

# Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor

Donald P. Visco Jr.<sup>a</sup>, Ramdas S. Pophale<sup>a</sup>, Mark D. Rintoul<sup>b</sup>, Jean-Loup Faulon<sup>b,\*</sup>

<sup>a</sup> Department of Chemical Engineering, Tennessee Technological University, Box 5013, Cookeville, TN, USA

<sup>b</sup> Sandia National Laboratories, P.O. Box 969, MS 9951 Livermore, CA, USA

## Abstract

The concept of signature as a molecular descriptor is introduced and various topological indices used in quantitative structure-activity relationships (QSARs) are expressed as functions of the new descriptor. The effectiveness of signature versus commonly used descriptors in QSAR analysis is demonstrated by correlating the activities of 121 HIV-1 protease inhibitors. Our approach to the inverse-QSAR problem consists of first finding the optimum sets of descriptor values best matching a target activity and then generating a focused library of candidate structures from the solution set of descriptor values. Both steps are facilitated by the use of signature. © 2002 Elsevier Science Inc. All rights reserved.

**Keywords:** Inverse-QSAR; Signature; Topological indices; Molecular descriptor

## 1. Introduction

In cheminformatics topological constraints are represented by topological descriptors or indices (TIs), which are graph invariants under isomorphism [1]. These descriptors are routinely utilized by biochemists, environmental chemists, and chemical engineers to derive quantitative structure-activity relationships (QSARs) and quantitative structure-property relationships (QSPRs) to predict biological, pharmaceutical, physical and chemical properties from molecular structures. The process is to evaluate the given TIs in a particular QSAR and input these values into an equation to solve for the unique solution for the activity of interest. Such a procedure is known as the forward QSAR problem. The inverse, or reverse, problem is much more taxing. Here, a given activity value is desired and the goal is to determine the structures that correspond to the specified value.

The inverse problem in QSAR can be written as an equation whose unknowns are integers (diophantine equation) wherein the researcher attempts to determine the solution sets (descriptor values) that correspond to the desired activity. After this set is determined, the solution procedure becomes the enumeration/sampling of molecular graphs corresponding to the given set of descriptor values. While retrieving graphs from their invariants may be solvable for specific invariants [2–4], the sheer number of structural

descriptors (several hundreds—see Table 1 for a sample list) that are currently used in cheminformatics makes this approach impractical. Ideally, one would like to find a universal descriptor that could easily be used with enumeration or sampling and from which other descriptors could be computed. We have developed such a descriptor, named signature.

In this work, we introduce the concept of a molecular signature and show how one can write many of the topological indices used in QSAR from these signatures. We provide a simple example showing that one need only count the appearance of a particular signature in a molecule as a descriptor versus calculating some TIs and demonstrate the equivalence of the resulting QSPRs. Finally, we provide a more strenuous test of the use of signature as descriptors in QSARs by correlating the activity of 121 HIV-1 protease inhibitors from signatures of various heights. We compare this to a QSAR determined from a commercially available package. We conclude by discussing the next step in this procedure, namely designing a focused library using signatures.

## 2. The signature descriptor

The *signature* is a systematic codification system over an alphabet of atom types, which can be compared to the SMILES notation system [5]. This concept, which was first presented and applied in the context of structural elucidation, [6] is generalized in the present paper. Prior to introducing

\* Corresponding author. Tel.: +1-925-294-1279; fax: +1-925-294-3020.  
E-mail address: jfaulon@sandia.gov (J.-L. Faulon).

the signature descriptor, some terminology and notation have to be defined.

### 2.1. 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. The function  $c_G()$  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 that 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:

$$\left. \begin{aligned} V({}^h\sigma_G(x)) &= \bigcup_{k=0}^h {}^kV(x), & E({}^h\sigma_G(x)) &= \bigcup_{k=0}^h {}^kE(x) \\ {}^0V(x) &= u_0, & {}^0E(x) &= \{\} \\ {}^kV(x) &= \bigcup_{u \in {}^{k-1}V(x)} {}^{+1}v_\sigma(u), & {}^kE(x) &= \bigcup_{u \in {}^{k-1}V(x)} {}^{+1}e_\sigma(u) \\ {}^{+1}v_\sigma(u) &= \bigcup_{y \in V_G(l_G(u))} u_y - {}^{-1}v_\sigma(u), & {}^{+1}e_\sigma(u) &= \bigcup_{v \in {}^{+1}v_\sigma(u)} [u, v] \end{aligned} \right| \quad (2)$$

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

The distance between two atoms  $x$  and  $y$  of  $G$ ,  $d_G(x, y)$ , is the minimum length of all the paths between  $x$  and  $y$ . Finally, we define,  ${}^hV_G(x)$ , as the set of neighbors of  $x$  that are at distance  $h$  from  $x$ . We write  $v_G(x) = {}^1V_G(x)$ , and we have  ${}^0V_G(x) = \{x\}$ .

### 2.2. Signature-tree

Let  $G = (V_G, E_G, C, c_G())$  be a molecular graph, the  $h$ -signature-tree, or  ${}^h\sigma$ -tree, of an atom  $x$  of  $V_G$ , is

Table 1

A list of the various types of topological indices available for use in QSPRs/QSARs

Descriptor	Reference (if applicable)
Atoms, bonds, molecular weight, cyclomatic number	
Connectivity index	[18]
Valence connectivity index	[19]
Variable connectivity index	[20]
Kier and Hall shape index	[21]
Intrinsic state sum	[10]
Valence shell	[22]
Platt number	[23]
Wiener number	[24]
Balaban $J$ index	[25]
Overall Wiener number	[26]
Path/Walk quotient	[27]
Electrotopological state	[10]

a tree describing the neighborhood of  $x$  in  $G$  up to distance  $h$ . More precisely, the  ${}^h\sigma$ -tree of  $x$ ,  ${}^h\sigma_G(x) = (V({}^h\sigma_G(x)), E({}^h\sigma_G(x)), C, c_\sigma())$  is a rooted-tree on  $x$ , where the first layer is composed of the neighbors of  $x$ , the second layer is composed of the neighbors of the first layer, and this recursively up to layer  $h$ . Any vertex,  $u_{i+1}$ , of layer  $i + 1$  must be the neighbor of a vertex  $u_i$  of layer  $i$ , and be different than the predecessor of  $u_i$  in layer  $i - 1$ . In other words, when developing signature-trees one is not allowed to backtrack on the walks initiated at the root. Note, however, that for cyclic compounds, a given atom  $y$  of  $G$  can be represented several times in  ${}^h\sigma_G(x)$ , and we write  $l_\sigma(y)$  the set of corresponding vertices in  $V({}^h\sigma_G(x))$ . Conversely, we write  $l_G(u)$  the single atom of  $G$  corresponding to vertex  $u$  of  $V({}^h\sigma_G(x))$ . Since a rooted-tree is a directed graph, let  ${}^{+k}v_\sigma(x)(u)$  be the set of  $k$ -distant children of  $u$  in  ${}^h\sigma_G(x)$ , and let  ${}^{-k}v_\sigma(u)$  be the unique  $k$ -distant parent of  $u$  in  ${}^h\sigma_G(x)$ . Using the above notation, the formal definition of the  ${}^h\sigma$ -tree,  ${}^h\sigma_G(x) = (V({}^h\sigma_G(x)), E({}^h\sigma_G(x)), C, c_\sigma())$  of an atom  $x$  of  $V_G$ , is:

where  $u_0$  and  $u_y$  verify  $l_G(u_0) = x$  and  $l_G(u_y) = y$ .

Several coloring functions  $c_\sigma()$  can be used with the vertices of  ${}^h\sigma_G(x)$  depending on how signature-trees are used. Let  $u_k$  be a vertex of layer  $k$  in  ${}^h\sigma_G(x)$ , i.e.  $u_k$  belongs to  ${}^kV(x)$ . In our original paper [6],  $u_k$  was colored by  $c_\sigma(u_k) = c_G(l_G(u_k))$ , hence, each vertex of  ${}^h\sigma_G(x)$  was colored with the type of the corresponding atom. In the present paper, we will be concerned only with acyclic compounds, thus the concepts of walk, trail and path are indistinguishable and our coloring function is only that of atom type. However, in general, one would need to expand the coloring function to include the number of appearances of  $l_G(u_k)$  in layer  $k$ , and the type of walk (i.e. a walk, trail or path—including a shortest and longest path) [7]. Examples of colored signature-trees are given in Fig. 1.

### 2.3. Signature of an atom

Let  $G = (V_G, E_G, C, c_G())$  be a molecular graph and let  $x$  be a atom of  $V_G$ . The signature of height  $h$  of  $x$  is a canonical representation of the  $h$ - $\sigma$ -tree,  ${}^h\sigma_G(x)$ , colored by the function  $c_\sigma()$ . Since there is a one-to-one mapping between signatures and signature-trees we use the same notation,  ${}^h\sigma_G(x)$ , to represent both objects. Rooted-tree canonization can be achieved in a time linearly proportional to the size of the tree [8]. In the present paper we have made use of an algorithm developed for labeled graphs and suitable for molecular graphs [9]. Once  ${}^h\sigma_G(x)$  has been canonized, the signature is written by reading the tree in a depth first-order and printing the character ‘(’ each time an edge parent-child

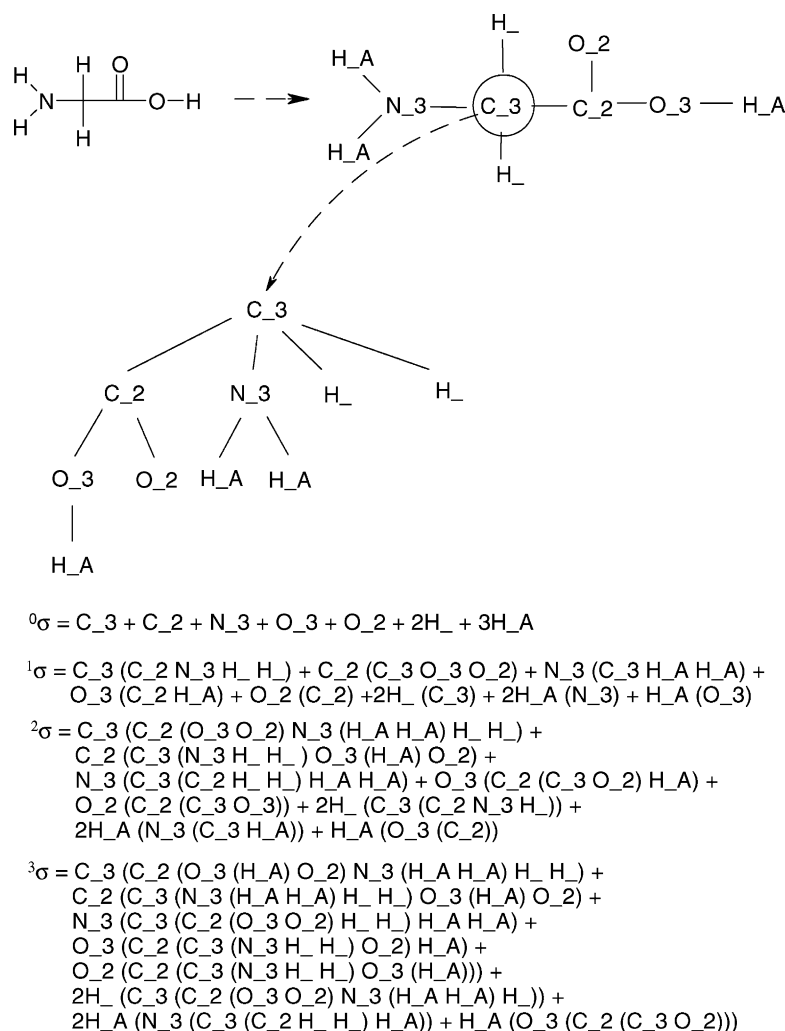


Fig. 1. The molecular graph of glycine colored using the Dreiding 2.21 force field [35]. We show the signature-tree of the atom colored C\_3 and the molecular signature of glycine for heights 0, 1, 2 and 3 which are the sum of the atomic signatures for each root point. Here, the hierarchy of labels is  $C\_3 > C\_2 > N\_3 > O\_3 > O\_2 > H\_ > H\_A$ .

is read, the character ‘)’ when the edge is read from child to parent, and a color if the item read is a vertex that has not already been visited.

#### 2.4. Signature of a molecule

According to the above definition, the signature of an atom can be viewed as a string of characters over an alphabet 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 as the linear combination of its atomic signatures:

$${}^h\sigma(G) = \sum_{x \in V_G} {}^h\sigma_G(x) = \sum_{i=1}^{{}^hK_G} {}^h\sigma_i {}^h\sigma_G({}^hX_i) \quad (3)$$

where  ${}^h\sigma_G({}^hX_i)$  is a base vector,  ${}^h\alpha_i$  is the number of atoms having the signature of the base vector, and  ${}^hK_G$  is the number of base vectors. Example of molecular signatures are given in Fig. 1 as a sum of atomic signatures.

### 3. Computing topological indices from signature

In this section, we provide the expressions that give various topological indices in terms of signatures, where applicable. We provide a more detailed demonstration of these relationships in another work [7].

In the following, we consider a *hydrogen-suppressed* covalent molecule  $G$  for acyclic molecular graphs but known molecular signatures up to height  $h$ ,  ${}^h\sigma(G) = \sum_{i=1}^{{}^hK_G} {}^h\alpha_i {}^h\sigma_G({}^hX_i)$ , where  $\sigma_G({}^hX_i)$ ,  $i = 1, \dots, {}^hK_G$ , are atomic  $h$ -signatures. We also define  $|{}^kV(x)|$  as the number of vertices in layer  $k \leq h$  in  ${}^h\sigma_G(x)$ , that is the number of vertices that are a distance  $k$  from  $x$ . The vertices of

the  $h$ -signature are colored with the function  $c_\sigma()$  defined earlier over the periodic table. Our results indicate that all chosen TIs can be computed from the  $h$ -signature of the unknown molecular graph for  $h$  no greater than  $n$ , the number of atoms computed from the 0-signature. We provide relationships for a few indices next and note that a more exhaustive list is presented elsewhere [7].

### 3.1. Number of atoms

The number of atoms,  $n$ , is a count of the coefficients of the atomic 0-signature.

$$n = \sum^0 \alpha_i \quad (4)$$

### 3.2. Number of bonds

The number of bonds,  $m$ , can be obtained by counting the number of vertices in the atomic 1-signature. If this is written in terms of base vectors, the coefficients are included as follows.

$$m = \sum^1 \alpha_i |^1 V(^1 X_i)| \quad (5)$$

### 3.3. Connectivity index—order 1 ( $^1\chi$ )

This index is defined by

$$^1\chi = \sum_{\text{paths}} [\deg(x_1)\deg(x_0)]^{-1/2} \quad (6)$$

using signature, we can write this index as

$$^1\chi = \frac{1}{2} \sum_{i=1}^{K_G} \alpha_i \sum_{u \in ^1 V_2(^1 X_i)} [\deg(u)\deg(^1 v_\sigma(u))]^{-1/2} \quad (7)$$

where  $h \geq 2$ . The higher-order indices follow from above, though note that the zero-order index does not have a double-counting factor, so there is no need for the  $1/2$  coefficient.

### 3.4. Kier–Hall shape index—order 1 ( $^1\kappa$ )

This index is defined as follows:

$$^1\kappa = \frac{n(n-1)^2}{(^1P)^2} \quad (8)$$

where  $^1P$  are the number of paths of length 1. Written in terms of signature, this relationship is

$$^1\kappa = \frac{(\sum^0 \alpha_i) ([\sum^0 \alpha_i] - 1)^2}{[1/2 \sum^h \alpha_i |^1 V(^1 X_i)|]^2} \quad (9)$$

where the denominator provides the number of paths and can be computed from any molecular signature of height  $\geq 1$ . The other Kier–Hall shape indices are determined in a similar manner.

### 3.5. Wiener index

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 number of path-distance vectors of the atoms,  $P_D = 1/2 \sum P_D(x)$ , where  $P_D(x) = (|^1 V(x)|, \dots, |^D V(x)|)$ .

### 3.6. Electrotological state (E-state) index

The E-state of atom  $x$  of the molecular graph  $G$  is expressed as  $S(x) = I(x) + \sum_y \Delta I_{xy}$ , 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  $m = 2$ ) [10]. The intrinsic state of an atom requires computation of all its paths of length 1, that is the set of neighbors of that atom which can be computed for all vertices in  $^h \sigma_G(x)$  that are not in layer  $h$ .

The E-state of atom  $x$  is derived as the path-distance vector:

$$S(x) = I(u_0) + \sum_{k=1}^n \sum_{u_k \in ^k V(x)} \frac{I(u_0) - I(u_k)}{k^m} \quad (10)$$

since intrinsic state must be calculated from all atoms including those at the maximum distance  $D$  from  $x$ ,  $S(x)$  must be computed from the  $D + 1$ -signature of  $x$ .

## 4. Equivalence example

Some topological descriptors can be written as linear combinations of the number of occurrences of atomic signatures of a certain height. To the extent that this is true, a QSPR performed on either set (the TIs or the atomic signatures) will, necessarily, be equivalent. To demonstrate this, we have chosen a small set of linear and branched alkanes as the compound set and the normal boiling point of these compounds as the property of interest (Fig. 2). This experimental data is provided in Table 2.

Since the compounds are alkanes, molecular 1-signatures created from the H-suppressed graphs are limited to four atomic 1-signatures, namely 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 the particular atomic 1-signature in the molecule. Accordingly, we will choose four topological indices that can be written in terms of atomic 1-signatures: the zero-order connectivity index ( $^0\chi$ ), the first-order Kier–Hall shape index ( $^1\kappa$ ), the sum of the intrinsic state for each node ( $S$ ) and the molecular weight ( $M_W$ ). Note that each type of the four descriptors ( $\chi$ -connectivity,  $\kappa$ -shape,  $S$ -electrotological and  $M_W$ ) [11–13] have been shown to be useful in correlating boiling points.

Table 2

The compounds and experimental boiling points used in the equivalence example

Compound	Normal boiling point (K) [28]
2,2-Dimethyl propane	280
2,2-Dimethyl butane	323
2,2,3-Trimethyl butane	354
2,2,3,3-Tetramethyl butane	379.7
2,2,3,3-Tetramethyl pentane	413.5
<i>n</i> -Decane	447.3
2,2,5,5-Tetramethyl hexane	410.7
<i>n</i> -Undecane	469.2
<i>n</i> -Dodecane	489.5
<i>n</i> -Nonane	424.0
2,2,3,4-Tetramethyl pentane	406.2
<i>n</i> -Octane	398.9
3-Ethyl-2-methyl pentane	388.8
<i>n</i> -Heptane	372.0
<i>n</i> -Hexane	342
<i>n</i> -Pentane	309.0
<i>Iso</i> -butane	261
2-Methyl butane	303
3-Methyl pentane	336
3-Ethyl pentane	367
3-Ethyl-3-methyl pentane	391.4
2,3,3,4-Tetramethyl pentane	414.7
4,4-Dimethyl octane	430.7
4-Methyl decane	460.1
2,7-Dimethyl octane	433.1

The QSPR created from the four molecular descriptors will have the form  $T_{MD} = a^0\chi + b^1\kappa + cS + dM_w$  where  $a$ ,  $b$ ,  $c$  and  $d$  represents the multiple linear regression (MLR) parameters, while  $T_{MD}$  is the predicted normal boiling point from the model in K. Likewise, the QSPR created from the number of occurrences of the atomic 1-signatures will have the form  $T_S = e\alpha_1 + f\alpha_2 + g\alpha_3 + h\alpha_4$ , where  $e$ ,  $f$ ,  $g$  and  $h$  represent the MLR parameters,  $T_S$  is the predicted normal boiling point from the model in K and  $\alpha_1$  represents the number of occurrences of C(C) in the molecule of interest,  $\alpha_2$  represents the number of occurrences of C(CC),  $\alpha_3$  represents the number of occurrences of C(CCC) and  $\alpha_4$  represents the number of occurrences of C(CCCC).

#### 4.1. Connectivity index

Following Eq. (7) (and using 1 instead of 1/2), we write

$${}^0\chi = \deg[C(C)]^{-1/2}\alpha_1 + \deg[C(CC)]^{-1/2}\alpha_2 + \deg[C(CCC)]^{-1/2}\alpha_3 + \deg[C(CCCC)]^{-1/2}\alpha_4 \quad (11)$$

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-dimethyl propane is 4C(C) + C(CCCC). This implies that  $\alpha_1 = 4$ ,  $\alpha_2 = \alpha_3 = 0$  and  $\alpha_4 = 1$  for this compound. 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

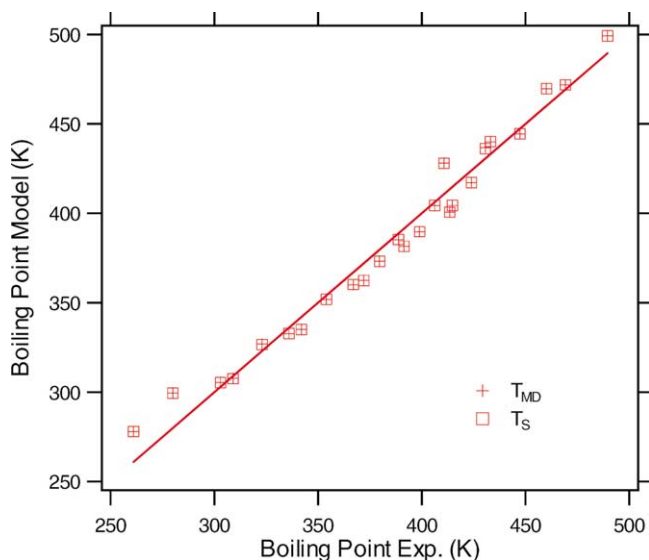


Fig. 2. The predicted boiling point from the two QSPRs plotted vs. the experimental boiling points. The cross indicates the QSPR as per the molecular descriptors while the empty square indicates the QSPR as per the height-1 signatures. A 45° line is included as a guide for the eye.

and, thus  $L_4 = 1/2$ . Therefore, the above equation properly predicts a value of 4.5 for the  ${}^0\chi$  of 2,2-dimethyl propane.

#### 4.2. Shape index

The first-order Kier–Hall shape index  ${}^1\kappa$  defined for the molecules of interest here become just a count of the number of carbon atoms. Accordingly, we can arrive at that value

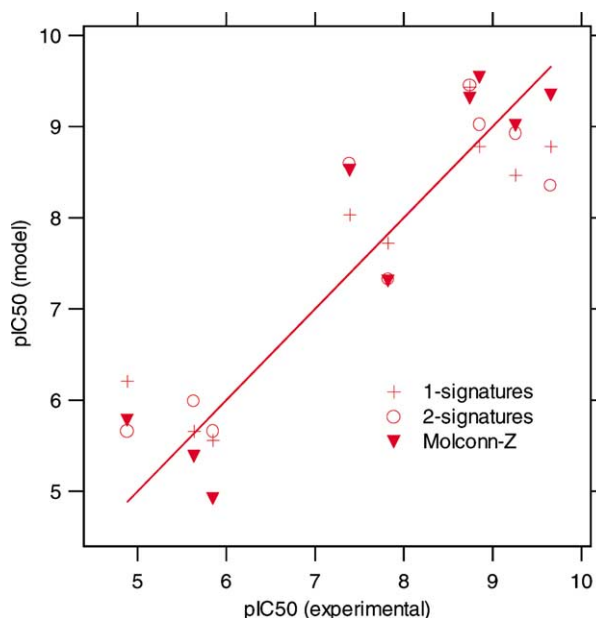


Fig. 3. The prediction of the  $pIC_{50}$  values for the nine test set compounds using each of the models.

by just summing the number of atomic 1-signatures in a molecule.

$${}^1\kappa = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \quad (12)$$

For the 2,2-dimethyl propane, the number of carbon atoms is five.

#### 4.3. Intrinsic state sum

The intrinsic state for the carbon atoms in an alkane is give as  $I_i = (\deg[u]^{-1/2})^2 + 1$ . In terms of the atomic 1-signatures, the sum of the intrinsic states of all the carbon atoms are given as

$$\begin{aligned} S = & \{(\deg[C(C)]^{-1/2})^2 + 1\}\alpha_1 + \{(\deg[C(CC)]^{-1/2})^2 + 1\}\alpha_2 + \{(\deg[C(CCC)]^{-1/2})^2 + 1\}\alpha_3 + \{(\deg[C(CCCC)]^{-1/2})^2 + 1\}\alpha_4 \\ S = & \{[L_1]^2 + 1\}\alpha_1 + \{[L_2]^2 + 1\}\alpha_2 + \{[L_3]^2 + 1\}\alpha_3 + \{[L_4]^2 + 1\}\alpha_4 \end{aligned} \quad (13)$$

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

#### 4.4. Molecular weight

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

$$M_w = 12(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \{2(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + 2\} \quad (14)$$

It is evident from the equations presented that the four molecular descriptors can be written as linear combination of the number of occurrences of the four atomic 1-signatures. Thus, both QSRs ( $T_{MD}$  and  $T_S$ ) will predict the same normal boiling point for the same compounds. That this is true can be seen in Fig. 3, where we plot the prediction of both models as a function of the experimental data.

### 5. QSAR on HIV-1 protease inhibitors

As a first attempt to probe the utility and robustness of QSARs based on signature, a large, biological system was chosen for a QSAR study. The system under investigation was 121 inhibitors (training set) of HIV-1 protease taken from six literature sources. The experimental activity of the ligands against the HIV-1 protease was reported in  $pIC_{50}$  ( $pIC_{50} = 9 - \log_{10} IC_{50}$ ) and spanned seven-orders of magnitude in activity. A test set of nine compounds that spanned the  $pIC_{50}$  of the training set was chosen with at least one test set compound coming from each of the literature sources used in the training set. Two QSARs, one using atomic 1-signatures and one using atomic 2-signatures, were developed. To provide some perspective on the signature results,

a QSAR created using a commercially available descriptor package, Molconn-Z from eduSoft LC, was used [14].

#### 5.1. Training set/test set

Since the QSAR study is only part of the overall coverage of this paper, we will not provide all of the structures of the ligands used here. Rather, we will provide a representative sample, but we will note specifically the compounds used from a particular literature source. This is found in Table 3. A complete list of all 121 compounds used for the training set as well as the nine compounds used in the test set are provided in Table 4.

#### 5.2. QSAR methodology

A standard forward-stepping, MLR [15] was performed in order to determine the best-fit parameters for each model. To provide a comparison between the three QSARs developed, a fixed number of parameters were chosen for each model. This value was set at seven descriptor parameters plus a constant term for a total of eight parameters in the MLR.

#### 5.3. QSAR using atomic signatures

The initial step in developing a QSAR using atomic signatures is to determine the number and type of atomic signatures in the training set as well as the value of each signature (i.e. its number of occurrences) for all the compounds in the training set. This was accomplished using an in-house C program that took a specified input format and created the molecular signature of the desired height. For the 121 compounds used in this study, there were 137 unique atomic signatures of height-1 and 548 unique atomic signatures of height-2. It was upon these initial descriptors that the MLR was performed.

#### 5.4. QSAR using Molconn-Z descriptors

Molconn-Z is a commercial package offered from eduSoft, LC [14] that contains many of the popular descriptors such as the various connectivity indices, Kier–Hall shape indices, fragments, electrotopological states, information indices, among others. In total, we started with 163 descriptors and it was upon this number that the MLR was performed.

#### 5.5. HIV-1 protease inhibitor QSAR results

For each model there were seven descriptors in the final QSAR. The relevant statistics for the three models are provided in Table 5, where  $R^2$  is the square of the correlation coefficient,  $F$  is the Fischer ratio and  $s$  is the root mean square error. Overall, the two QSARs created from the signatures were able to correlate the training set data with errors only slightly larger than that from the QSAR obtained

Table 3  
The scaffolds of the HIV-1 protease inhibitors used in this study

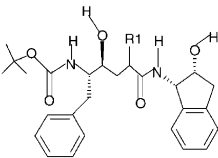

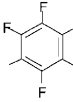

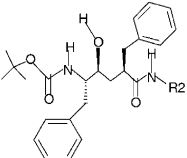



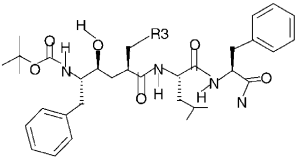


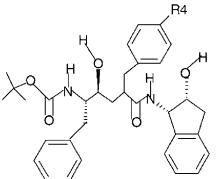



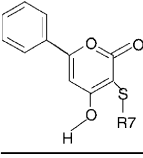


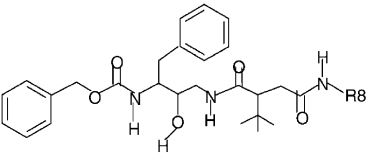


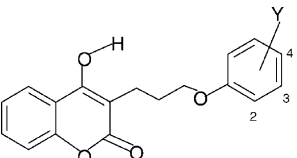



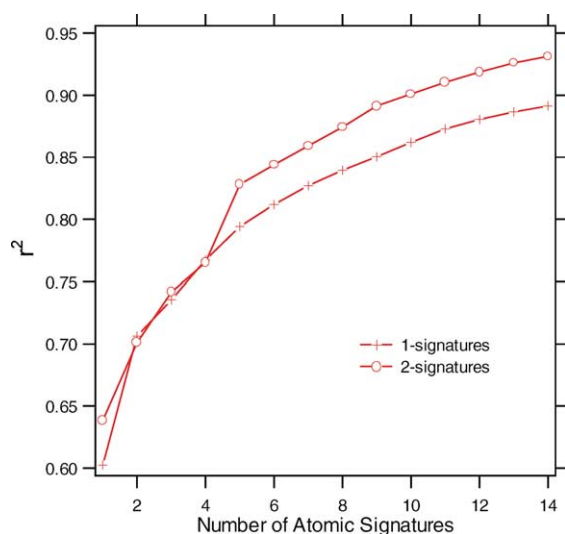
Scaffold	R1	pIC <sub>50</sub>	Source
		9.60	[29]
		9.22	
		8.27	
		7.41	[29]
		8.02	
		4.52	
		9.36	[30]
		8.92	
	H	8.22	
		9.70	[31]
		10.04	
		8.69	
		5.52	[32]
		5.89	
		6.8	[33]
		7.8	
	Y		
	(2) 	4.3	[34]
	(3) 	6.1	
	(4) 	5.1	

Table 4

A complete list of the compounds 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)		pIC <sub>50</sub> (test set)
	Training set	Test set	
[29]	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
[30]	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
		46	8.854
[34]	1, 12, 13, 16, 18, 19, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38	15	4.886
		27	5.638
[31]	19, 22, 23, 24, 29, 31, 32, 33, 34, 36, 37, 38, 39, 40, 43, 44, 45	30	9.260
		35	8.745
[32]	1, 2, 3, 5, 6, 7	4	5.851
[33]	9, 10, 12, 13, 14, 15	11	7.824

Fig. 4. A comparison of the  $r^2$  values from QSARs created using height-1 signatures to that from height-2 signatures.

using the Molconn-Z parameters. To test the predictive ability of the models, these QSARs were used on a test set not used in the parameter fitting. These results are provided in Fig. 3 and Table 5. Very similar results were obtained from the three models, though the QSAR developed using the height-1 signatures slightly outperformed the other models.

The larger initial set of signatures available for height-2 relative to height-1 (548 versus 137) coupled with the more probative value of the longer paths for height-2 relative to height-1 should allow the researcher to correlate the same data set using height-2 signatures better than with height-1 signatures. To explore this feature, we examined the evolution of the square of the correlation coefficient ( $r^2$ ) as a

function of the number of signatures used in a MLR (using the forward-stepping method). Fig. 4 provides this information and we can see the trend we expect, which is a larger  $r^2$  value for the height-2 signatures using the same number of descriptors.

## 6. Focused library design from inverse-QSAR

Once a QSAR has been determined using signature, the next step is to use this equation to solve for the sets of atomic signatures corresponding to a desired activity. For example, given a QSAR of the form  $A(\vec{\alpha}) = \sum_k b_k \alpha_k$ , where  $A$  is an activity function,  $\alpha_k$  are the number of occurrences of a particular atomic signature  $k$  and  $b_k$  are the corresponding regression coefficients, the goal is to find the roots of the equation  $f$  (i.e. the sets of the solution vector  $\vec{\alpha}$ ) such that given a target value for the activity of interest,  $A^{\text{TAR}}$ , the equation  $f(\vec{\alpha}) = \sum_k b_k \alpha_k - A^{\text{TAR}}(\vec{\alpha}) = 0$  is satisfied. Algorithms to both enumerate [16] and sample [17] the chemical structures corresponding to given solution vectors  $\vec{\alpha}$  have already been developed. If the QSAR developed is from atomic 1-signatures, the number of structures determined from the solution space will be necessarily larger than if the QSAR developed is from signatures of a greater height. Therefore, the researcher has the ability to control, to some extent, the size of the library of structures corresponding to a given  $A^{\text{TAR}}$ .

## 7. Concluding remarks

In this work we presented the concept of signature and how signature notation can be used to denote the molecular graph of a compound. We showed how some popular TIs used in QSAR modeling can be written as functions of the atomic signatures. A simple example using aliphatic alkanes was provided to show an equivalence between a QSAR created using four TIs and that from four height-1 signatures.

We tested the height-1 and height-2 signatures for use as descriptors in a QSAR by correlating the pIC<sub>50</sub> values of

Table 5

The regression results for the three QSARs on HIV-1 protease inhibitors

Model	$r^2$	$F$	$s$ (training)	$s$ (test)
Height-1 signatures	0.8272	77.3	0.8021	0.7677
Height-2 signatures	0.8591	98.5	0.7242	0.8259
Molconn-Z	0.8682	106.4	0.7004	0.7759



121 HIV-1 protease inhibitors. This result was compared to a QSAR developed from a commercial package, Molconn-Z. All three QSARs provided similar results for the training set as well as the test set. Though this was by no means an exhaustive test, such a result indicates that useful QSARs can be created with signatures as descriptors. Future work along this line is ongoing to explore the robustness of using signatures as molecular descriptors on large data sets.

The main advantage of signature versus other descriptors is its readiness for inverse problems. While in classical QSAR analysis, inverse-QSAR must be carried out for each descriptor one at a time and the inverse problem remains unsolved for most descriptors, with signature-based QSAR analysis, algorithms have already been developed to enumerate and/or sample the molecular structures corresponding to a given signature. Thus, the success of signature in solving inverse problems should only depend on how well signature can be used to predict activity/property which, we believe, has been demonstrated in the present work.

## Acknowledgements

Funding for this work was provided by the US Department of Energy and Sandia National Laboratories under grant no. DE-AC04-76DP00789. RSP would like to acknowledge additional support from the Center for the Management, Utilization and Protection of Water Resources at Tennessee Technological University.

## References

- [1] N. Trinajstić, Chemical Graph Theory, 2nd Edition, CRC Press, Boca Raton, FL, 1992.
- [2] D. Goldman, S. Istrail, G. Lancia, A. Piccolboni, B. Walenz, Algorithmic strategies in combinatorial chemistry, *SODA* 11 (2000) 275–284.
- [3] M.I. Skvortsova, I.I. Baskin, O.L. Slovokhotova, V.A. Palyulin, N.S. Zefirov, Inverse problem in QSAR/QSPR studies for the case of topological indices characterizing molecular shape (Kier indices), *J. Chem. Inform. Comput. Sci.* 33 (1993) 630–634.
- [4] I.I. Baskin, M.I. Skvortsova, I.V. Stankevich, N.S. Zefirov, On the basis of invariants of labeled molecular graphs, *J. Chem. Inform. Comput. Sci.* 35 (1995) 527–531.
- [5] D. Weininger, A. Weininger, SMILES: a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inform. Comput. Sci.* 28 (1988) 31–40.
- [6] J.-L. Faulon, Stochastic generator of chemical structure. 1. Application to the structure elucidation of large molecules, *J. Chem. Inform. Comput. Sci.* 34 (1994) 1204–1218.
- [7] J.-L. Faulon, D.P. Visco Jr., R.S. Pophale, The Signature Molecular Descriptor. 1. Extended Valence Sequences and Topological Indices, *J. Chem. Inform. Comput. Sci.*, 2002, submitted.
- [8] J.E. Hopcroft, R.E. Tarjan, Isomorphism of Planar Graphs, Complexity of Computer Computation, Plenum Press, New York, 1972, pp. 131–152.
- [9] L. Kucera, Combinatorial Algorithms, Adam Hilger, Bristol, 1990.
- [10] L.B. Kier, L.H. Hall, Molecular Structure Description, Academic Press, San Diego, CA, 1999.
- [11] S.C. Basak, B.D. Gute, G.D. Grunwald, A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient, *J. Chem. Inform. Comput. Sci.* 36 (1996) 1054–1060.
- [12] G. Espinosa, D. Yaffe, Y. Cohen, A. Arenas, F. Giral, Neural network based quantitative structural property relations (QSPRs) for predicting boiling points of aliphatic hydrocarbons, *J. Chem. Inform. Comput. Sci.* 40 (2000) 859–879.
- [13] L.H. Hall, C.T. Story, Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks, *J. Chem. Inform. Comput. Sci.* 36 (1996) 1004–1014.
- [14] L.H. Hall, Molconn-Z Hall Associates Consulting, Quincy, MA, 1991.
- [15] N.R. Draper, H. Smith, Applied Regression Analysis, 2nd Edition, Wiley, New York, 1981.
- [16] J.-L. Faulon, On using graph-equivalent classes for the structure elucidation of large molecules, *J. Chem. Inform. Comput. Sci.* 32 (1992) 338–348.
- [17] J.-L. Faulon, Stochastic generator of chemical structure. 2. Using simulated annealing to search the space of constitutional isomers, *J. Chem. Inform. Comput. Sci.* 36 (1996) 731–740.
- [18] M. Randić, On the characterization of molecular branching, *J. Am. Chem. Soc.* 97 (1975) 6609–6615.
- [19] L.B. Kier, L.H. Hall, Derivation and significance of valence molecular connectivity, *J. Pharm. Sci.* 70 (1981) 583–589.
- [20] M. Randić, On computation of optimal parameters for multivariate analysis of structure-property relationship, *J. Comput. Chem.* 12 (1991) 970–980.
- [21] L.H. Hall, L.B. Kier, The Molecular and Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling, Reviews in Computational Chemistry, VCH Publishers, New York, 1991, pp. 367–422.
- [22] M. Randić, Graph valence shells as molecular descriptors, *J. Chem. Inform. Comput. Sci.* 41 (2001) 627–630.
- [23] J.R. Platt, Influence of neighbor bonds on additive bond properties in paraffins, *J. Chem. Phys.* 15 (1947) 419.
- [24] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* 69 (1947) 17–20.
- [25] A.T. Balaban, Topological index  $J$  for heteroatom-containing molecules taking into account periodicities of element properties, *Math. Chem. (MATCH)* 21 (1986) 115–122.
- [26] D. Bonchev, The overall Wiener index—a new tool for characterization of molecular topology, *J. Chem. Inform. Comput. Sci.* 41 (2001) 582–592.
- [27] M. Randić, Novel shape descriptors for molecular graphs, *J. Chem. Inform. Comput. Sci.* 41 (2001) 607–613.
- [28] W.G. Mallard, P.J. Linstrom, NIST Chemistry Webbook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg, MD, 2000.
- [29] C. Perez, M. Pastor, A.R. Ortiz, F. Gago, 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.* 41 (1998) 839–852.
- [30] S.D. Young, L.S. Payne, W.J. Thompson, N. Gaffin, T.A. Lyle, S.F. Britcher, S.L. Graham, T.H. Schultz, A.A. Deana, P.L. Darke, J. Zugay, W.A. Schleif, J.C. Quintero, E.A. Emini, P.S. Anderson, J.R. Huff, 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.* 35 (1992) 1702–1709.
- [31] W.J. Thompson, P.M.D. Fitzgerald, M.K. Holloway, E.A. Emini, P.L. Darke, B.M. McKeever, W.A. Schleif, J.C. Quintero, J. Zugay, T.J. Tucker, J.E. Schwering, C.F. Homnick, J. Nunberg, J.P. Springer, J.R. Huff, 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.* 35 (1992) 1685–1701.

- [32] J.V.N. Vara Prasad, K.S. Para, E.A. Lunney, D.F. Ortwine, J.J.B. Dunbar, D. Ferguson, P.J. Tummino, D. Hupe, B.D. Tait, J.M. Domagala, C. Humblet, T.N. Bhat, B. Liu, D.M.A. Guerine, E.T. Baldwin, J.W. Erickson, T.K. Sawyer, Novel series of achiral, low molecular weight, and potent HIV-1 protease inhibitors, *J. Am. Chem. Soc.* 116 (1994) 6989–6990.
- [33] P.L. Beaulieu, D. Wernic, A. Abraham, P.C. Anderson, T. Bogri, Y. Bousquet, G. Croteau, I. Guse, D. Lamarre, F. Liard, W. Paris, D. Thibeault, S. Pav, L. Tong, Potent HIV protease inhibitors containing a novel (hydroxyethyl)amide isostere, *J. Med. Chem.* 40 (1997) 2164–2176.
- [34] E.A. Lunney, S.E. Hagen, J.M. Domagala, C. Humblet, J. Kosinski, B.D. Tait, J.S. Warmus, M. Wilson, D. Ferguson, D. Hupe, P.J. Tummino, E.T. Baldwin, T.N. Bhat, B. Liu, J.W. Erickson, A novel non-peptide HIV-1 protease inhibitor: elucidation of the binding mode and its application in the design of related analogs, *J. Med. Chem.* 37 (1994) 2664–2677.
- [35] S.L. Mayo, B.P. Olafson, I. Goddard, W.A. Schleif, Dreiding: a generic force field, *J. Phys. Chem.* 94 (1990) 8897–8909.