be more neighboring points between our point of

interest and the center of the sphere than between that point and the outer surface of the sphere. In the case of a point along that outer surface, if we surround that point with a small sphere, less than half of that sphere will be filled with neighboring points (the intersection of the small sphere and the structure space). So what does this mean? While we can continue to make predictions about chemicals that lie along or near the edge of our structure space (or even slightly outside that defined space), if we are being cautious (or need to be conservative in our predictions) the "domain of applicability" or prediction space for the structure space should be defined as a subset of the structure space which is well characterized; i.e., a subset well within the bounds of the full structure space where even compounds on the outer edge of the prediction space are well-defined in terms of surrounding neighbors in the greater structure space.

And finally, in the future it will be interesting to see whether integrated structure spaces consisting of chemodescriptors and biodescriptors (descriptors of relevant biological activity) allow for increased selectivity in analog selection and estimation of biomedicinal and toxicological properties.

11. Acknowledgements

This is contribution number 325 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper has been supported by several grants from the United States Air Force. The most recent work was funded by U.S. Air Force grant FA9550-05-1-0456.

12. References

1. CAS Registry Number and Substance Counts. http://www.cas.org/cgi-bin/regreport.pl (accessed Mar 2006).
2. Auer, C.M.; Zeeman, M.; Nabholz, J. V.; Clements, R. G. SAR—the U.S. regulatory perspective. *SAR QSAR Environ. Res.* **1994,** *2*, 29.
3. *Goodman and Gilman's The Pharmacological Basis of Therapeutics;* Goodman, A. G., Wall, T. W., Nies, A. S., Taylor, P., Eds.; Pergamon Press: New York, 1990.
4. Thornber, C. W. Isosterism and molecular modification in drug design. *Chem. Soc. Rev.* **1979,** *8*, 563.
5. *Advances in Molecular Similarity. Vol. 2;* Carbo-Dorca, R., Mezey, P. G., Eds.; JAI Press: Stamford, CN., 1998.
6. Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity;* John Wiley & Sons, Inc.: New York, 1990.

7. Whyte, L. L. Atomism, structure and form: a report on the natural philosophy of form. In *Structure in Art and Science;* Kepes, G., Ed.; George Braziler, Inc.: New York, 1965; pp 20–28.
8. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math*. **1988,** *19*, 17.
9. Weininger, S. J. The molecular structure conundrum: Can classical chemistry be reduced to quantum chemistry? *J. Chem. Educ.* **1984,** *61*, 939.
10. Woolley, R. G. Must a molecule have a shape? *J. Am. Chem. Soc*. **1978,** *100*, 1073.
11. Basak, S. C. Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Med. Sci. Res*. **1987**, *15*, 605.
12. Kier, L. B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976.
13. Richards, W. G. *Quantum Pharmacology;* Butterworths: London, 1977.
14. Grossman, S. C. Chemical ordering of molecules—A graph theoretical approach to structure-property studies. *Int. J. Quantum Chem*. **1985,** *28*, 1.
15. Bunge, M. *Method, Model and Matter;* D. Reidel Publishing Co.: Dordrecht-Holland/Boston, 1973.
16. Sylvester, J. J. On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics. *Amer. J. Math*. **1878,** *1*, 64.
17. Balaban, A. T. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **1985,** *25*, 334.
18. Basak, S. C.; Niemi, G. J.; Veith, G. D. Recent developments in the characterization of chemical structure using graph-theoretic indices. In *Computational Chemical Graph Theory;* Rouvray, D. H., Ed.; Nova Science Publishers, Inc.: Commack, NY, 1990; pp 235–277.
19. Johnson, M.; Basak, S.C.; Maggiora, G. A characterization of molecular similarity methods for property prediction. *Mathl. Comput. Modelling* **1988,** *11*, 630.
20. O'Brien, S. E.; Popelier, P. L. A. Quantum molecular similarity. Part 2: The relation between properties in BCP space and bond length. *Can. J. Chem*. **1999,** *77*, 28.
21. O'Brien, S. E.; Popelier, P. L. A. Quantum molecular similarity. 3. QTMS descriptors. *J. Chem. Inf. Comput. Sci*. **2001,** *41*, 764.
22. O'Brien, S. E.; Popelier, P. L. A. Quantum topological molecular similarity. Part 4. A QSAR study of cell growth inhibitory properties of substituted (*E*)-1-phenylbut-1-en-3-ones. *J. Chem. Soc., Perkin Trans*. **2002,** *2*, 478.
23. Popelier, P. L. A. Quantum molecular similarity. 1. BCP space. *J. Phys. Chem*. **1999,** *103*, 2883.
24. Trinajstic, N. *Chemical Graph Theory,* 2nd Edn..; CRC Press: Boca Raton, FL, 1992.
25. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis;* Research Studies Press: Letchworth, Hertfordshire, U.K., 1986.
26. Kier, L. B.; Murray, W. J.; Randic, M.; Hall, L. H. Molecular connectivity. V. Connectivity series concept applied to diversity. *J. Pharm. Sci*. **1976,** *65*, 1226.

27. Wiener, N. *Cybernetics;* Wiley: New York, 1948.
28. Trinajstic, N. *Chemical Graph Theory, Vols. I & II.;* CRC Press: Boca Raton, Florida, 1983.
29. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948,** *27*, 379.
30. Rashevsky, N. Life, information theory and topology. *Bull. Math. Biophys*. **1955,** *17*, 229.
31. Trucco, E. On the information content of graphs: compound symbols: different states for each point. *Bull. Math. Biophys*. **1956,** *18*, 237.
32. Trucco, E. A note on the information content of graphs. *Bull. Math. Biophys*. **1956,** *18*, 129.
33. Sarkar, R.; Roy, A. B.; Sarkar, R. K. Topological information content of genetic molecules - I. *Math. Biosci.* **1978,** *39*, 299.
34. Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood complexities and symmetry of chemical graphs and their biological applications. In *Mathematical Modelling in Science and Technology;* Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: Elmsford, NY, 1984; pp 745–750.
35. Magnuson, V. R.; Harriss, D. K.; Basak, S. C. Topological indices based on neighborhood symmetry: Chemical and biological applications. In *Chemical Applications of Topology and Graph Theory;* King, R. B., Ed.; Elsevier: New York, 1983; pp 178–191.
36. Mowshowitz, A. Entropy and the complexity of graphs: I. An index of the relative complexity of a graph. *Bull. Math. Biophys.* **1968,** *30*, 175.
37. Mowshowitz, A. Entropy and the complexity of graphs: II. The information content of digraphs and infinite graphs. *Bull. Math. Biophys*. **1968,** *30*, 225.
38. Mowshowitz, A. Entropy and the complexity of graphs: III. Graphs with prescribed information content. *Bull. Math. Biophys*. **1968,** *30*, 387.
39. Mowshowitz, A. Entropy and the complexity of graphs: IV. Entropy measures and graphical structure. *Bull. Math. Biophys*. **1968,** *30*, 533.
40. Basak, S. C.; Roy, A. B.; Ghosh, J. J. Study of the structure-function relationship of pharmacological and toxicological agents using information theory. In *Proceedings of the Second International Conference on Mathematical Modelling;* Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, Missouri, 1980; pp 851.
41. Basak, S. C.; Magnuson, V. R. Molecular topology and narcosis: A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim.-Forsch./Drug Res*. **1983,** *33*, 501.
42. Bonchev, D.; Trinajstic, N. Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **1977,** *67*, 4517.
43. Schreider, J.A. *Equality, Resemblance, and Order;* Mir Publishers: Moscow, 1975.
44. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. Topological indices: Their nature, mutual relatedness, and applications. *Mathl. Model.* **1987,** *8*, 300.
45. Russom, C. L.; Anderson, E. B.; Greenwood, B. E.; Pilli, A. ASTER: An integration of the AQUIRE data base and the QSAR system for use in ecological risk assessments. *Sci. Total Environ.* **1991,** *109/110*, 667.

46. Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. Correlations between chemical structures and normal boiling points of halogenated alkanes C_1-C_4. *J. Chem. Inf. Comput. Sci*. **1992,** *32*, 233.
47. Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. Correlation between structure and normal boiling points of haloalkanes C_1-C_4 using neural networks. *J. Chem. Inf. Comput. Sci*. **1994,** *34*, 1118.
48. Needham, D. E.; Wei, I. C.; Seybold, P. G. Molecular modelling of the physical properties of alkanes. *J. Am. Chem. Soc*. **1988,** *110*, 4186.
49. Mekenyan, O.; Bonchev, D.; Trinajstic, N. Chemical graph theory: Modeling the thermodynamic properties of molecules. *Int. J. Quant. Chem*. **1980,** *18*, 369.
50. *Spectral Atlas of Polycyclic Aromatic Hydrocarbons, Vol. 2;* Karcher, W., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 1988.
51. Leo, A.; Weininger, D. *CLOGP Version 3.2 User Reference Manual;* Pomona College: Claremont, CA, 1984.
52. Basak, S. C. *H-Bond;* University of Minnesota: Duluth, MN, 1988.
53. Brooke, D. N.; Dobbs, A. J.; Williams, N. Octanol:water partition coefficients (*P*): Measurement, estimation and interpretation, particularly for chemicals with $P > 10.5$. *Ecotoxicol. Environ. Safety* **1986,** *11*, 251.
54. De Brujin, J.; Busser, F.; Seinen, W.; Hermens, J. Measuring octanol/water partition coefficients with a "slow-stirring method". *Environ. Toxicol. Chem*. **1989,** *8*, 499.
55. Basak, S. C.; Grunwald, G. D. Use of topological space and property space in selecting structural analogs. *Mathl. Model. Sci. Computing* **1994,** *4*, 464.
56. Kamlet, M. J. White Oak Laboratory, Silver Spring, MD. Personal communication, 1987.
57. AFOSR JP-8 Jet Fuel Workshop, University of Arizona, Tucson, Arizona, 2000.
58. Hansch, C.; Yoshimoto, M. Structure-activity relationships in immunochemistry. 2. Inhibition of complement by benzamidines. *J. Med. Chem.* **1974,** *17*, 1160.
59. Baker, B. R.; Cory, M. Irreversible enzyme inhibitors. CLXIV. Proteolytic enzymes. XIV. Inhibition of guinea pig complement by meta-substituted benzamidines. *J. Med. Chem.* **1969,** *12*, 1049.
60. Baker, B. R.; Cory, M. Irreversible enzyme inhibitors. CLXV. Proteolytic enzymes. XV. Inhibition of guinea pig complement by derivatives of m-phenoxypropoxybenzamidine. *J. Med. Chem*. **1969,** *12*, 1053.
61. Baker, B. R.; Cory, M. Irreversible enzyme inhibitors. 180. Irreversible inhibitors of the C'1a component of complement derived from *m*-(phenoxypropoxy) benzamidine and phenoxyacetamide. *J. Med. Chem*. **1971,** *14*, 119.
62. Baker, B. R.; Cory, M. Irreversible enzyme inhibitors. 186. Irreversible inhibitors of the C'1a component of complement derived from *m*-(phenoxypropoxy) benzamidine by bridging to a terminal sulfonyl fluoride. *J. Med. Chem*. **1971,** *14*, 805.
63. Cohen, G. M.; Mannering, G. J. Involvement of a hydrophobic site in the inhibition of the microsomal *p*-hydroxylation of aniline by alcohols. *Mol. Pharmacol*. **1973,** *9*, 383.

64. Hall, L.H.; Kier, L.B.; Phipps, G. Structure-activity relationship studies on the toxicities of benzene derivatives: I. An additivity model. *Environ. Toxicol. Chem*. **1984,** *3*, 355.
65. Soderman, J. V. *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database. Vol. 1;* CRC Press: Boca Raton, FL, 1982.
66. Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of graph theoretic parameters in risk assessment of chemicals. *Toxicol. Lett*. **1995,** *79*, 239.
67. Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in Salmonella typhimurium TA98 and TA100. *Environ. Mol. Mutagen*. **1992,** *19*, 37.
68. Benigni, R.; Andreoli, C.; Giuliani, A. QSAR models for both mutagenic potency and activity: Application to nitroarenes and aromatic amines. *Environ. Mol. Mutagen*. **1994,** *24*, 208.
69. McCann, J.; Choi, E.; Yamasaki, E.; Ames, B. N. Detection of carcinogens as mutagens in the *Salmonella*/microsome test: Assay of 300 chemicals. *Proc. Natl. Acad. Sci.* **1975,** *72*, 5135.
70. Yamaguchi, T.; Hasegawa, R.; Hagiwara, A.; Hirose, M.; Imaida, K.; Ito, N.; Shirai, T. Results for 277 chemicals in the medium term liver carcinogenesis bioassay of rats, 1999.
71. Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A comparative study of molecular similarity, statistical, and neural network methods for predicting toxic modes of action of chemicals. *Environ. Toxicol. Chem*. **1998,** *17*, 1056.
72. Basak, S. C.; Gute, B. D.; Grunwald, G. D. Quantitative comparison of five molecular structure spaces in selecting analogs of chemicals. *Mathl. Modelling Sci. Comput.*, in press.
73. Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Application of graph theoretical parameters in quantifying molecular similarity and structure-activity relationships. *J. Chem. Inf. Comput. Sci*. **1994,** *34*, 270.
74. Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of graph invariants in QMSA and predictive toxicology, DIMACS Series 51. In *Discrete Mathematical Chemistry;* Hansen, P., Fowler, P., Zheng, M., Eds.; American Mathematical Society: Providence, RI, 2000; pp 9–24.
75. Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and application of molecular similarity methods using nonempirical parameters. *Mathl. Modelling Sci. Comput.*, in press.
76. Basak, S. C.; Grunwald, G. D. Molecular similarity and risk assessment: analog selection and property estimation using graph invariants. *SAR QSAR Environ. Res*. **1994,** *2*, 289.
77. Basak, S. C.; Grunwald, G. D. Estimation of lipophilicity from molecular structural similarity. *New J. Chem*. **1995,** *19*, 231.
78. Basak, S. C.; Grunwald, G.D. Molecular similarity and estimation of molecular properties. *J. Chem. Inf. Comput. Sci*. **1995,** *35*, 366.
79. Basak, S. C.; Grunwald, G. D. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: A similarity based study. *Chemosphere* **1995,** *31*, 2529.

80. Basak, S. C.; Grunwald, G. D. Tolerance space and molecular similarity. *SAR QSAR Environ. Res*. **1995,** *3*, 265.
81. Basak, S. C.; Gute, B. D. Characterization of molecular structures using topological indices. *SAR QSAR Environ. Res*. **1997,** *7*, 1.
82. Basak, S. C.; Gute, B. D. Use of graph theoretic parameters in predicting inhibition of microsomal hydroxylation of anilines by alcohols: a molecular similarity approach. In *Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health;* Johnson, B. L., Xintaras, C., Andrews, J. S., Eds.; Princeton Scientific Publishing Co, Inc.: Princeton, NJ, 1997; pp 492–504.
83. Gute, B. D.; Grunwald, G. D.; Mills, D.; Basak, S. C. Molecular similarity based estimation of properties: A comparison of structure spaces and property spaces. *SAR QSAR Environ. Res.* **2001,** *11*, 363.
84. Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985,** *25*, 64.
85. Basak, S. C.; Grunwald, G. D. *APProbe*; University of Minnesota: Duluth, MN, 1993.
86. *SAS/STAT User's Guide*, Release 6.03; SAS Institute Inc.: Cary, NC, 1988.
87. Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ. Health Perspect*. **1990,** *87*, 183.
88. Basak, S. C.; Gute, B. D.; Grunwald, G. D. Characterization of the molecular similarity of chemicals using topological invariants. In *Advances in Molecular Similarity*, *Vol. 2;* Carbo-Dorca, R., Mezey, P. G., Eds.; JAI Press: Stamford, CT, 1998; pp 171–185.
89. Gute, B. D.; Basak, S. C. Molecular similarity-based estimation of properties: A comparison of three structure spaces. *J. Mol. Graphics and Model.* **2001,** *20*, 95.
90. Basak, S. C.; Gute, B. D.; Mills, D.; Hawkins, D. M. Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: A comparison of arbitrary versus tailored similarity spaces. *J. Mol. Struct. (THEOCHEM)* **2003,** *622*, 127.
91. Gute, B. D.; Basak, S. C.; Mills, D.; Hawkins, D. M. Tailored similarity spaces for the prediction of physicochemical properties. *Internet Electr. J. Mol. Design* **2002,** *1*, 374.
92. Basak, S. C.; Gute, B. D.; Mills, D. Quantitative molecular similarity analysis (QMSA) methods for property estimation: A comparison of property-based, arbitrary, and tailored similarity spaces. *SAR QSAR Environ. Res.* **2002,** *13*, 727.
93. Gute, B. D.; Basak, S. C. 2006. Optimal neighbor selection in molecular similarity: Comparison of arbitrary versus tailored prediction spaces. *SAR QSAR Environ. Res*. **2006,** *17*, 37.
94. Hawkins, D.; Basak, S.; Shi, X. QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.* **2001,** *41*, 663.

95. Basak, S. C.; Mills, D. Quantitative structure-property relationships (QSPRs) for the estimation of vapor pressure: A hierarchical approach using mathematical structural descriptors. *J. Chem. Inf. Comput. Sci.* **2001,** *41*, 692.
96. Lajiness, M. Molecular similarity-based methods for selecting compounds for screening. In *Computational Chemical Graph Theory;* Rouvray, D. H., Ed.; Nova Science Publishers, Inc.: Commack, NY, 1990; pp 299–316.
97. Basak, S. C.; Mills, D.; Gute, B. D.; Balaban, A. T.; Basak, K.; Grunwald, G. D. Use of mathematical structural invariants in analyzing combinatorial libraries: A case study with Psoralen derivatives. In *Some Aspects of Mathematical Chemistry;* Sinha, D. K., Basak, S. C., Mohanty, R. K., Basumallick, I. N., Eds.; Visva-Bharati University: Shantineketan, India, in press.

Authors' Biographical Data

Subhash C. Basak was born in Calcutta, India. He received a B.S. degree in Chemistry (1966), a M.S. degree in Biochemistry (1968), and a Ph.D. in Biochemistry (1980) from the University of Calcutta. In 1982, he was appointed Faculty Research Associate for the Department of Chemistry at the University of Minnesota Duluth. He has worked as a Faculty Research Associate for the Natural Resources Research Institute of the University of Minnesota Duluth since 1987 and also holds adjunct professorships for the Department of Chemistry and the Department of Biochemistry and Molecular Biology. He is the president of the International Society for Mathematical Chemistry and an editorial board member for the journals *SAR and QSAR in Environmental Research* and *Journal of Computer Information and Modeling*. He has authored 165 papers, reviews, and book chapters. He is also a co-chair of the Indo-US Workshop on Mathematical Chemistry, with Applications to Drug Discovery, Environmental Toxicology, Cheminformatics and Bioinformatics conference series. Areas of research interest include chemical graph theory and topology, development of topological indices, mathematical chemistry, predictive toxicology, and structure-activity relationship (SAR) modeling in drug design and environmental chemistry.

Brian D. Gute was born in Owatonna, Minnesota, USA. He received a B.A. degree in Chemistry and English (1993) and a M.S. degree in Toxicology (2001) from the University of Minnesota. He has been working as part of Dr. Subhash Basak's research group since 1995 and became a full-time researcher at the University of Minnesota in 2000. Since then, he has continued his work as a member of Dr. Basak's research team as a Research Fellow (2004) at the Natural Resources Research Institute of the University of Minnesota Duluth. His research publications include 30 journal articles and 13 book chapters. Mr. Gute's research interests include chemical graph theory and topology, quantitative molecular similarity analysis, development of biodescriptors, predictive toxicology, and computer-aided drug design and environmental chemistry.

Denise Mills was born in Iron River, Michigan, USA. She received a B.S. degree in Chemistry, cum laude, from the University of Minnesota Duluth in 1992. She joined the research group of Subhash Basak at the Natural Resources Research Institute of the University of Minnesota Duluth, working in the field of computational chemistry. Publications include 40 journal articles and book chapters. Research interests include quantitative structure-activity relationship modeling in chemical design and environmental chemistry.