

The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies

Jean-Loup Faulon*

Sandia National Laboratories, P.O. Box 969, MS 9951, Livermore, California 94551

Donald P. Visco, Jr. and Ramdas S. Pophale

Department of Chemical Engineering, Tennessee Technological University, Box 5013, Cookeville, Tennessee 38502

Received October 29, 2002

We present a new descriptor named signature based on extended valence sequence. The signature of an atom is a canonical representation of the atom's environment up to a predefined height h . The signature of a molecule is a vector of occurrence numbers of atomic signatures. Two QSAR and QSPR models based on signature are compared with models obtained using popular molecular 2D descriptors taken from a commercially available software (Molconn-Z). One set contains the inhibition concentration at 50% for 121 HIV-1 protease inhibitors, while the second set contains 12865 octanol/water partitioning coefficients (Log P). For both data sets, the models created by signature performed comparable to those from the commercially available descriptors in both correlating the data and in predicting test set values not used in the parametrization. While probing signature's QSAR and QSPR performances, we demonstrate that for any given molecule of diameter D , there is a molecular signature of height $h \leq D+1$, from which any 2D descriptor can be computed. As a consequence of this finding any QSAR or QSPR involving 2D descriptors can be replaced with a relationship involving occurrence number of atomic signatures.

INTRODUCTION

By representing the two-dimensional structure of a molecule as a graph complete with vertices (atoms) and edges (bonds), an endless number of operators on this graph can be composed in an attempt to characterize the properties of the molecule. These operators, known as descriptors, need not have any theoretical basis and many times have only qualitative meaning, at least at the outset of their introduction, and a more-refined interpretation can come much later.¹ The numerical values resulting from the operation of a given descriptor on a graph are normally used in quantitative-structure relationships (QSR) as independent variables to correlate and predict various experimental physical properties (QSPR) or biological activities (QSAR).

Though the three-dimensional structure, or topography, of a molecule contains more information about the spatial relationships of the atoms and bonds than the two-dimensional chemical graph, the topological, or 2D descriptors relative to 3D descriptors have been shown in some studies to contain more information content.²

The number of descriptors available to use for a QSR are numerous, and the advent of commercial packages with QSR capabilities make the selection of the proper descriptors problem-dependent.³ Since the graph of a molecule itself contains limited information relative to the independent variables (descriptors) available, many of the descriptors in a set are highly correlated. Add to this the issue that there often is not one-to-one mapping between a descriptor and a

molecular property³ and the final QSR which is reported becomes researcher-dependent, though not necessarily of varying legitimacy.⁴

In a recent work Randić and Basak ask "Do we need additional descriptors although hundreds of molecular descriptors...are already available...? How can we establish whether the available molecular descriptors suffice for a complete characterization of molecules for QSPR and QSAR?"⁴ A step toward addressing these important, yet seemingly overlooked issues is in the generation of a finite set of descriptors that are not highly correlated and form a complete set from which all other descriptors can be calculated. In this work, we introduce the concept of a molecular signature and show how one can write many of the topological indices used in QSR from these signatures.

This paper sets up as follows. First we introduce the concept of the signature and show how it is calculated for a molecular graph. We then evaluate the computational complexity of storing and calculating atomic and molecular signatures. We also probe the relationship between molecular graph canonical representations and molecular signatures. The next section presents QSAR and QSPR studies using signature on two different problems, prediction of HIV-1 protease inhibition concentration, and Log P calculations for a large set of organic compounds. The signature results are compared with similar results obtained using 2D descriptors taken from the Molconn-Z commercial package. To explain their comparable performance, we demonstrate in the section that follows how popular 2D descriptors can be computed directly from signature without knowing the molecular graph

*Corresponding author phone: (925)294-1279; fax: (925)294-3020; e-mail: jfaulon@sandia.gov.

of the studied molecules. As a direct application of the relationships between signature and 2D descriptors, we next present a simple study showing the equivalence between a QSAR developed using four indices and that using four atomic signatures for the boiling point of 25 aliphatic hydrocarbons. Finally, we make concluding remarks and discuss future work involving signatures as molecular descriptors.

THE SIGNATURE DESCRIPTOR

The *signature* is a systematic codification system over an alphabet of atom types, describing the extended valence (i.e., neighborhood) of the atoms of a molecule. This concept, which was first presented and applied in the context of structural elucidation⁵ and recently defined for acyclic compounds and used in QSAR analysis,⁶ is further detailed and generalized in the present paper. Prior to introducing the signature descriptor, some terminology and notation have to be defined.

Molecular Graph. A molecule is represented by a graph $G = (V_G, E_G, C, c_G())$, where the elements of V_G are the atoms and the edges of E_G are the bonds. The atoms of a molecular graph are colored by C , a set of atom types, which, for instance, can be the set of elements of the periodic table or any set of atom types provided by a molecular force field. $c_G()$ is the function that associates an atom of G to an atom type. Every atom type has a valence, which is the number of covalent bonds that can be formed with this atom. A **molecular graph** is a graph representing a molecule which is not necessarily saturated. Formally, a molecular graph $G = (V_G, E_G, C, c_G())$ is an undirected graph colored with the function $c_G()$ over the elements of C verifying the equation

$$\forall x \in V_G, \deg(x) \leq \text{valence}(c_G(x)) \quad (1)$$

Signature of an Atom. Let G be a molecular graph and let x be a atom of G . The signature of height h of x , $^h\sigma_G(x)$, is a canonical representation of the subgraph of G containing all atoms that are at distance h from x . This canonical representation takes the form of a tree and is constructed following a five-step procedure detailed next and illustrated in Figure 1.

(1) A subgraph $^hG(x)$ is extracted from G containing all atoms and all bonds between these atoms that are at distance h from x .

(2) The vertices of $^hG(x)$ are labeled in a canonical order, atom x having label 1.

(3) A tree spanning all the edges of $^hG(x)$ is constructed. The root of the tree is atom x itself. The first layer of the tree is composed of the neighbors of x , and the second layer is composed of the neighbors of the vertices of the first layer except atom x . To each vertex added to the tree, one associates the color and the canonical label of the corresponding atom. The construction proceeds one layer at a time up to layer h . Assume the tree has been constructed up to layer $k < h$, layer $k+1$ is constructed considering each vertex y of layer k . Let z be a neighbor of y in G , vertex z and edge $[y,z]$ are added to layer $k+1$ if the edges $[y,z]$ or $[z,y]$ are not already present in the tree. The neighbors of y present in layer $k+1$ are sorted in decreasing lexicographic order using the colors and the canonical labels of the corresponding atoms. As shown in Figure 1 in the first layer, C,3 appears

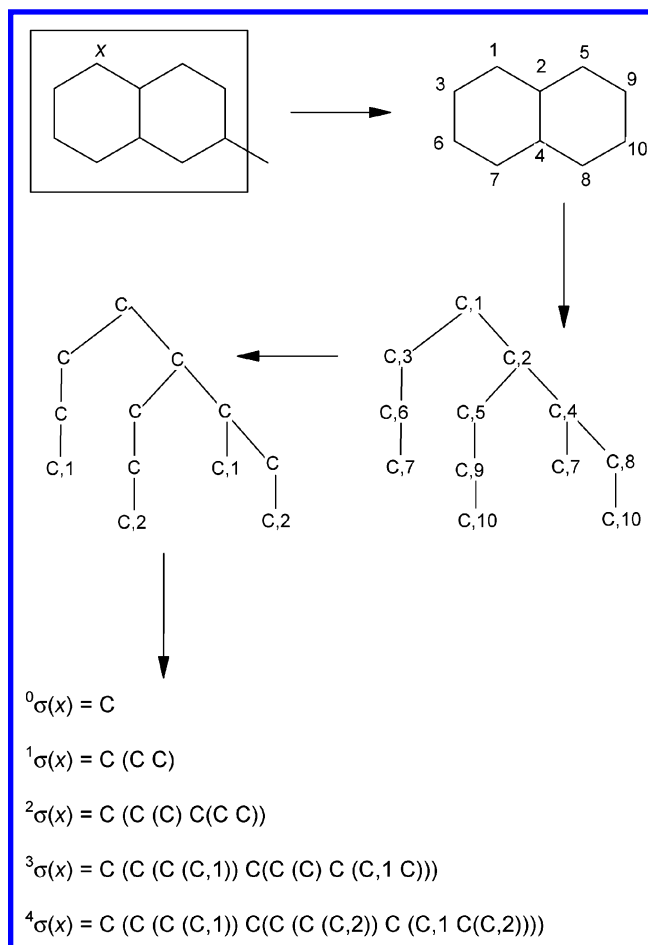


Figure 1. Atomic signature. The figure illustrates the five step procedure for computing the atomic signature of atom x in methyl-naphthene. (1) The subgraph containing all atoms at distance 4 from atom x is extracted. (2) The subgraph is canonized atom x having label 1. (3) A tree spanning all edges of the subgraph is constructed. (4) All labels appearing only once are removed, and the remaining labels are renumbered in the order they appear. (5) The signature is printed reading the tree in a depth-first order. Here we show atomic signatures from $h = 0$ to $h = 4$.

before C,2. Note that a given vertex z may appear several times in the tree (such as C,7 in Figure 1) since it can be the neighbor of several vertices present in the previous layer; however, as imposed by the construction procedure any given edge appears only one time.

(4) Once the tree has been constructed up to layer h , all canonical labels that appear only one time are removed. The remaining labels are renumbered in the order they appear while reading the tree in a depth-first order.

(5) The signature is written by reading the tree in a depth-first order and printing the character '(' each time an edge parent-child is read, the character ')' when the edge is read from child to parent and the vertex color followed by a label if the vertex appears several time in the tree.

Signature of a Molecule. According to the above definition the signature of an atom can be viewed as a string of characters over the alphabet C of atom types. Note that for a given height h , the list of all possible atomic signatures, although large, is of finite size. Consequently, any molecule of the chemical universe can be represented by its coordinates in a vectorial space where the base vectors are the distinct atomic signatures. We thus define the signature of a molecule

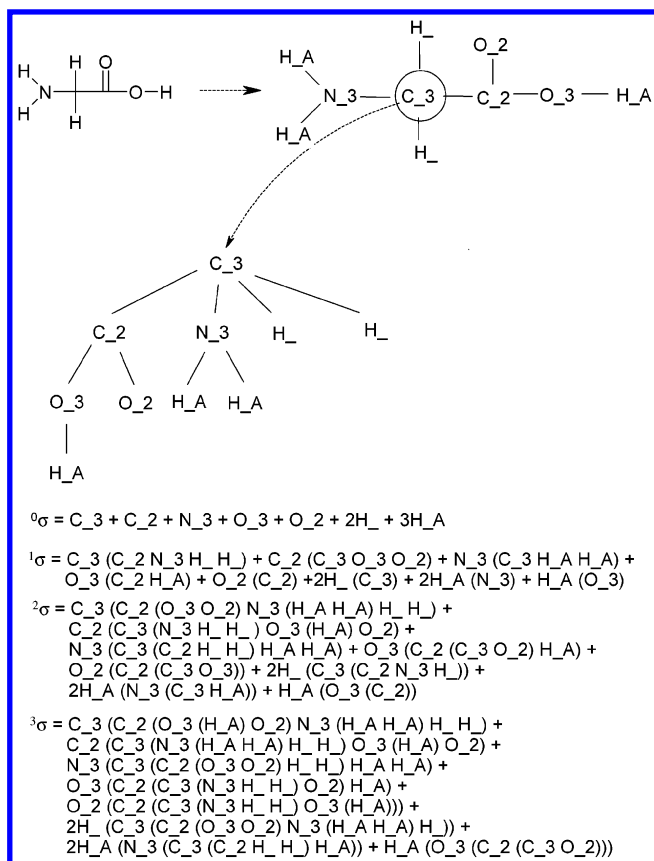


Figure 2. Molecular signature. The molecular graph of glycine colored using the Dreiding 2.21 force field.⁶² We show the signature tree of the atom colored C₃ and the molecular signature of glycine for heights 0, 1, 2, and 3. Here, the hierarchy of labels is C₃ > C₂ > N₃ > O₃ > O₂ > H₋ > H_A.

as the linear combination of its atomic signatures

$${}^h\sigma(G) = \sum_{x \in V_G} {}^h\sigma_G(x) = {}^h\alpha_G {}^h\Sigma \quad (2)$$

where ${}^h\Sigma$ is the basis set of all atomic signatures of height h , and ${}^h\alpha_G$ is the vector of occurrence number of atomic h -signatures of graph G . Examples of molecular signatures are given in Figure 2.

Signature of a Bond. Let $G = (V_G, E_G, C, c_G())$ be a molecular graph and let b be a bond/edge of E_G . Let $G-b = (V_G, E_G - \{b\}, C, c_G())$ be the molecular graph in which the bond b has been removed. The h -signature of b is defined as follows

$${}^h\sigma(b) = {}^h\sigma(G) - {}^h\sigma(G-b) \quad (3)$$

Signature of a Reaction. Let $B = (V_B, E_B, C, c_B())$ and $E = (V_E, E_E, C, c_E())$ be two molecular graphs representing the reactants and products of the reaction $R: B \rightarrow E$. Note that signatures can be computed on graphs that are not necessarily connected, hence B and E can both be composed of several molecules. The h -signature of reaction R is given by the equation

$${}^h\sigma(R) = {}^h\sigma(E) - {}^h\sigma(B) \quad (4)$$

SIGNATURE COMPUTATIONAL SPACE AND TIME EFFICIENCY

In the following G is a molecular graph of n atoms, x is an arbitrary atom of G , and h is the height of the signatures

considered. Let D be the diameter of G , that is the largest distance between all pairs of atoms in G . Note that a priori D may be equal to n . ${}^hG(x)$ is the subgraph composed of all vertices of G that are at most h -distant from x , ${}^h\sigma_G(x)$ is the atomic h -signature of x , and ${}^h\sigma(G)$ is the molecular h -signature of G .

Proposition 1. The number of characters used to print ${}^h\sigma_G(x)$ scales $O(n)$.

The number of characters used to print ${}^h\sigma_G(x)$ is linearly proportional to the number of edges of ${}^hG(x)$ since as described in the previous section every edge of ${}^hG(x)$ appears only once in ${}^h\sigma_G(x)$. The number of edges in ${}^hG(x)$ is linearly proportional to the number of atoms of ${}^hG(x)$ since molecules are bounded valence graphs. Hence, the number of characters in ${}^h\sigma_G(x)$ scales $O(n)$.

Proposition 2. Let b be a bond of G and let R be the reaction $R: B \rightarrow E$. The numbers of characters used to print ${}^h\sigma(G)$, ${}^h\sigma(b)$, and ${}^h\sigma(R)$ scale $O(n^2)$.

It is trivial to prove that ${}^h\sigma(G)$ scales $O(n^2)$, since G may comprise up to n atoms having different atomic signatures each scaling $O(n)$. From eqs 3 and 4, it is obvious that the space taken by ${}^h\sigma(b)$ for any bond b of G scales at most $O(n^2)$ and the space occupied by the signature ${}^h\sigma(R)$ of the reaction $R: B \rightarrow E$ scales $O(n^2)$, n being the sum of the vertices of the molecular graphs B and E .

Proposition 3. For all $h \geq D+1$, ${}^h\sigma(G) = {}^{D+1}\sigma(G)$.

Let x be an arbitrary atom of G , we prove ${}^h\sigma_G(x) = {}^{D+1}\sigma_G(x)$ for all $h \geq D+1$. Let us recall that in ${}^h\sigma_G(x)$ each bond of the subgraph ${}^hG(x)$ appears only once. Thus, the maximum height h above which all signatures will be identical is the one that covers all the bonds of G . All atoms y of ${}^hG(x)$ are at most at distance D from x . Thus, in ${}^h\sigma_G(x)$, every atom of ${}^hG(x)$ first appears in a layer $h \leq D$. First, assume y appears in a layer $h < D$, then all neighbors z of y should appear at layer D if not earlier, and, consequently, all bonds $[y, z]$ or $[z, y]$ are listed in ${}^D\sigma_G(x)$. Now, assume y appears for the first time in layer D , that is, y is at distance D from x . Because no atom is at a distance greater than D from x , and because y first appears at layer D , all neighbors z of y either first appear at layer $D-1$ or layer D . If atom z first appears at layer $D-1$, then bond $[z, y]$ is listed in ${}^D\sigma_G(x)$. The only remaining case is when both y and z appear for the first time at layer D . In this instance all bonds to which y or z belong will be listed in developing the signature one additional layer, that is $D+1$. Since we have examined all cases covering all atoms of G , all bonds of G are listed in ${}^{D+1}\sigma_G(x)$.

Proposition 4. Once initialized, the adjacency matrix of G can be constructed in $O(n)$ steps from ${}^{D+1}\sigma(G)$.

We assume the adjacency matrix of G , M_G , has been initialized with zeros. Note that this task requires $O(n^2)$ steps since the matrix has n rows and n columns. Let x be an arbitrary atom of G , we prove that M_G can be filled reading ${}^{D+1}\sigma_G(x)$ three times. Since ${}^{D+1}\sigma_G(x)$ contains $O(n)$ characters, M_G can be filled in $O(n)$ steps. ${}^{D+1}\sigma_G(x)$ is read a first time to determine the maximum number following any comma in the signature string. We call this number the initial label number. If no number appears in the signature string, the initial label number is initialized to zero. We also monitor the current label number, with initial value equal to the initial label number augmented by one. ${}^{D+1}\sigma_G(x)$ is read again, but this time for each atom type read that is not followed by

a comma, one inserts after the atom type a comma and the current atom number. For each insertion the current atom number is incremented by one. The results of this procedure is that every atom is now labeled with a number. Finally $^{D+1}\sigma_G(x)$ is read a third time while maintaining a last in/first out stack of atom labels and two labels x and y . The first character of the signature string (i.e., an atom type), the following comma, and the number that follows are initially read. This first number is added to the stack of labels. Then the string is read character by character with the following rules. If the current character read is '(', then let x be value on the top of the stack. If the current character read is ')', then the top element of the stack is popped, and x takes the value of the new top. For any other character read (e.g., atom type) one reads the comma and the atom label that follows. Variable y takes the value of the newly read atom label and is added to the stack. The adjacency matrix is updated by setting $M_G[x,y]$ and $M_G[y,x]$ to 1.

Proposition 5. $^{D+1}\sigma(G)$ is a canonical representation of G .

We already know from proposition 4 that graph G is fully described by $^{D+1}\sigma(G)$; it remains to be proven that two isomorphic graphs G_1 and G_2 have identical height $D+1$ signatures. Let π be an isomorphism between G_1 and G_2 and let x_1 be an atom of G_1 and x_2 an atom of G_2 such that $x_2 = \pi(x_1)$. $^{D+1}G_1(x_1)$ is isomorphic to $^{D+1}G_2(x_2)$ since G_1 and G_2 are isomorphic and x_1 and x_2 are mapped by π . Because $^{D+1}G_1(x_1)$ and $^{D+1}G_2(x_2)$ are isomorphic, they have the same canonical representation and consequently the same signature. Since each atom x_1 of G_1 is mapped by π to an atom x_2 of G_2 , to each atom of G_1 corresponds an atom of G_2 having the same height $D+1$ signature, therefore, G_1 and G_2 have the same height $D+1$ signature.

Proposition 6. Let x_1 and x_2 be two vertices of G , if $^{D+1}\sigma_G(x_1) = ^{D+1}\sigma_G(x_2)$, then x_1 and x_2 are automorphic in G (e.g., symmetrical).

Proof. We create two identical copies of G , G_1 , and G_2 . Because $^{D+1}\sigma_G(x_1) = ^{D+1}\sigma_G(x_2)$, $^{D+1}G_1(x_1)$, and $^{D+1}G_2(x_2)$ have the same canonical representation and consequently an isomorphism π can be found between $^{D+1}G_1(x_1)$ and $^{D+1}G_2(x_2)$ such that $x_2 = \pi(x_1)$. Since $^{D+1}G_1(x_1) = G_1$, $^{D+1}G_2(x_2) = G_2$ and G_1 and G_2 are identical, π is an automorphism of G .

Proposition 7. $^h\sigma(G)$ can be computed with a number of steps bounded by a polynomial function in n .

A molecular signature may be composed of n different atomic signatures, hence the cost of computing a molecular signature is n times the cost of computing an atomic signature. Computing an atomic signature involves five steps as described in the previous section, and their computational complexity is evaluated next.

(1) $^hG(x)$ *Computation.* This is achieved by removing from G all atoms that are at a distance greater than h from x . This task can be performed in $O(n)$ time following a classical distance in graph measurement algorithm.⁷

(2) $^hG(x)$ *Canonization.* Up to date, graph canonization cannot be performed in polynomial time for general graphs. But because molecules are bounded valence graph, it has been shown previously that all molecular graphs can indeed be canonized in polynomial time.⁸ For instance, Babai and Luks canonization algorithm can be used with an expected running time of $O(n^{k \log k})$ for a graph of maximum valence $k+1$.⁹ Note that most organic compounds have a maximum

valence 4, leading to a complexity of $O(n^{3 \log 3})$, furthermore all molecules have a maximum valence that is bounded by a constant independent of n , thus resulting in a polynomial time complexity for canonization. In the present paper we are making use of McKay's Nauty program, which is believed to be the fastest practical canonization algorithm.¹⁰

(3) *Edge Spanning Tree Construction.* This tree spans all the edges of $^hG(x)$ reading all vertices of $^hG(x)$ starting from x , proceeding with the neighbors of x , the neighbors of the neighbors, and so forth until all edges have been visited. Consequently, the tree is constructed in $O(n)$ step since the number of edges scales $O(n)$. At each step of the construction procedure one maintains a list of all edge visited. This list can contains $O(n)$ element. The total cost of the tree construction is therefore $O(n^2)$.

(4) *Renumbering of the Labels.* This renumbering is performed reading the tree in first-depth order with a cost linearly proportional to the tree size, $O(n)$.

(5) *Signature Writing.* The signature is written reading the tree in first-depth order with a cost $O(n)$.

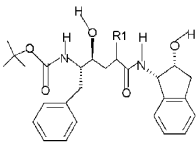



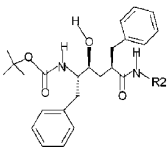



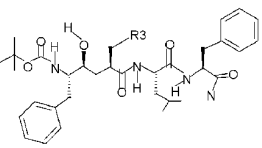


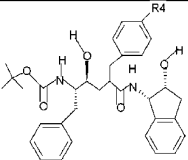



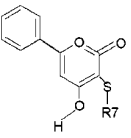


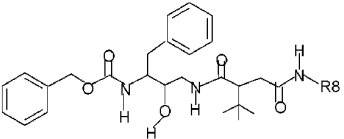


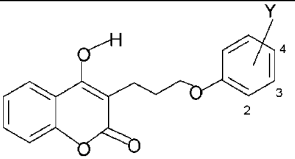



USING SIGNATURE IN QSAR AND QSPR

The logical next step in this study is to explore the utility and robustness of signature as descriptors in quantitative structure activity/property relationships. To this end we have chosen two data sets to explore the various features of signature for QSAR/QSPR modeling. The first data set comes from a biological source and measures the activity for 121 compounds used as HIV-1 protease inhibitors. The second, a much larger set, is for the octanol/water partition coefficient ($\log P$) of 12865 compounds. We compare the results of the models developed and some of the features of the descriptors (i.e., orthogonality and stability) to that from a commercially available program, Molconn-Z, from eduSoft, LC,¹¹ to provide some perspective on the results.

Molconn-Z (v 3.50) has over 200 descriptors to choose from for use in QSAR/QSPRs. These descriptors include connectivity indices, shape indices, information indices, fragment counts, and electrotopological states among others.^{11,12} When using atomic signatures as descriptors, the occurrence number of a particular atomic signature is used as the descriptor value. However, the overall number of unique atomic signatures in the data set is not known a priori and is a function of the height used as well as the size of the training set. In the following QSAR and QSPR studies we will use atomic signatures of height 0 (0-signature), height 1 (1-signature), and height 2 (2-signature).

For both the Molconn-Z models and the signature-based models, the procedure to determine the overall number of descriptors available involved determining whether two or more descriptor vectors (the elements of the vectors are the descriptor values for each compound in the training set) were equivalent (i.e.: had a pair-pair correlation coefficient of unity). If this occurred, the descriptor that had the lowest overall absolute pair-pair correlation coefficient with the rest of the descriptors left in the training set was kept and the other descriptor(s) discarded. The overall number of descriptors used in each part of this study will be provided in the relevant section. All models used were of the linear form, and parametrization of the models was performed using the forward-stepping regression technique of multiple linear

Table 1. Scaffolds of the HIV-1 Protease Inhibitors Used in This Study

Scaffold	R1	pIC ₅₀	Source
		9.60	56
		9.22	
		8.27	
Scaffold	R2	pIC ₅₀	Source
		7.41	56
		8.02	
		4.52	
Scaffold	R3	pIC ₅₀	Source
		9.36	57
		8.92	
	H	8.22	
Scaffold	R4	pIC ₅₀	Source
		9.70	58
		10.04	
		8.69	
Scaffold	R7	pIC ₅₀	Source
		5.52	59
		5.89	
Scaffold	R8	pIC ₅₀	Source
		6.8	60
		7.8	
Scaffold	Y	pIC ₅₀	Source
	(2) 	4.3	61
	(3) 	6.1	
	(4) 	5.1	

regression (MLR).¹³ Briefly, forward-stepping regression is a technique used for MLR wherein independent variables (here descriptors) are added to the model in a systematic way. In this work, we insert variables into the model by choosing that descriptor with the largest partial correlation coefficient of the variables not currently in the model.

Before examining the quality of the fit produced from any of the models, it is of interest to look at some measure of orthogonality within the descriptor sets. If one uses orthogonal descriptors, the benefit of stability during parametrization occurs for the regression coefficients. If the descriptor vectors are all orthogonal to each other, then the individual regression

Table 2. A Complete List of the Compounds Used in the Training Set (121) and Test Set (9) Used for the QSAR on the HIV-1 Protease Inhibitors

source	compound number (as per reference)		test set	pIC ₅₀ (test set)
	training set			
56	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,30,31,33,34,35,36,37,38,39,40,41,44,45,46,47,48,49,50		29	7.393
57	1,7,9,10,11,12,16,19,20,24,25,26,35,36,37,39,40,41,42,43,44,45,47,49,50,51		34	9.658
61	1,12,13,16,18,19,21,22,23,24,25,26,28,29,30,31,33,34,35,36,37,38		46	8.854
58	19,22,23,24,29,31,32,33,34,36,37,38,39,40,43,44,45		27	5.638
59	1,2,3,5,6,7		30	9.260
60	9,10,12,13,14,15		35	8.745
			4	5.851
			11	7.824

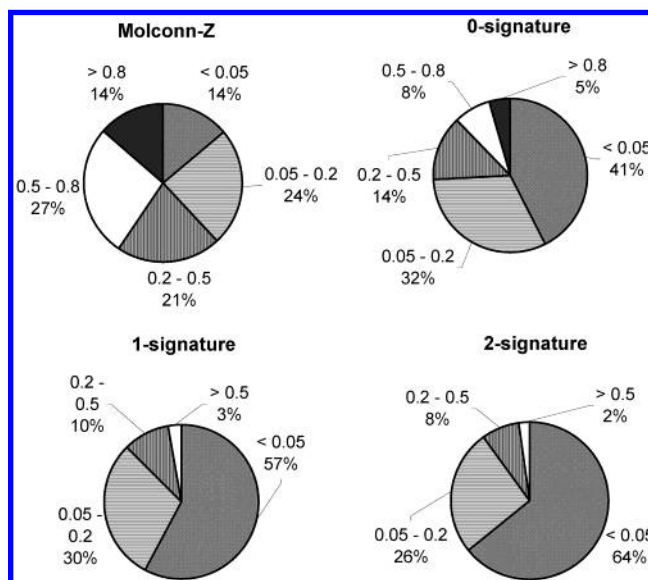
coefficients will remain unchanged regardless of whether other descriptors are added to or removed from the overall descriptor set. Orthogonal descriptors can be made from the original descriptor set through a variety of means, such as Principal Component Analysis¹⁴ or Dominant Component Analysis,¹⁵ but in each case the new orthogonal descriptors are linear combinations of the original descriptors. Although the benefit of stability is achieved through making descriptors orthogonal, ambiguity may arise if one tries to attribute physical meaning to the presence (or absence) of these orthogonal descriptors in the QSAR/QSPR.¹⁶ We will examine the extent of the orthogonality of the original descriptor sets for all of the models using both data sets. We must note, however, that an orthogonal set of descriptors does not imply that those descriptors correlate well with the dependent data.

QSAR for HIV-1 Protease Inhibitors. We report below a QSAR on HIV-1 protease inhibitors. While a QSAR has already been published using this data set (Tables 1 and 2),⁶ we have since then updated the definition of the signature descriptor. In our previous paper, bonds could be repeated several times in the signature notation, while in the present work bonds appear only once. Note that the updated definition allows one to store signatures efficiently (cf. propositions 1 and 2). The overall results are similar, but there are some differences and, in particular, we do not have the same number of descriptors because of the above different definitions. Additionally, the number of overall Molconn-Z descriptors between the two studies for this data set (previous work – 163; this work – 152) are different because some of the Molconn-Z descriptors were redundant and, hence, removed from the overall number upon which the MLR was performed.

Training Set/Test Set. From six literature sources a total of 121 HIV-1 protease inhibitors were isolated as a training set. The experimental activity reported was in units of pIC₅₀ (pIC₅₀ = 9 – log IC₅₀) and spanned 7 orders of magnitude. Representative compounds in this training set are given in Table 1 with a complete list provided in Table 2. Nine compounds, at least one from each of the six literature sources, were chosen as test set compounds. These compounds were not used in creating the models and were used only to evaluate the predictive ability of the models against compounds not used in the parameter fitting. A list of the compounds used in the test set are provided in Table 2. After

Table 3. Total Available Number of Descriptors for Each Model Using the HIV-1 Data Set

model	no. of available descriptors for HIV-1 data set	model	no. of available descriptors for HIV-1 data set
Molconn-Z	152	1-signatures	121
0-signatures	12	2-signatures	290

**Figure 3.** The pair-wise correlation coefficients (R_{ij}^2) for each of the four QSARs created on the HIV-1 protease inhibitors data set.**Table 4.** Statistics for the Models Created for the HIV-1 Protease Inhibitors Data Set

model	no. of descriptors	training set ($n = 121$)		test set ($n = 9$) s
		r^2	s	
Molconn-Z	152	~ 1.0	3.038E-6	145.1
Molconn-Z ^a	31	0.951	0.410	0.575
0-signature	12	0.166	1.762	1.435
0-signature ^a	8	0.161	1.767	1.414
1-signature	121	0.991	0.186	701.9
1-signature ^a	18	0.915	0.503	0.633
2-signature	290	~ 1.0	3.819E-7	355.8
2-signature ^a	37	0.988	0.214	0.766

^a Number of descriptors giving the smallest test set s values obtained using the forward-stepping algorithm mentioned in the text.

using the methods provided in the previous section we arrived at the final number of descriptors used in the parameter fitting. This is given in Table 3.

Orthogonality. From Figure 3 we see that the atomic signatures, by percentage, are less intercorrelated than those from Molconn-Z. Additionally, as the height of the atomic signatures increases, the amount of correlation among the descriptors decreases. For the atomic height-2 signatures, 90% of the pair-wise correlations have a low correlation level (<0.2).

QSAR Statistics. Using forward-stepping regression, we parametrize the models relative to the training set and report those results in Table 4. For the four models, we first examine the statistics of the fit using all available parameters, per model. It is seen that the models generated from Molconn-Z and from the atomic height-2 signatures can correlate the training set almost perfectly. This is not surprising given the fact that the number of parameters, in both cases, are greater

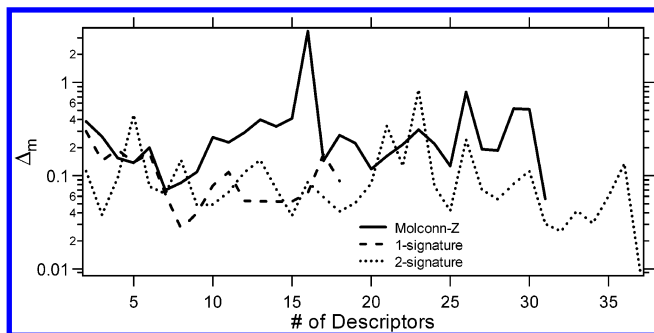


Figure 4. A comparison of the stability of the MLR coefficients for three of the QSARs created on the HIV-1 protease inhibitors data set.

than the number observations. Using a model with such a large parameter-to-observation ratio results in overfitting, and, accordingly, the model is neither robust nor predictive. This is observed by looking at the test set s value (root-mean-square error) for both models when using all available parameters. Then, for each model, the number of parameters (and their relevant statistics) needed to generate the smallest test set s value is also shown. These numbers of parameters or descriptors (31 for Molconn-Z, 8 for 0-signature, 18 for 1-signature, and 37 for 2-signature) were computed using the forward-stepping regression mentioned earlier. For this data set, the models formulated from atomic height-1 and height-2 signatures compare well with that developed from the Molconn-Z descriptors, though the latter does predict the test set activities slightly better than the former two models.

Previously we mentioned the importance of stability of the coefficients in models formed by MLR and the utility of orthogonal sets of descriptors. To examine this feature in the models created, we evaluate the *change* in the values of the coefficients from step to step in a regression. This deviation metric we give the symbol Δ and define it as

$$\Delta = \frac{1}{m-1} \sum_{i=1}^{m-1} \left| \frac{\alpha_i - \alpha_i^*}{\alpha_i^*} \right| \quad (5)$$

where m is the number of descriptors in the model, and α_i is the regression coefficient for the i th descriptor with the superscript “*” indicating the value of regression coefficient for that descriptor in the *previous* step. Note that if a descriptor is being added to the model in that step (i.e.: the m th descriptor), it is not considered in calculating Δ .

From Figure 4 we see that, overall, the Molconn-Z model results in larger values of Δ as a function of the number of descriptors added to the model relative to the atomic height 1 and height 2-signatures. The larger the value of Δ , the more the values of the coefficients from the regression change from step-to-step. However, even though the model from the height-2 signatures had 90% of the pair-pair correlations with less than a 20% correlation, this apparently is far away from a completely orthogonal descriptor set as seen in by the delta values for the atomic signatures in Figure 4.

QSPR for Log P . Training Set/Test Set. From the Syracuse Research Corporation¹⁷ we obtained a large database of octanol/water partition coefficients (Log P) for a myriad of compounds. From this large database, we segregated 12865 compounds to act as a training set and 123

Table 5. Total Available Number of Descriptors for Each Model Using the Log P Data Set

model	no. of available descriptors for Log P data set	model	no. of available descriptors for Log P data set
Molconn-Z	192	1-signatures	1161
0-signatures	16		

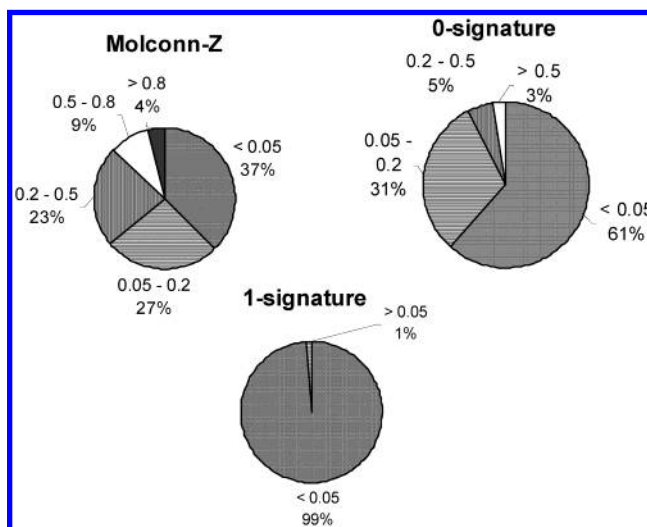


Figure 5. The pair-pair correlation coefficients (R_{ij}^2) for each of the three QSPRs created on the Log P data set.

Table 6. Statistics for the Models Created for the Log P Data Set

model	no. of descriptors	training set ($n = 12865$)		test set ($n = 123$) s
		r^2	s	
Molconn-Z	192	0.881	0.629	0.722
Molconn-Z ^a	123	0.878	0.638	0.709
0-signature	16	0.635	1.103	1.124
1-signature	1161	0.922	0.508	0.770
1-signature ^a	265	0.914	0.536	0.632

^a Number of descriptors giving the smallest test set s values obtained using the forward-stepping algorithm mentioned in the text.

compounds to act as a test set. After using the methods provided in the previous section we arrived at the final number of descriptors used in the parameters fitting. This is given in Table 5.

Orthogonality. The pair-pair correlation coefficients for the Molconn-Z descriptors as well as the atomic height 0 and height 1-signatures are shown in Figure 5. Similar trends as that from the activity data set are seen in the Log P data set. Note the very small pair-pair correlation coefficients for the 1-signatures here with 99% of the pairs having less than 5% correlation. For a QSPR built using the atomic 1-signatures, we should find small Δ values and, accordingly, a very stable model.

QSPR Statistics. Using forward-stepping regression, we parametrize the models relative to the training set and report those results in Table 6. For the three models, we examine the statistics of the fit using all available parameters, per model. Since the number of observations is large (12865) relative to the maximum number of descriptors available, overfitting is not an issue for height-1 signatures. Once again, we show the results for a QSPR created using all the available descriptors and, also, the model corresponding to the smallest error in the prediction of the test set values.

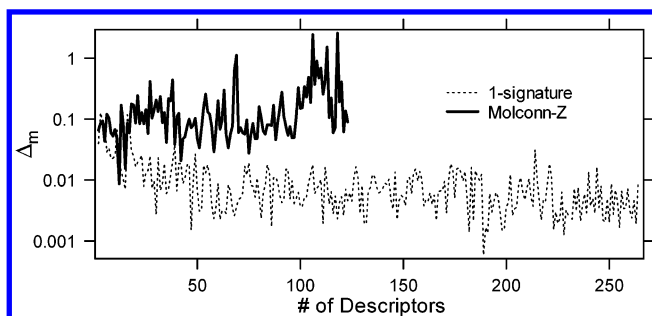


Figure 6. A comparison of the stability of the coefficients for two of the QSPRs created on the Log P data set.

Note that the smallest test set error for the atomic height 0-signatures occurs when all 16 signatures are used in the model.

For this data set, the models formulated from atomic height-1 signatures compare well with that developed from the Molconn-Z descriptors, though this time the former predicts the test set log P values better than that obtained from the Molconn-Z descriptors.

The stability of the QSPR created from the atomic height-1 signatures, as implied from the pair–pair correlation plot, is readily seen from Figure 6 that plots the Δ values as a function of the number of descriptors used in the model. The QSPR created from the 1-signatures has a Δ value an order of magnitude less than that from the Molconn-Z QSPR.

COMPUTING TOPOLOGICAL INDICES FROM SIGNATURE

In the previous section, we have shown that signature produces meaningful QSAR/QSPRs. To this end, we created one QSAR and one QSPR using atomic signatures on different types of data sets—a relatively small set for biological activity (QSAR) and a relatively large set for Log P (QSPR). To provide perspective on our results, we compared these models developed using signature to those developed using popular molecular descriptors from the Molconn-Z package. Our results indicate that signature performs comparable to the Molconn-Z model for both data sets. This result may appear surprising since Molconn-Z encompasses hundreds of descriptors characterizing many aspects of molecular topology, including connectivity, shape, electronic configuration, and information content, while signature only capture the extended valence sequence of a molecule. The intent of this section is to provide an explanation of this somewhat surprising result by demonstrating how Molconn-Z descriptors as well as other recent and popular TIs used in QSR applications can be computed from signature. Since it is not practical to screen all possible topological indices, we chose a sample belonging to different descriptor families. Note that some of the relationships we provide next are valid only for alkanes and cycloalkanes, such as the relationships given for the Wiener index and the Balaban J index. Other relationships are valid for compounds with heteroelements such as those derived for the Kier and Hall alpha-shape indices or the total topological index.

In the following we consider a hydrogen suppressed covalent molecule G , of n atoms and diameter D . Molecule G has an unknown molecular graph but known molecular

Table 7. List of the TIs Computable from the Molecular h -Signature

h	TIs
0	n : number of atoms, MF : molecular formula, MW : molecular weight, fragments of one atom
1	MF and MW for hydrogen suppressed graphs, m : number of bonds, μ : cyclomatic number, I : intrinsic state, mwc_1 : molecular walk count of length 1, $^1\kappa$, $^1\kappa_\alpha$: shape indices of length 1, $^0\chi$, $^0\chi_\alpha$: connectivity indices of length 0, p_1/w_1 : Randić shape index of length 1, fragments of diameter 1
2	mwc_2 , $^2\kappa$, $^2\kappa_\alpha$, F : Platt number, $^1\chi$, $^1\chi_\alpha$, p_1/w_1 , s_1 : shell index of length 1, fragments of diameter 2
1	mwc_l , $^l\kappa$, $^l\kappa_\alpha$, $^{l-1}\chi$, $^{l-1}\chi_\alpha$, s_{l-1} , p_l/w_l , fragments of diameter l
$l = D$	W , M , WW , H : Wiener-type indices, Balaban J index, I_E^P : information index
$l = D+1$	S : E-state index, τ total topological index for hydrogen suppressed graphs, twc : total walk count, ω , $\omega\omega$: detour-type indices, Wt : all-path based Wiener index

^a D is the diameter of the molecular graph, and n is the number of atoms.

signatures up to height h , $^h\sigma(G) = ^h\alpha_G \cdot ^h\Sigma$, where $^h\Sigma$ is the basis set of all atomic signatures of height h , and $^h\alpha_G$ is the vector of occurrence number of atomic h -signatures of molecule G over the basis set. As shown in Proposition 4, from the height $h = D+1$ molecular signature, the adjacency matrix of the entire molecular graph itself can be computed. Consequently, any TI can be computed from a molecular signature of height $D+1$, since any TI can be computed from an adjacency matrix. A less trivial result is the minimum height for which a given TI can be computed. Table 7 summarizes our results for a subset of popular TIs used in QSR. As detailed next, for all chosen TI, there exist a height $h \leq D+1$ for which the index can be expressed as a dot product between two vectors, $^h\alpha_G$, the vector of occurrence number of atomic h -signatures of molecule G , and $TI(\text{root}(^h\Sigma))$, the vector of TI values computed for each root of the atomic h -signatures of $^h\Sigma$:

$$TI(G) = \text{constant } ^h\alpha_G \cdot TI(\text{root}(^h\Sigma)) \quad (6)$$

We arbitrarily separate topological indices into seven categories: simple indices, walk-based, path-based, distance-based, hybrids (a combination of two or more), information-based and fragment-based. For a detailed coverage, see refs 18 and 19. We use the term “arbitrary” since some descriptors can be equivalently characterized by more than one category, depending on how the operator is formulated. We will briefly introduce each type here, highlight some of the popular descriptors, and provide the relationships between the descriptors and signatures.

(1) *Simple Indices.* Simple indices are parameters such as the number of atoms, bonds, the molecular formula, the molecular weight, and the cyclomatic number. The number of atoms, $n(G)$, is simply the sum of the coordinates of the molecular 0-signature, $n(G) = ^0\alpha \cdot 1$ where 1 is a unit vector. The number of bonds, $m(G)$, is computed from the number of bonds of the atomic base vectors 1-signatures, $m(G) = \frac{1}{2} {}^1\alpha \cdot m(\text{root}(^1\Sigma))$ where $m(\text{root}(^1\Sigma))$ is a vector counting the number of bonds of each root of $^1\Sigma$. The cyclomatic number is $\mu = m - n + 1$. The molecular formula, MF , is derived from the atom type of the root of the 0-signatures.

For hydrogen-suppressed graphs, the number of hydrogens is computed from the number of bonds and the atom types in order to saturate the molecule. The molecular weight, MW , is computed from the molecular formula. MF and MW are thus computed from the molecular 1-signature.

(2) *Walk-Based Indices*. A walk on a graph is an alternating sequence of vertices and edges wherein the sequence may include vertices and/or edges more than once. Though walk-based indices do not have a rich history that demonstrates their use in QSRs, recent work by Rucker looking at atomic, molecular, and total walk counts on a graph may change that view owing to the simplicity of calculating these indices.^{20–23} The atomic walk count of length k , awc_k , of any atom can be computed considering all atoms that are at most k distant from that atom. Thus, the atomic walk count of an atom is the atomic walk count of the root of its height k signature. Furthermore, note that any height $h \geq k$ signature gives the same atomic walk count for its root than the height k signature. The molecular walk count of length k , mw_c_k , is the sum of awc_k taken over all atoms, and $mw_c_k(G) = {}^k\alpha \cdot mw_c_k(\text{root}({}^k\Sigma))$. Summation of all mw_c_k with length from 1 to $n-1$ divided by 2 gives the total walk count, twc . This summation requires to compile signatures up to height $L = \text{MIN}(n-1, D+1)$, since according to Proposition 3, ${}^h\sigma(G) = {}^{D+1}\sigma(G)$ for all $h \geq D+1$. Hence,

$$twc(G) = 1/2 {}^L\alpha_G \cdot \left[\sum_{k=1}^{n-1} mw_c_k(\text{root}({}^L\Sigma)) \right] = 1/2 {}^{D+1}\alpha_G \cdot \left[\sum_{k=1}^{n-1} mw_c_k(\text{root}({}^{D+1}\Sigma)) \right]$$

(3) *Path-Based Indices*. Path-based indices are a very-popular and well-studied class of descriptors. A walk is a path if all its vertices are distinct. In addition to the count of paths of various lengths, modifications to path-based descriptors have been made to incorporate the local environment of the vertices in the path. For example, the connectivity index (denoted χ), introduced by Randić²⁴ nearly 30 years ago, weights a particular path by the product of the square root of the degree of each vertex. Various perturbations of the connectivity index have been introduced in the intervening years and include an accounting for the valence electrons at a node,²⁵ to a variable index^{15,26} that can change during the parameter regression. Other path-based connectivity indices that purport to discriminate molecular shape are the various κ indices of Kier.^{27,28} These descriptors encode a measure of cyclic nature (first-order index), spatial density of atoms (second-order index), and centrality of branching (third-order index).²⁹

The ${}^l\kappa$ shape indices require computation of the total number of atoms and, lP , the total number of paths of length l . The number of atoms is derived from the 0-signature of the molecule. Evidently the number of l -path for any atom can be computed considering all atoms that are at most l -distance from that atom. Hence, the number of l -paths of an atom is the number of l -paths computed for the root of any height $h \geq l$ signature of that atom. The total number of l -path of a molecule G is computed from any molecular h -signature with $h \geq l$: ${}^lP(G) = 1/2 {}^h\alpha \cdot {}^lP(\text{root}({}^h\Sigma))$. The modified shape indices ${}^l\kappa_\alpha$ encode atom types into the α -valence coefficients. Like the molecular formula, these

coefficients can be derived from the 0-signature of the molecule. Therefore both ${}^l\kappa$ and ${}^l\kappa_\alpha$ can be computed from the molecular l -signature. Derived from the shape indices, the flexibility index is the product of ${}^1\kappa_\alpha$ and ${}^2\kappa_\alpha$ divided by the number of atoms. This index can thus be computed from the molecular 2-signature.

The popular connectivity index, ${}^l\chi$, is computed by compiling the atom degrees along all the paths of length l in the studied molecule. ${}^l\chi$ is given by the formula ${}^l\chi(G) = \sum_{\text{paths of } G} [\deg(x_l)\deg(x_{l-1})\cdots\deg(x_0)]^{-1/2}$. As with lP , for any given atom, ${}^l\chi$ is computed by considering all the paths of length l that start at the root of its atomic signature. However, for hydrogen suppressed graphs, the signature height h must be greater than l , the length of the path, since the degrees of the vertices l distant from the root are needed, thus for any $h \geq l+1$, ${}^l\chi(G) = 1/2 {}^h\alpha \cdot {}^l\chi(\text{root}({}^h\Sigma))$. The valence connectivity index, ${}^h\chi^v$, is computed by replacing the degree function with a $\delta()$ function, which is fully defined from atom degrees and the atom types. Thus, both ${}^l\chi$ and ${}^h\chi^v$ can be computed from the $l+1$ -signature of the molecule.

Recently, Randić introduced a new descriptor, s_l , based on the concept of neighbor shells.³⁰ The difference between paths and shells is that instead of counting for each atom the number of neighbors at length l , one adds the degrees of the neighbors that are at the given length. As with ${}^l\chi$ this descriptor can easily be computed from h -signature for any $h \geq l+1$: ${}^l s(G) = 1/2 {}^h\alpha \cdot {}^l s(\text{root}({}^h\Sigma))$.

The Platt index, F ,³¹ is equal to the total sum of the edge-degrees of a molecular graph. The index can be reformulated in terms of vertex degrees, $F(G) = \sum_{\text{bond } [x,y] \text{ in } G} [\deg(x) + \deg(y) - 2]$. Like the ${}^l\chi$ connectivity index, the Platt index can be computed from molecular 2-signature, $F(G) = 1/2 {}^2\alpha \cdot F(\text{root}({}^2\Sigma))$.

(4) *Distance-Based Indices*. The distance matrix, from which distance-based descriptors are determined, is a symmetrical matrix describing the shortest path between two vertices. The first distance-based index was the Wiener Number^{32,33} introduced more than a half century ago, which sums the shortest path between all atoms. Other distance-based indices exist which attempt to provide an averaging or weighting to the sum in an attempt to increase its discriminatory nature, such as the Balaban J Index,³⁴ overall Wiener index,³⁵ detour index,³⁶ hyper-detour index,^{36,37} Pasareti index,³⁸ Verhalom index,^{38,39} reversed Wiener Index,⁴⁰ and Harary Index,⁴¹ for example. For a recent review, see ref 42.

The Wiener index, W , is defined as half of the off-diagonal elements of the molecular distance matrix. It can also be expressed as the dot product between the path-distance vector, P_D , and the path-length vector, $L = (1, \dots, D)$, where D is the diameter of molecular graph. The path-distance vector for the graph is half of the sum of the path-distance vectors of the atoms, $P_D(G) = 1/2 \sum P_D(x) \cdot P_D(x)$ can evidently be computed from the height D signature of atom x , and $P_D(G) = 1/2 {}^D\alpha \cdot P_D(\text{root}({}^D\Sigma))$. Therefore, W can be computed from the molecular D -signature, $W(G) = 1/2 {}^D\alpha \cdot W(\text{root}({}^D\Sigma))$ where $W(\text{root}({}^D\Sigma)) = (P_D(\text{root}({}^D\Sigma)), \dots, P_D(\text{root}({}^D\Sigma))) \cdot (1, \dots, D)$.

The reverse Wiener index, $M = P_D \cdot L^*$, where $L^* = (D, \dots, 1)$, the Hyper Wiener index, $WW = P_D \cdot (L+L^*)$, and the Harary index, $H = P_D \cdot L'$, where $L' = (1/1, 1/2, \dots, 1/D)$, are also computable from the molecular D -signature. The Wiener product, π , is equal to the product of the shortest

distances between all pairs of atoms

$$\pi = \prod_{x < y} d(x, y) = \left(\prod_{x, y} d(x, y) \right)^{-1/2} = \left(\prod_{i=1}^n ({}^iL)^{iP_D} \right)^{-1/2}$$

where iL , $({}^iP_D)$ is the i th coordinate of vector L (P_D). Thus, the Wiener product can be computed from the molecular D -signature.

Balaban J Index. The index is defined as follows: $J = m/(\mu+1)\Sigma_{\{x,y\}}(d_x d_y)^{-1/2}$, where m is the number of edges, μ is the cyclomatic number, and d_x is the sum of all entries in the x -row of the distance matrix. Clearly, $d_x = P_D(x)(1, \dots, D)$. Accordingly, the Balaban J index can be computed from the molecular D -signature.

Pasaréti Index or Total Wiener Index. This index is an all-path version of the Wiener index and can be defined by the following dot product: $P = ({}^1P, \dots, {}^nP) \cdot L$, where kP is the total number of path of length k and $L = (1, \dots, n)$. As already seen with the shape indices, kP can be calculated from any molecular h -signature with $h \geq k$. According to proposition 3, for all height $h \geq D+1$ we have ${}^h\sigma(G) = {}^{D+1}\sigma(G)$, thus the Pasaréti index can be computed from the molecular $D+1$ -signature, $P(G) = 1/2 \cdot {}^{D+1}\alpha \cdot P(\text{root}({}^{D+1}\Sigma))$ where $P(\text{root}({}^{D+1}\Sigma)) = (P_1(\text{root}({}^{D+1}\Sigma)), \dots, P_n(\text{root}({}^{D+1}\Sigma))) \cdot (1, \dots, n)$.

Detour Index. The detour index, ω , is equal to the half-sum of the off-diagonal elements of the detour matrix. The detour matrix compiles the lengths of the longest paths between all pairs of atoms. As with the Wiener index, ω can be expressed as a dot product, $\omega = P_\Delta \cdot L$, where $P_\Delta = 1/2 \Sigma_x P_\Delta(x)$, where $P_\Delta(x) = ({}^1P_\Delta(x), \dots, {}^nP_\Delta(x))$ and ${}^kP_\Delta(x)$ is the number of distinct atoms for which the longest path in G starting at x has length k . The detour index too can be computed from the molecular $D+1$ -signature: $\omega(G) = 1/2 {}^{D+1}\alpha \cdot \omega(\text{root}({}^{D+1}\Sigma))$ where $\omega(\text{root}({}^{D+1}\Sigma)) = ({}^1P_\Delta(\text{root}({}^{D+1}\Sigma)), \dots, {}^nP_\Delta(\text{root}({}^{D+1}\Sigma))) \cdot (1, \dots, n)$. The same conclusion can be drawn for the hyper detour index ($\omega\omega$).

(5) Hybrid Indices. A hybrid descriptor is one that has more than one of the attributes from the six categories given above. Recently, Randić has introduced a descriptor that is the quotient of atomic paths to atomic walks as well as molecular paths to molecular walks. Such descriptors are an attempt to encode molecular shape.⁴³ The electrotopological state of Kier and Hall^{12,44} encode atoms intrinsic state as well as intrinsic state perturbations calculated through a distance-type matrix. Such a descriptor has attributes of both a path index and distance-type index. Finally the total topological index also introduced by Kier and Hall⁴⁵ and computed from the geometric mean of the valence delta values of the vertices in the paths of a distance matrix is a hybrid descriptor mixing path and distance information.

Randić Molecular Shape p/w Quotient Index. The p_k/w_k quotient index is the ratio of the number of paths by the number of walks of a length k . Formally, $p_k/w_k = \Sigma_x {}^kP(x)/awc_k(x)$, where ${}^kP(x)$ ($awc_k(x)$) is the number of paths (walks) in the molecule of length k starting at atom x . Clearly, this index can be computed from any molecular h -signature with $h \geq k$.

Electrotopological State (E-State) Index. The E-state can be defined as a path/distance index. The E-state of atom x of the molecular graph G is expressed as $S(x) = I(x) + \Sigma_y \Delta I_{xy}$, where $\Delta I_{xy} = [I(x) - I(y)]/d_G(x, y)^m$, $I(x)$ is the intrinsic

state of atom x , and m is a constant (Kier and Hall use $m = 2$). The intrinsic state¹² of an atom (i.e., its local electronic environment) is computed from its degree. Since the E-state of atom x requires it to compute the intrinsic state of all atoms y linked to x , this calculation has to be performed on the maximum signature height, i.e., height $D+1$, and $S(x) = S(\text{root}({}^{D+1}\sigma_G(x)))$. The sum of the intrinsic states of all of the atoms can be used as a descriptor and is thus categorized as a path-based index.

The total topological state index τ is $\tau = \Sigma_{k=1}^n t_{ii} + \Sigma_{i=1}^n \Sigma_{j>i} t_{ij} = [GM_{ij}]^a d^b$, where n is the number of atoms, a and b are constant integers (usually, $a = 1$ and $b = -2$), d is the distance between atoms i and j , and GM_{ij} is the mean of the valence delta values ($\delta()$) of the atoms in the $\langle i, \dots, j \rangle$ path. Since t_{ij} is computed for all pairs of atoms i, j including those at distance D from each other, and because the valence delta value of an atom requires to know its degree, for hydrogen suppressed graphs, the total topological state index has to be computed from the height $D+1$ signature, thus $\tau = {}^{D+1}\alpha_G \cdot (t_{ii}(\text{root}({}^{D+1}\Sigma)) + 1/2 \cdot t_{ij}(\text{root}({}^{D+1}\Sigma)))$.

(6) Information-Theoretic Indices. Information theory provides another class of TIs that have been used in QSRs.¹⁸ Here, a partitioning of the compounds into subsets based on certain features is performed, and the information content contained therein is determined. Also, information content within descriptor sets can be used for an estimate of descriptor variance. Popular information theory TIs include the Shannon entropy index,⁴⁶ the structural information content, and the complementary information content.¹⁹

The information content I of a system of n elements is defined by $I = n \log n - \Sigma n_k \log n_k$, where n_k is the number of elements in the system having the same specific property. For instance with the structural information content index $I_E^D = n \log n - \Sigma n_k \log n_k$, n_k is the number of pairs of atoms at distance k from each other. Clearly $n_k = {}^kP_D$, kP_D being the k -element of path-distance vector $P_D = 1/2 \Sigma P_D(x)$, with $P_D(x)$ having the same definition than with the Wiener index. As all indices based on distance, I_E^D can be computed from the molecular D -signature.

(7) Fragment-Based Indices. Fragments, or structural keys, are generally a large number of small, important atom-groupings determined a priori and listed in a database. The molecule of interest is scanned for these important groups in the database, and the existence or absence of this group in the molecule is noted (normally by a 1 or 0 or by the number of occurrences of that fragment). Such examples of fragments are found in the commercial products by MDL (MACCS or SSKEYS)⁴⁷ and EduSoft, LC (MOLCONN-Z).¹¹ Note that signatures fall into this category of indices. To count the number of occurrences of a given list of fragments one computes the signature of the fragments and counts the number of occurrences of the fragment signatures in the given molecular signature. This requires to perform subtree isomorphism checks between the fragment signatures and the molecular signature. According to proposition 3, to capture all the atoms of the fragments the height of the molecular signature has to be equal to the diameter of the largest fragment augmented by one.

EQUIVALENCE EXAMPLE

This section serves as an illustration to the previous demonstration. Here we show an equivalence between a

Table 8. Compounds and Experimental Boiling Points Used in the Equivalence Example

compound	normal boiling point (K)
2,2-dimethylpropane	280
2,2-dimethylbutane	323
2,2,3-trimethylbutane	354
2,2,3,3-tetramethylbutane	379.7
2,2,3,3-tetramethylpentane	413.5
<i>n</i> -decane	447.3
2,2,5,5-tetramethylhexane	410.7
<i>n</i> -undecane	469.2
<i>n</i> -dodecane	489.5
<i>n</i> -nonane	424.0
2,2,3,4-tetramethylpentane	406.2
<i>n</i> -octane	398.9
3-ethyl,2-methylpentane	388.8
<i>n</i> -heptane	372.0
<i>n</i> -hexane	342
<i>n</i> -pentane	309.0
isobutane	261
2-methylbutane	303
3-methylpentane	336
3-ethyl pentane	367
3-ethyl,3-methylpentane	391.4
2,3,3,4-tetramethylpentane	414.7
4,4-dimethyloctane	430.7
4-methyldecane	460.1
2,7-dimethyloctane	433.1

QSPR using molecular descriptors to that using signatures. Accordingly, we have chosen a small set of linear and branched alkanes as the compound set and the normal boiling point as the property of interest. These experimental data are provided in Table 8. Note that we are using this example just to *show equivalence* and acknowledge the fact that many QSPRs for boiling points (among other properties) exist in the literature already. Hence, we do not report any statistics (other than regression coefficients) for this QSPR since the purpose of this section is not to demonstrate the correlative ability of signatures used as descriptors by themselves.

Since the compounds are alkanes, 1-signatures created from the H-suppressed graphs are limited to four atomic 1-signatures: C(C), C(CC), C(CCC), and C(CCCC). There are, thus, four descriptors in a QSPR created from the atomic 1-signatures; with values equal to the number of occurrences of that atomic 1-signature in the molecule. Accordingly, any comparison made with this QSPR to establish equivalence must contain four molecular descriptors.

In order for the comparison to be useful, the selected molecular descriptors must be able to be written in terms of the four atomic 1-signatures. From this small subset of all molecular descriptors, four were chosen: the ${}^0\chi$ index (equivalent to valence connectivity, ${}^0\chi^v$, here), the ${}^0\kappa$ index (equivalent to zero-order alpha shape-index, ${}^0\kappa_\alpha$, here), the sum of the intrinsic state for each node (S), and the molecular weight (MW). Note that each type of the four descriptors (χ -connectivity,⁴⁸ κ -shape,⁴⁹ S-electrotopological,⁵⁰ and MW) have been shown to be useful in correlating boiling points.

The QSPR created from the four molecular descriptors will have the form $T_{MD} = MD_0 + a^0\chi + b^1\kappa + cS + dM$ where MD_0 , a , b , c , and d represent the multiple linear regression (MLR) parameters, while T_{MD} is the predicted normal boiling point from the model in Kelvin. Using the step forward MRL technique presented earlier with the data listed in Table 8, we find $T_{MD} = 531.79 - 286.75^0\chi + 2317.69^1\kappa + 133.79S - 163.45M$. Likewise, the QSPR

created from the number of occurrences of the atomic 1-signatures will have the form $T_S = S_0 + e\alpha_1 + f\alpha_2 + g\alpha_3 + h\alpha_4$ where S_0 , e , f , g , and h represent the MLR parameters, T_S is the predicted normal boiling point from the model in Kelvin, α_1 represents the number of occurrences of C(C) in the molecule of interest, α_2 represents the number of occurrences of C(CC), α_3 represents the number of occurrences of C(CCC) and α_4 represents the number of occurrences of C(CCCC).

To establish equivalence between the two models, we write each of the four molecular descriptors as a linear function of the number of occurrences of the atom 1-signatures.

Connectivity Index. ${}^0\chi = \deg[C(C)]^{-1/2} \cdot \alpha_1 + \deg[C(CC)]^{-1/2} \cdot \alpha_2 + \deg[C(CCC)]^{-1/2} \cdot \alpha_3 + \deg[C(CCCC)]^{-1/2} \cdot \alpha_4$

For simplicity, we will denote L_1 as $\deg[C(C)]^{-1/2}$, L_2 as $\deg[C(CC)]^{-1/2}$, L_3 as $\deg[C(CCC)]^{-1/2}$, and L_4 as $\deg[C(CCCC)]^{-1/2}$. For example, the molecular 1-signature of 2,2-dimethylpropane is $4C(C) + C(CCCC)$. This implies that $\alpha_1 = 4$, $\alpha_2 = \alpha_3 = 0$, and $\alpha_4 = 1$. For the atomic 1-signature C(C), the root has a degree of 1 and, thus, $L_1 = 1$ while for C(CCCC), the root point has a degree of 4 and, thus $L_4 = 1/2$. Therefore, the above equation correctly predicts a value of 4.5 for the ${}^0\chi$ of 2,2-dimethylpropane.

Shape Index. The first-order Kier-Hall shape index ${}^1\kappa$ defined for the molecules of interest here becomes just a count of one-bond paths. Accordingly, we can arrive at that value by just summing the number of atomic 1-signatures in a molecule and subtract 1. ${}^1\kappa\{G\} = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - 1$. For the 2,2-dimethylpropane, the number of one-bond paths is 4.

Intrinsic State Sum. Since the principal quantum number of carbon is 2, the expression for the Intrinsic State Sum in terms of the atomic 1-signatures is simplified to

$$S = \{(\deg[C(C)]^{-1/2})^2 + 1\} \cdot \alpha_1 + \{(\deg[C(CC)]^{-1/2})^2 + 1\} \cdot \alpha_2 + \{(\deg[C(CCC)]^{-1/2})^2 + 1\} \cdot \alpha_3 + \{(\deg[C(CCCC)]^{-1/2})^2 + 1\} \cdot \alpha_4$$

$$S = (\{[L_1]^2 + 1\} \cdot \alpha_1 + \{[L_2]^2 + 1\} \cdot \alpha_2 + \{[L_3]^2 + 1\} \cdot \alpha_3 + \{[L_4]^2 + 1\} \cdot \alpha_4)$$

For 2,2-dimethylpropane, the sum of the intrinsic state is correctly given as 9.25.

Molecular Weight. The molecular weight of a molecule can be given for the aliphatic hydrocarbons in terms of the atomic 1-signatures by

$$MW = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) \cdot 12 + \{2 \cdot (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + 2\} \cdot 1$$

It is evident from the above expressions that the four molecular descriptors can be written as linear combination of the number of occurrences of the four atomic 1-signatures. Substituting the previous expressions into the equation $T_{MD} = 531.79 - 286.75^0\chi + 2317.69^1\kappa + 133.79S - 163.45M$, leads to $T_{MD} = T_S = -236.41 + 230.94\alpha_1 + 27.37\alpha_2 - 178.37\alpha_3 - 388.0\alpha_4$. Thus, both QSPRs (T_{MD} and T_S) will predict the same values for the same compounds. That this

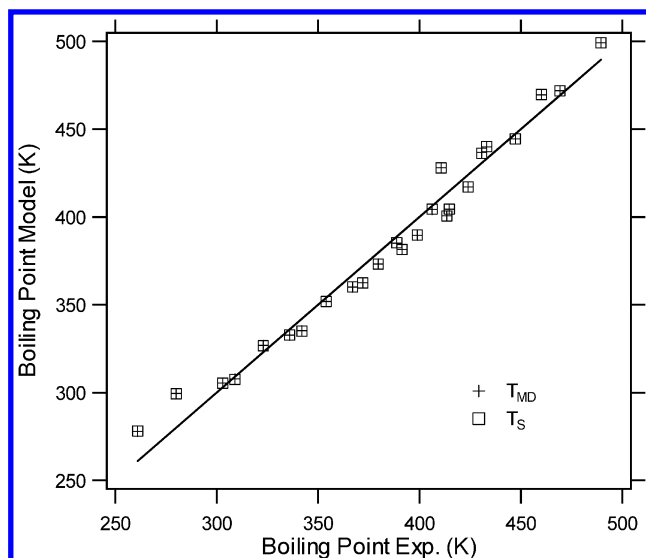


Figure 7. The predicted boiling point from the two QSARs plotted versus the experimental boiling points. The cross indicates the QSAR as per the molecular descriptors, while the empty square indicates the QSAR as per the height-1 signatures. A 45° line is included as a guide for the eye.

is true can be seen in Figure 7 where we plot the prediction of both models as a function of the experimental data.

DISCUSSION AND CONCLUDING REMARKS

In this work we introduced the concept of signature as a new molecular descriptor. Atomic and molecular signatures can be computed in polynomial time and store respectively in $O(n)$ and $O(n^2)$ space. Put to the test in two QSAR and QSPR studies, signature performed quite well compared with the Molconn-Z commercially available 2D descriptors. For both studies, the statistics of signature and the 2D descriptors were found similar (Tables 4 and 6), while signature outperformed the 2D descriptors in term of orthogonality (Figures 3 and 5) and stability (Figures 4 and 6). Comparable conclusions have already been drawn while evaluating HQSAR fragment based holograms⁵¹ with the CODESA descriptors to predict the binding of chemicals to estrogen receptors. Analogous to a molecular signature, a molecular hologram is composed of the substructural fragments constituting a molecule. Unlike signatures, holograms are stored into a bin array of predefined length, and thus the structure of the fragments are lost since several structurally different fragments may appear in the same bin.⁵¹ Additionally, like the height with signature, holograms have a predefined size; however, unlike atomic signatures, holograms are not centered on a particular atom. Yet, it is surprising that both signatures and holograms, solely based on fragments, perform as well as hundreds of molecular descriptors encoding features as varied as molecular topology, geometry, electrostatics, and information content. A simple explanation to this result is that 2D descriptors can in fact be computed from molecular fragments. Precisely, we have proven in this paper that there exists a signature height no greater than the diameter of the molecule augmented by one, for which any given molecular descriptor can be computed (Table 7). As a consequence of this finding, any existing QSAR or QSPR equation can be replaced with an equivalent equation involving occurrence numbers of substructural fragments (e.g., atomic signatures).

At this point one may be tempted to make a parallel between signature and group contribution and draw the general conclusion that group contribution can perform as well if not outperform any QSAR/QSPR based on 2D descriptors. However, it is important to point out that in group contribution the fragments or groups are chosen through expert knowledge prior to performing the QSAR/QSPR analysis and do not necessarily represent all the possible atomic arrangements found in a training set. Conversely, with signature or molecular holograms, there is no expert knowledge involved in choosing the fragments or groups, but the fragments considered span the entire set of all possible atomic arrangements of a predefined size. Thus, while it appears that a given QSAR/QSPR involving 2D descriptors can be replaced with an equation of similar performance involving signature or holograms, the statement is not necessarily true for group contribution in general.

The search for a “universal” molecular descriptor from which other can be derived is not a new idea. Baskin et al.⁵² proved that any TI can be uniquely represented as a linear combination of occurrence numbers of some substructures. This results were first established for molecular graphs comprising n atoms with a set of N substructures equal to the set of all possible labeled graphs composed of n vertices. Precisely, it was proven that for any graph H , any TI, $f(H)$, is uniquely represented in the form

$$f(H) = \sum_{j=1}^N c_j g_j(H) \quad (7)$$

where c_j is a constant independent of H and dependent on f , and $g_j(H)$ is the occurrence number of a graph H_j in the graph H (i.e., the number of times H_j appears as a subgraph of H). Each coefficient c_j is computed by solving the system of N equations and N unknowns $F = BC$, where F is the TI vector $(f(H_1), f(H_2), \dots, f(H_N))$, B is the square matrix with elements $b_{ij} = g_j(H_i)$, and C is the coefficient vector (c_1, c_2, \dots, c_N) . Consequently, Baskin et al. established that the matrix B is a basis algebra for graph invariants; that is, a molecular descriptor from which all others can be computed. Unfortunately, this theoretical result is of limited applicability since the size of matrix B is N^2 where N is the total number of labeled graphs of n vertices, which scales at least 2^n .⁵³ In their paper, Baskin et al. list already 20 labeled graphs with only three vertices. Storage consideration is not the only shortcoming of the above result. To compute matrix B one has to perform a subgraph isomorphism for each matrix entry, and this cannot be performed efficiently since subgraph isomorphism is known to be NP-complete.⁵⁴

In a more recent paper by the same authors⁵⁵ the storage limitation of the technique was recognized, and it was established that the system $F = BC$ could be solved for any matrix B with $\det B \neq 0$. Thus, in this later result for any given graph H of n vertices, the set of subgraphs to be considered is not necessarily the set of all labeled graphs of size n but is any set for which $\det B \neq 0$. While the authors present in their paper sets and subsequent B matrices of limited size, they do not provide a procedure on how to determine the B matrices such that $\det B \neq 0$. Instead, the selection of the set of subgraphs appears to be problem specific, and it does seem it can be automated.

In contrast, computing the signature of a given molecular graph is systematic and scales polynomially both in time and space. Furthermore, we too provide a general equation (eq 8) from which any TI can be calculated. Both eqs 8 and 9 basically express the fact that any TI can be computed from the dot product of a vector of occurrence numbers of some substructures and a vector of constants. While Baskin et al. solve a system of equations to compute the constants, we provide not only a simple scheme to compute these constants but also their significance. The vector of constant is the vector of TI values applied to an atom the substructure, it is computed the same way the TI would be on any given atom of a molecular graph.

While we have established the utility of signature in QSAR and QSPR study, in the subsequent paper of the series we explore how signature can be used to inverse design molecule. Specifically, we outline an algorithm that enumerates all molecular graphs corresponding to a given signature. This algorithm is then used to probe the degeneracy of signature versus other descriptors.

A program to compute signatures from various input formats is available from the authors upon request.

ACKNOWLEDGMENT

Funding for this work was provided by the U.S. Department of Energy and Sandia National Laboratories under Grant number DE-AC04-76DP00789. J.L.F. is also pleased to acknowledge funding provided by the Math Information and Computer Science program of the U.S. Department of Energy. R.S.P. would like to acknowledge additional support from the Center for the Management, Utilization and Protection of Water Resources at Tennessee Technological University.

REFERENCES AND NOTES

- (1) Kier, L. B.; Hall, L. H. Intermolecular accessibility: The meaning of molecular connectivity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792–795.
- (2) Brown, R. D.; Martin, Y. C. The information content in 2D and 3D structural descriptors relevant to ligand–receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (3) Randić, M.; Zupan, J. On interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550–560.
- (4) Randić, M.; Basak, S. C. A new descriptor for structure–property and structure–activity correlations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 650–656.
- (5) Faulon, J.-L. Stochastic generator of chemical structure: 1. Application to the structure elucidation of large molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1204–1218.
- (6) Faulon, J.-L.; Visco, J. D. P.; Pophale, R. S. Developing a Methodology for an Inverse Quantitative Structure–Activity Relationship Using the Signature Molecular Descriptor. *J. Molecular Graphics Modeling* **2002**, *20*, 429–438.
- (7) Kucera, L. *Combinatorial algorithms*; Adam Hilger: Bristol, 1990.
- (8) Faulon, J. L. Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 432–444.
- (9) Babai, L.; Luks, E. M. Canonical Labeling of Graphs. *Proc. 15th ACM Symp. Theory Comput.* **1983**, 171–183.
- (10) Miyazaki, T. *The Complexity of McKay's Canonical Labeling Algorithm*. In *Groups and Computation II*; Finkelstein, L., Kantor, W. M., Eds.; Amer. Math. Soc.: Providence, RI, 1997; pp 239–256.
- (11) Hall, L. H. *MOLCONN-Z*; Hall Associates Consulting: Quincy, MA, 1991.
- (12) Kier, L. B.; Hall, L. H. *Molecular structure description*; Academic Press: San Diego, CA, 1999.
- (13) Draper, N. R.; Smith, H. In *Applied Regression Analysis*, 2nd ed.; John Wiley & Sons: New York, 1981.
- (14) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **1933**, *24*, 417–441.
- (15) Randić, M. On computation of optimal parameters for multivariate analysis of structure–property relationship. *J. Comput. Chem.* **1991**, *12*, 970–980.
- (16) Randić, M. Resolution of ambiguities in structure–property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
- (17) Corporation, S. R. *SRC PHYSPROP Database*. p.http://esc.syrres.com/.
- (18) Trinajstić, N. *Chemical Graph Theory*, 2nd ed. In *Mathematical Chemistry*; Klein, D. J., Randić, M., Eds.; CRC Press: Boca Raton, FL, 1992.
- (19) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. *Use of graph-theoretical and geometrical molecular descriptors in structure–activity relationships*. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Publishing Corp.: New York, 1997; pp 73–116.
- (20) Rucker, G.; Rucker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683–695.
- (21) Rucker, G.; Rucker, C. On topological indices, boiling points, and cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788–802.
- (22) Rucker, G.; Rucker, C. Walk counts, labyrinthicity, and complexity of acyclic and cyclic graphs and molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 99–106.
- (23) Gutman, I.; Rucker, C.; Rucker, G. On walks in molecular graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 739–745.
- (24) Randić, M. On the characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
- (25) Kier, L. B.; Hall, L. H. Derivation and significance of valence molecular connectivity. *J. Pharm. Sci.* **1981**, *70*, 583–589.
- (26) Randić, M. Novel graph theoretical approach to heteroatoms in QSAR. *Chemometrics Intel. Lab. Syst.* **1991**, *10*, 213–227.
- (27) Kier, L. B. A shape index from molecular graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109.
- (28) Kier, L. B. Indexes of molecular shape from chemical graphs. *Acta Pharm. Jugosl.* **1986**, *36*, 171.
- (29) Hall, L. H.; Kier, L. B. The molecular and connectivity chi indexes and kappa shape indexes in structure–property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; pp 367–422.
- (30) Randić, M. Graph valence shells as molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 627–630.
- (31) Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.* **1947**, *15*, 419.
- (32) Hosoya, H. Topological Index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332–2339.
- (33) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
- (34) Balaban, A. T. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)* **1986**, *21*, 115–122.
- (35) Bonchev, D. The overall Wiener index - A new tool for characterization of molecular topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 582–592.
- (36) Lukovits, I. The detour index. *Croat. Chem. Acta* **1996**, *69*, 873–882.
- (37) Linert, W.; Lukovits, I. Formulas for the hyper-Wiener and hyper-detour indices of fused bicyclic structures. *Comm. Math. Comput. Chem. (MATCH)* **1997**, *35*, 65–74.
- (38) Lukovits, I. An all-path version of the Wiener index. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 125–129.
- (39) Lucic, B.; Lukovits, I.; Nikolic, S.; Trinajstić, N. Distance-related indexes in the quantitative structure–property relationship modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 527–535.
- (40) Randić, M. Linear combinations of path numbers as molecular descriptors. *New J. Chem.* **1997**, *21*, 945–951.
- (41) Ivanciuc, O.; Balaban, T. S.; Balaban, A. T. Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. *J. Math. Chem.* **1993**, *12*, 309–318.
- (42) Ivanciuc, O. QSAR comparative study of Wiener descriptors for weighted molecular graphs. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412–1422.
- (43) Randić, M. Novel shape descriptors for molecular graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 607–613.
- (44) Kier, L. B.; Hall, L. H. An electrotopological state index for atoms in molecules. *Pharm. Res.* **1990**, *7*, 801.
- (45) Hall, L. H.; Kier, L. B. Determination of Topological Equivalence in Molecular Graphs from the Topological State. *Quant. Struct.-Act. Relat.* **1990**, *9*, 115.
- (46) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1963.
- (47) MDL Information Systems, I.: San Leandro, CA.

- (48) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054–1060.
- (49) Espinosa, G.; Yaffe, D.; Cohen, Y.; Arenas, A.; Giral, F. Neural network based quantitative structural property relations (QSPRs) for predicting boiling points of aliphatic hydrocarbons. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 859–879.
- (50) Hall, L. H.; Story, C. T. Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004–1014.
- (51) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Shhehan, D. M. Evaluation of Quantitative Structure–Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669–67.
- (52) Baskin, I. I.; Skvortsova, M. I.; Stankevich, I. V.; Zefirov, N. S. On the basis of invariants of labeled molecular graphs. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 527–531.
- (53) Harary, F.; Palmer, E. M. *Graphical Enumeration*; Academic Press: New York, 1973.
- (54) Garey, M. R.; Johnson, D. S. *Computers and Intractability. A Guide to the Theory of NP-completeness*; W. H. Freeman & Company: New York, 1979.
- (55) Skvortsova, M. I.; Baskin, I. I.; Skvortsova, L. A.; Palyulin, V. A.; Stankevich, I. V.; Zefirov, N. S. Chemical graphs and their basis invariants. *Theochem: J. Mol. Struct.* **1999**, *466*, 211–217.
- (56) Perez, C.; Pastor, M.; Ortiz, A. R.; Gago, F. Comparative binding energy analysis of HIV-1 protease inhibitors: Incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Med. Chem.* **1998**, *41*, 839–852.
- (57) Young, S. D.; Payne, L. S.; Thompson, W. J.; Gaffin, N.; Lyle, T. A.; Britcher, S. F.; Graham, S. L.; Schultz, T. H.; Deana, A. A.; Darke, P. L.; Zugay, J.; Schleif, W. A.; Quintero, J. C.; Emini, E. A.; Anderson, P. S.; Huff, J. R. HIV-1 protease inhibitors based on hydroxyethylene dipeptide isosteres: An investigation into the role of the P1' side chain on structure–activity. *J. Med. Chem.* **1992**, *35*, 1702–1709.
- (58) Thompson, W. J.; Fitzgerald, P. M. D.; Holloway, M. K.; Emini, E. A.; Darke, P. L.; McKeever, B. M.; Schleif, W. A.; Quintero, J. C.; Zugay, J.; Tucker, T. J.; Schwering, J. E.; Homnick, C. F.; Nunberg, J.; Springer, J. P.; Huff, J. R. Synthesis and antiviral activity of a series of HIV-1 protease inhibitors with functionality tethered to P1 or P1' phenyl substituents: X-ray crystal structure assisted design. *J. Med. Chem.* **1992**, *35*, 1685–1701.
- (59) Vara Prasad, J. V. N.; Para, K. S.; Lunney, E. A.; Ortwine, D. F.; Dunbar, J., J. B.; Ferguson, D.; Tummino, P. J.; Hupe, D.; Tait, B. D.; Domagala, J. M.; Humblet, C.; Bhat, T. N.; Liu, B.; Guerine, D. M. A.; Baldwin, E. T.; Erickson, J. W.; Sawyer, T. K. Novel series of achiral, low molecular weight, and potent HIV-1 protease inhibitors. *J. Am. Chem. Soc.* **1994**, *116*, 6989–6990.
- (60) Beaulieu, P. L.; Wernic, D.; Abraham, A.; Anderson, P. C.; Bogri, T.; Bousquet, Y.; Croteau, G.; Guse, I.; Lamarre, D.; Liard, F.; Paris, W.; Thibeault, D.; Pav, S.; Tong, L. Potent HIV protease inhibitors containing a novel (hydroxyethyl)amide isostere. *J. Med. Chem.* **1997**, *40*, 2164–2176.
- (61) Lunney, E. A.; Hagen, S. E.; Domagala, J. M.; Humblet, C.; Kosinski, J.; Tait, B. D.; Warmus, J. S.; Wilson, M.; Ferguson, D.; Hupe, D.; Tummino, P. J.; Baldwin, E. T.; Bhat, T. N.; Liu, B.; Erickson, J. W. A novel non-peptide HIV-1 protease inhibitor: Elucidation of the binding mode and its application in the design of related analogues. *J. Med. Chem.* **1994**, *37*, 2664–2677.
- (62) Mayo, S. L.; Olafson, B. P.; Goddard, I., W. A. Dreiding: A generic force field. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

CI020345W