ELSEVIER

A widely applicable set of descriptors

Paul Labute

Chemical Computing Group Inc., Montreal, Quebec, Canada

Three sets of molecular descriptors computable from connection table information are defined. These descriptors are based on atomic contributions to van der Waals surface area, log P (octanol/water), molar refractivity, and partial charge. The descriptors are applied to the construction of QSAR/QSPR models for boiling point, vapor pressure, free energy of solvation in water, solubility in water, thrombin/trypsin/factor Xa activity, blood-brain barrier permeability, and compound classification. The wide applicability of these descriptors suggests uses in QSAR/QSPR, combinatorial library design, and molecular diversity work.

Keywords: molecular descriptors, QSAR, molecular diversity

INTRODUCTION

The pioneering work of Hansch et al.[1]and Leo et al.[2] was an attempt to describe biological phenomena in a "language" consisting of a small set of physical molecular properties, in particular, logP (octanol/water), pK_a, and molar refractivity. Early QSAR efforts centered on deriving linear regression[3] relationships between such descriptors and biological activity. Not surprisingly, these models (first-order relationships among few descriptors) were limited to analogue series. Subsequent efforts sought to increase the applicability of linear models by introducing more[4–8] (and more) descriptors. As the number of descriptors available increased, methods were required to automatically select the appropriate descriptors from a large pool,[9] often several hundred. Unfortunately, as epidemiologists are well aware: *as the number of descriptors increases, so too does the likelihood of finding chance relationships in the data.* This has led to an increased reliance on validation methods to identify spurious models (e.g., leave-one-out or *k*-fold cross-validation). Of course, it is assumed that such validation methods can indeed identify spurious models in the first place (a fact which cannot be proved). The situation is even more confusing with combined variable selection and model validation methods such as CART[10]: the introduction of validation into the regression determination procedure may destroy the significance (if any) of the validation. Notwithstanding the use of complicated regression methods and descriptor selection procedures, the consistent production of effective QSAR models remains elusive. For this reason, we have chosen to return to the thinking of Hansch and Leo: by fixing a relatively small set of descriptors for use in many (hopefully all) situations we can, perhaps, a) reduce the problems of variable selection, and b) consistently produce meaningful QSAR models. Moreover, we will attempt to stay true to the Hansch and Leo concepts in order to make direct use of these well thought-out descriptors.

The idea of using a fixed collection of descriptors in QSAR is related to the definition of a "chemistry space" for use in molecular diversity studies. In such work, a compound is mapped to a *k*-dimensional vector that is used as a surrogate when comparing compounds. Validating a chemistry space can be difficult, especially when it is proposed for diversity analysis. Two different chemistry spaces generally will induce two different notions of diversity. Unless one has a reference diversity metric for comparison, it will not be clear which space is better. Often, cluster analysis is used to justify a chemistry space: if compounds with similar biological behavior cluster together in the proposed space, then it seems reasonable to conclude that the chemistry space is good. Unfortunately, these results often depend on the choice of clustering algorithm. An alternative to cluster-based justification is QSAR-based justification: if a collection of descriptors can be used to construct reasonable models of many properties of interest, using many modeling techniques, it seems reasonable to conclude that the chemistry space induced by the descriptors is meaningful for diversity analysis. This has been the case with other efforts to construct meaningful chemistry spaces. For example, the BCUT[11] or WHIM[12] descriptor collections were designed for one purpose (e.g., diversity) but quickly found application in the other[13,14] (e.g., QSAR) and vice versa. In the present approach, we will adopt the QSAR-based justification of the chemistry space induced by the descriptors defined herein.

This article is divided into several sections. In the Methods section, we define three collections of descriptors. In the subsequent sections, we apply these descriptors to the prediction of several physical and biological properties and describe further applications of the descriptors. We draw conclusions in the final section.

METHODS

Approximate Surface Area

The surface area of an atom in a molecule is the amount of surface area of that atom not contained in any other atom of the molecule (Figure 1). If we take the shape of each atom to be a

Corresponding author: Paul Labute, Chemical Computing Group Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada H3B 3X3. Tel.: 514-393-1055; fax: 514-874-9538. *E-mail address*: paul@chemcomp.com (P. Labute).

Journal of Molecular Graphics and Modelling 18, 464–477, 2000

655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PII S1093-3263(00)00068-1

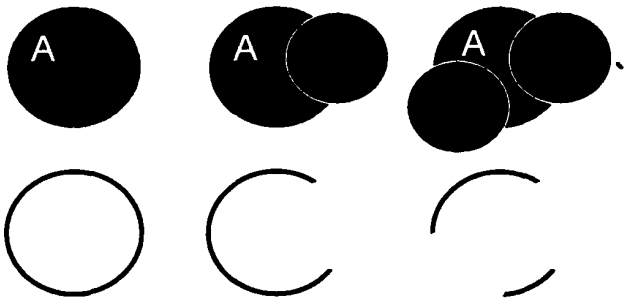


Figure 1. Assuming spherical atoms, the surface area of atom A *is the amount of surface area not contained in other atoms.*

sphere with radius equal to the van der Waals radius, we obtain the van der Waals surface area (VSA) for each atom. The sum of the VSA of each atom gives the molecular VSA.

Consider two spheres A and B with radii r and s, respectively, and centers separated by distance d. The amount of surface area of sphere A not contained in sphere B, denoted by V_A, is given by:

$$V_A = \begin{cases} 4\pi r^2 - \pi r d^{-1}[s^2 - (r-d)^2] & \text{if } |r-s| < d < r+s \\ 4\pi r^2 & \text{otherwise} \end{cases}$$

The case of more than two spheres is more complicated because a portion of sphere A may be contained in several other spheres. However, we will neglect this complication (in the hope that the error introduced will not be large). Thus, we approximate the VSA for sphere A with n neighboring spheres B_i with radii s_i and at distances d_i as:

$$V_A = 4\pi r^2 - \pi r \sum_{i=1}^{n} \frac{s_i^2 - (r-d_i)^2}{d_i} \delta(|r-s_i| < d_i < r + s_i)$$

where the generalized delta function, $\delta(P)$, adopts a value of 1 if the condition P is satisfied and 0 otherwise. This formula is similar to the pairwise approximations used in approximate overlap volume calculations[15] and approximate surface area calculations for generalized Born implicit solvent models.[16] Consider, now, a molecule of n atoms each with van der Waals radius R_i and let B_i denote the set of all atoms bonded to atom i. We will neglect the effect of atoms not related by a bond and define the VSA for atom i, denoted by V_i, to be:

$$V_i = 4\pi R_i^2 - \pi R_i \sum_{j \in B_i} \frac{R_j^2 - (R_i - d_{ij})^2}{d_{ij}}$$

$$d_{ij} = \min\{\max\{|R_i - R_j|, b_{ij}\}, R_i + R_j\}$$

where b_{ij} is the ideal bond length between atoms i and j. Thus, the VSA for each atom can be calculated from connection table information alone assuming a dictionary of van der Waals radii and ideal bond lengths. In the present work, the radii are derived from MMFF94[17] with certain modifications for polar hydrogen atoms and are presented in Table 1. In the results to follow, the ideal bond length b_{ij} between atoms i and j was calculated according to the formula $b_{ij} = s_{ij} - o_{ij}$, where s_{ij} is a reference bond length derived from MMFF94 parameters (Table 2) that depends on the two elements involved and o_{ij} is a correction that depends on the bond order: 0 for single, 0.1 for aromatic, 0.2 for double, and 0.3 for triple. Finally, the approximate VSA for an entire molecule is just the sum of the V_i for each atom i in the molecule.

Table 1. Van der Waals radii used for VSA calculations

H-O	0.8	O (other)	1.779
H-N,P	0.7	F	1.496
H-other	1.485	P	2.287
C	1.950	S	2.185
N	1.950	Cl	2.044
O (oxide)	1.810	Br	2.166
O (acid)	2.152	I	2.358

Radii are given in Angstroms.

To test the accuracy of the approximate VSA calculation, a database of 1,947 small organic molecules was assembled. The molecular weights fell in the range (300,1600). For each molecule, a 3D extended conformation was calculated using the 2D to 3D converter of MOE[18] version 1999.05. The MMFF94 parameter set was used to energy minimize the structures to an RMS gradient less than 0.001. Using the radii from Table 1, the molecular VSA was calculated using a dot-based method: each atom was surrounded with a large number of points on its surface. Points inside any other atom were eliminated, and the

Table 2. Reference bond lengths used for VSA calculation

Br	Br	2.540	F	F	1.280
Br	C	1.970	F	H	0.870
Br	Cl	2.360	F	I	2.040
Br	F	1.850	F	N	1.410
Br	H	1.440	F	O	1.320
Br	I	2.650	F	P	1.500
Br	N	1.840	F	S	1.640
Br	O	1.580	H	I	1.630
Br	P	2.370	H	N	1.010
Br	S	2.210	H	O	0.970
C	C	1.540	H	P	1.410
C	Cl	1.800	H	S	1.310
C	F	1.350	I	I	2.920
C	H	1.060	I	N	2.260
C	I	2.120	I	O	2.140
C	N	1.470	I	P	2.490
C	O	1.430	I	S	2.690
C	P	1.850	N	N	1.450
C	S	1.810	N	O	1.460
Cl	Cl	2.310	N	P	1.60
Cl	F	1.630	N	S	1.760
Cl	H	1.220	O	O	1.470
Cl	I	2.560	O	P	1.570
Cl	N	1.740	O	S	1.570
Cl	O	1.410	P	P	2.260
Cl	P	2.010	P	S	2.070
Cl	S	2.070	S	S	2.050

Lengths are given in Angstroms.

remaining number of points were used to estimate the exposed surface area. Conformational analysis of several randomly chosen flexible molecules revealed that the dot-based van der Waals surface area of individual conformations differ < 2%. From this observation, we decided that it was reasonable to compare the approximate VSA to the dot-based surface area of a single extended conformation of each molecule. A scatter plot of the results is shown in Figure 2. The correlation coefficient was $r^2 = 0.9666$ and the relative error was <10%. Most of the errors occurred for the larger molecules and in molecules with many atoms in fused ring systems. No systematic error appeared to be present (other than the increase in error with molecular weight), and it was concluded that the approximate VSA calculation was reasonably accurate. We then made the inference that the individual contributions to the approximate VSA also were reasonably accurate.

Thus, we have defined V_i, the contribution of atom i to the approximate VSA of a molecule. This contribution is reasonably accurate, but it has the advantage that it can be calculated much more rapidly than a 3D VSA contribution and without a 3D conformation (just connection table information). The approximate molecular VSA is very much a $2\frac{1}{2}$ descriptor: it is (highly correlated to) a conformation independent 3D property that requires only 2D connection information.

Descriptors

Suppose that for each atom i in a molecule, we are given a numerical property P_i. Our fundamental idea is to create a descriptor for a specific range $[u,v)$[19] of the property values P. This descriptor will be the sum of the atomic VSA contributions of each atom i with P_i in $[u,v)$. More precisely, we define the quantity P−VSA(u,v) to be:

$$\text{P\_VSA}(u, v) = \sum_i V_i \delta(P_i \in [u, v])$$

where V_i is the atomic contribution of atom i to the VSA of the molecule (defined in the previous section). We now define a set of n descriptors associated with the property P as follows:

$$\text{P\_VSA}_k = \sum_i V_i \delta(P_i \in [a_{k-1}, a_k]) \quad k = 1,2,\ldots n$$

where $a_0 < a_k < a_n$ are interval boundaries such that $[a_0,a_n)$ bound all values of P_i in any molecule. Figure 3 is an example of the calculation of a hypothetical set of descriptors from a chemical structure. Each VSA-type descriptor can be characterized as *the amount of surface area with P in a certain range.* If, for a given set of descriptors, the interval ranges span all values, then the sum of the descriptors will be the VSA of the molecule. Therefore, the VSA-type descriptors correspond to a subdivision of the molecular surface area.

Wildman and Crippen's recent methods[20] for calculating logP (octanol/water) and molar refractivity (MR) provide a good basis for VSA analogues of logP and MR because these methods were parameterized with atomic contributions in mind. Both methods assign a numerical contribution to each atom in a molecule. We implemented both methods in the SVL programming language of MOE. To determine the interval boundaries, we obtained statistics on a database of 44,795 small organic compounds from the Maybridge[21] catalog (the entire October 1998 HTS database less 2,000 randomly selected compounds used for testing). We chose the interval boundaries so that the resulting intervals were equally populated over the database (resulting in nonuniform boundaries). This led to 10 descriptors for logP and 8 descriptors for MR. The respective interval boundaries for logP were $(-\infty, -0.4, -0.2, 0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, \infty)$. The interval boundaries for MR were $(0, 0.11, 0.26, 0.35, 0.39, 0.44, 0.485, 0.56, \infty)$. The third set of descriptors we will define is based on atomic partial charge: the P property for each atom is the partial charge. We chose the Gasteiger[22] (PEOE) method of calculating partial charges, which is based on the iterative equalization of atomic orbital electronegativities. This method was implemented in MOE using the SVL programming language. Fourteen descriptors resulted from the use of uniform interval boundaries in $(-\infty, -0.3, -0.25, -0.20, -0.15, -0.10, -0.05, 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, \infty)$.

We have thus defined three sets of molecular descriptors:

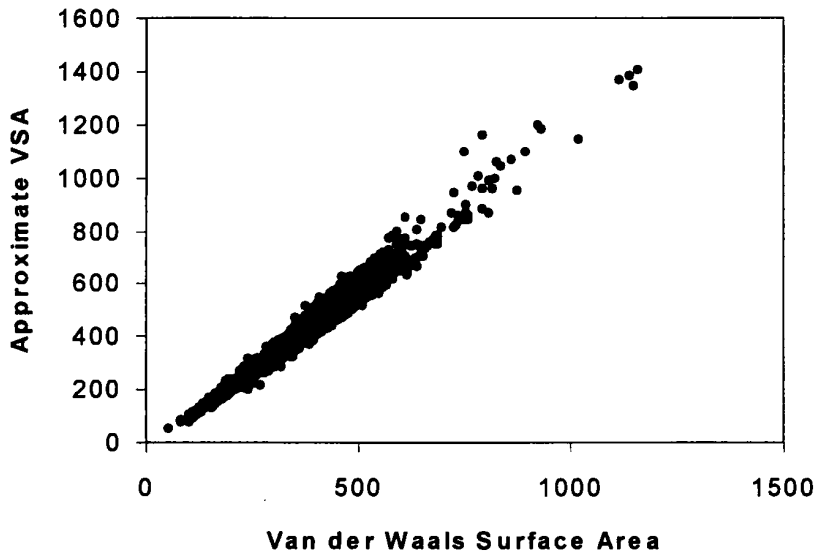
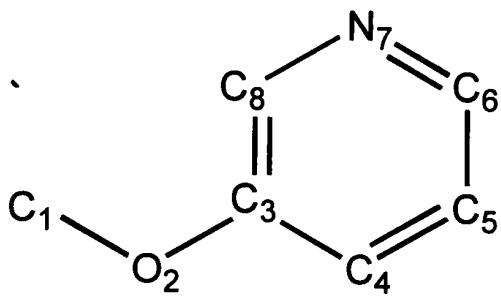


Figure 2. Scatter plot comparing a 3D surface area (x-axis) calculation and the approximate VSA calculation (y-axis).



$P_1 = 2.4$ $V_1 = 9.2$
$P_2 = 1.2$ $V_2 = 6.3$
$P_3 = 4.5$ $V_3 = 2.2$
$P_4 = 5.9$ $V_4 = 4.5$
$P_5 = 5.7$ $V_5 = 4.5$
$P_6 = 3.1$ $V_6 = 4.4$
$P_7 = 0.2$ $V_7 = 4.6$
$P_8 = 3.9$ $V_8 = 4.4$

Figure 3. Calculation of a hypothetical set of five VSA descriptors $D_1,...D_5$ based on a property P. The chemical structure consists of eight atoms, each with the given property value P_i and VSA contribution V_i.

$D_1 = P\_VSA(0,2) = V_2+V_7 = 6.3 + 4.6 = 10.9$
$D_2 = P\_VSA(2,3) = V_1 = 9.2$
$D_3 = P\_VSA(3,4) = V_6+V_8 = 4.4 + 4.4 = 8.8$
$D_4 = P\_VSA(4,5) = V_3 = 2.2$
$D_5 = P\_VSA(5,6) = V_4+V_5 = 4.5 + 4.5 = 9.0$

- SlogP−VSA$_k$ (10) intended to capture hydrophobic and hydrophilic effects either in the receptor or on the way to the receptor
- SMR−VSA$_k$ (8) intended to capture polarizability
- PEOE−VSA$_k$ (14) intended to capture direct electrostatic interactions.

Each of these descriptor sets is derived from, or related to, the Hansch and Leo descriptors, with the expectation that they would be widely applicable. Taken together, the VSA descriptors define, nominally, a 10 + 8 + 14 = 32 dimensional chemistry space.

RESULTS

Self-Correlation

A database of 2,000 structures from the Maybridge HTS database (not used in the statistics collection) was selected at random to test the correlation between the descriptors we have defined. Figure 4 presents the correlation between each of the eight MR descriptors. The largest r value of 0.6 (r^2 = 0.36) appeared once; the remaining pairs exhibited r values <0.27 (r^2 = 0.07). Thus, MR descriptors are, for the most part, weakly correlated with each other. Figure 5 presents the correlation between each of the 10 logP descriptors. The largest r value of 0.42 (r^2 = 0.18) appeared once; the remaining pairs exhibited r values <0.27 (r^2 = 0.07). Thus, the logP descriptors are, for the most part, weakly correlated with each other. Figure 6 presents the correlation between each of the 14 PEOE−VSA descriptors on the same 2,000 compounds used to measure correlation in the MR and logP descriptors. The largest r value was 0.65 (r^2 = 0.42). Thus, the PEOE−VSA descriptors are, for the most part, weakly correlated with each other. Each of the three descriptor sets described herein has shown weak intraset correlation; however, a question remains about interset correlation. Figure 7 presents the full correlation matrix containing r values between the PEOE−VSA, SlogP−VSA, and SMR−VSA descriptors. Generally, the intercorrelation is weak; however, seven r values were >0.7 (r^2 = 0.49). At first glance, the sets seem to exhibit higher correlation than in the intraset cases. However, it must be remembered that, for a given molecule, each PEOE−VSA, SlogP−VSA and SMR−VSA descriptor collection sums to the VSA of the molecule; hence, there is one less dimension than the nominal 14 + 10 + 8 = 32.

SMR_VSA1		12	21	20	18	1	20	22
SMR_VSA2	12		1	60	21	-9	6	-1
SMR_VSA3	21	1		3	12	-6	3	-3
SMR_VSA4	20	60	3		27	12	11	17
SMR_VSA5	18	21	12	27		-11	3	-15
SMR_VSA6	1	-9	-6	12	11		12	-13
SMR_VSA7	20	6	3	11	3	12		8
SMR_VSA8	22	-1	-3	17	-15	-13	8	

Figure 4. Correlation matrix of the eight SMR−VSA descriptors. Each entry in the matrix is the r *value (signed) in percent. The order of the columns is the same as the order of the rows.*

Correlation with Other Descriptors

To test the extent to which the VSA descriptors encode other popular descriptors, a database of 1,932 small organic compounds[23] with molecular weights in the range (28,800) was assembled. For each molecule, the SlogP−VSA, SMR−VSA, and PEOE−VSA descriptors were calculated, as well as 64 other descriptors, all of which were calculated using MOE version 1999.05. For each of the latter 64 descriptors, a principal components regression was calculated to produce a linear model for each descriptor as a function of the 32 VSA descriptors. In all cases, 31 of 32 principal components were retained. The results are summarized in Figure 8. Out of the 64 descriptors tested, 32 showed an r^2 of 0.90 or better; 49 had an r^2 of 0.80 or better; and 61 showed an r^2 of 0.5 or better. These results suggest that the 32 VSA descriptors encode much of the information contained in most of the 64 popular descriptors.

SlogP_VSA1		2	42	1	5	2	-9	-16	2	-2
SlogP_VSA2	2		14	-6	12	1	-2	4	-14	3
SlogP_VSA3	42	14		4	1	4	-9	-16	23	27
SlogP_VSA4	1	-6	4		0	1	1	12	-4	2
SlogP_VSA5	5	12	1	0		6	0	13	15	8
SlogP_VSA6	2	1	4	1	6		-6	7	-14	-11
SlogP_VSA7	-9	-2	-9	1	-	-6		21	-10	14
SlogP_VSA8	-16	4	-16	12	13	7	21		-16	-22
SlogP_VSA9	2	-14	23	-4	15	-14	-10	-16		6
SlogP_VSA10	-2	3	27	2	8	-11	14	-22	6	

Figure 5. Correlation matrix of the 10 SlogP−VSA descriptors. Each entry in the matrix is the r *value (signed) in percent. The order of the columns is the same as the order of the rows.*

Free Energy of Solvation in Water

The free energy of solvation in water is the change in free energy upon transfer from gas phase to water phase. A database of 291 small organic molecules with associated experimental free energies of solvation was created from the literature.[24] Even though the SlogP−VSA descriptors are based on a transfer free energy, it must be pointed out that the descriptor values themselves are surface areas; hence, it is reasonable to attempt to create such a model. For each structure the PEOE−VSA, SlogP−VSA and SMR−VSA descriptors were calculated and a principal components regression was calculated using MOE. Descriptors with small (normalized) coefficients were discarded until only the PEOE−$VSA_{2,8,13}$, SlogP−$VSA_{1,2,3,4,6,8,10}$, and SMR−$VSA_{2,8}$ descriptors remained. A principal components regression was calculated using the remaining 12 descriptors, and the resulting r^2 was 0.90 with an RMSE of 0.78 kcal/mol (scatter plot in Figure 9). The leave-one-out cross-validated r^2 was 0.89 with an RMSE of 0.82 kcal/mol. A single leave-100-out cross-validation test produced a prediction r^2 of 0.88 on the 100 randomly chosen compounds left out of the training.

Boiling Point

We assembled a database of 298 small organic molecules[25] with associated experimental boiling points taken from the CRC Handbook. For each molecule, the SlogP−VSA and SMR−VSA descriptors were calculated with MOE. A principal components regression was estimated to create a linear model of the boiling point as a function of the 18 descriptors resulting in an r^2 of 0.93. A scatter plot of the predicted vs experimental boiling points showed a quadratic relationship. We then created a linear model of the square of the boiling point as a function of the same descriptors. Upon taking square roots of the predicted squared boiling point, the resulting predictions were well correlated with the experimental values with an r^2 of 0.96 and RMSE of 15.53 Kelvin. The leave-one-out cross-validated r^2 was 0.94 with an RMSE of 21.37 Kelvin. Figure 10 is a scatter plot of the predicted and experimental boiling points. A single leave-100-out cross-validation test produced a prediction r^2 of 0.94 on the 100 randomly chosen compounds left out of the training.

Blood-Brain Barrier Permeability

Permeability at the blood-brain barrier is an important factor in the design of safer and more effective therapeutic compounds active in the central nervous system. For a given compound, the experimental determination of the ratio of concentration in the brain and concentration in the blood is a time-consuming and expensive process requiring appropriate amounts of the pure compound, often in radiolabeled form. We assembled a collection of 75 compounds and experimental log BB concentration ratios from the literature.[26] Acids were deprotonated and bases were protonated. The PEOE−VSA, SlogP−VSA, and SMR−VSA descriptors were calculated for each structure. A principal components regression was performed to estimate a linear model of log BB as a function of the descriptors. Descriptors with small (normalized) coefficients were discarded until only the PEOE−$VSA_{4,9,13}$, SlogP−$VSA_{1,2,3,5,8,9}$, and SMR−VSA_{1-5} descriptors remained. A principal components regression was calculated using the remaining 15 descriptors, and the resulting r^2 was 0.83 with an RMSE of 0.32. The leave-one-out cross-validated r^2 was 0.73 with an RMSE of 0.43. Figure 11 is a scatter plot of the predicted and experimental log BB concentration ratios.

Solubility in Water

A collection of 1,438 small organic molecules with associated experimental water solubilities was assembled from the Syracuse Research Corporation[27] (SRC) 1999 physical property database. The SRC database contains experimental and estimated solubilities; accordingly, we selected all compounds with experimental values measured at 25°C. The solubility measurements were converted to a log scale. The PEOE−VSA, SlogP−VSA, and SMR−VSA descriptors were calculated for each structure (a total of 32 descriptors). A principal components regression model was calculated (31 principal components were retained), and the resulting r^2 was 0.75 with an RMSE of 2.4. The leave-one-out cross-validated r^2 was 0.74 with an RMSE of 2.5. Figure 12 presents a scatter plot of predicted solubilities and experimental solubilities.

Vapor Pressure

A collection of 1,771 small organic molecules with associated experimental vapor pressure was assembled from the SRC physical property database. The SRC database contains experimental and estimated vapor pressures; accordingly, we selected all compounds with experimental values measured at 25°C. The vapor pressures were converted to a log scale. For each molecule, the SlogP−VSA, SMR−VSA, and

PEOE_VSA1		10	-2	-8	7	13	-13	43	15	-2	-13	-7	25	-5
PEOE_VSA2	10		19	0	16	11	-4	23	14	-7	-1	-4	9	29
PEOE_VSA3	-2	19		21	30	1	2	-8	-27	2	4	11	5	43
PEOE_VSA4	-8	0	21		26	9	16	-16	-20	5	0	24	31	26
PEOE_VSA5	7	16	30	26		8	-9	4	-7	-1	-2	-11	42	53
PEOE_VSA6	13	11	1	9	8		-2	8	-8	7	-4	2	61	17
PEOE_VSA7	-13	-4	2	16	-9	-2		-23	-20	1	22	65	0	4
PEOE_VSA8	43	23	-8	-16	4	8	-23		11	8	-9	-19	17	-1
PEOE_VSA9	15	14	-27	-20	-7	-8	-20	11		-3	-11	-24	1	-21
PEOE_VSA10	-2	-7	2	5	-1	7	1	8	-3		-5	4	5	-1
PEOE_VSA11	-13	-1	4	0	-2	-4	22	-9	-11	-5		-1	-8	-3
PEOE_VSA12	-7	-4	11	24	-11	2	65	-19	-24	4	-1		-6	-4
PEOE_VSA13	25	9	5	31	42	61	0	17	1	5	-8	-6		3
PEOE_VSA14	-5	29	43	26	53	17	4	-1	-21	-1	-3	-4	3	

Figure 6. Correlation matrix of the 14 PEOE−VSA descriptors. Each entry in the matrix is the r *value (signed) in percent. The order of the columns is the same as the order of the rows.*

	1	2	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	28	29	30	31	32
1:		10	2	8	7	13	13	43	15	2	13	7	25	5	0	4	24	35	23	37	11	1	54	27	22	1	9	21	12	41	43	42
2:	10		19	0	16	11	4	23	14	7	1	4	9	29	21	5	28	49	12	27	12	44	20	2	17	28	3	42	11	23	50	16
3:	2	19		21	30	1	2	8	27	2	4	11	5	43	54	12	31	8	5	15	9	14	12	4	16	46	9	48	43	18	8	2
4:	8	0	21		26	9	16	16	20	5	0	24	31	26	41	37	25	4	6	1	6	11	8	0	34	19	47	11	40	14	0	4
5:	7	16	30	26		8	9	4	7	1	2	11	42	53	76	15	52	0	8	2	5	12	9	5	37	65	20	50	38	3	8	4
6:	13	11	1	9	8		2	8	8	7	4	2	61	17	10	31	46	3	34	2	4	5	4	21	62	12	5	7	3	1	10	22
7:	13	4	2	16	9	2		23	20	1	22	65	0	4	1	24	4	1	9	0	2	16	15	8	31	1	15	13	6	25	1	5
8:	43	23	8	16	4	8	23		11	8	9	19	17	1	1	4	11	15	30	20	11	29	26	39	7	5	5	25	6	43	22	45
9:	15	14	27	20	7	8	20	11		3	11	24	1	21	19	20	1	6	2	11	22	56	26	19	8	22	13	4	15	64	9	14
10:	2	7	2	5	1	7	1	8	3		5	4	5	1	2	3	7	4	2	4	3	12	0	23	5	2	2	0	5	10	7	19
11:	13	1	4	0	2	4	22	9	11	5		1	8	3	3	5	6	7	6	2	24	4	10	3	9	27	7	1	3	11	1	9
12:	7	4	11	24	11	2	65	19	24	4	1		6	4	5	23	0	4	3	0	5	16	12	9	20	4	11	17	0	22	0	2
13:	25	9	5	31	42	61	0	17	1	5	8	6		3	23	27	69	4	40	0	6	0	20	22	78	15	25	15	11	13	18	19
14:	5	29	43	26	53	17	4	1	21	1	3	4	3		67	32	39	7	0	1	6	14	4	2	36	51	14	46	42	14	2	10
15:	0	21	54	41	76	10	1	1	19	2	3	5	23	67		2	42	1	5	2	9	16	2	2	27	71	24	65	57	10	7	2
16:	4	5	12	37	15	31	24	4	20	3	5	23	27	32	2		14	6	12	1	2	4	14	3	51	14	7	12	23	4	0	2
17:	24	28	31	25	52	46	4	11	1	7	6	0	69	39	42	14		4	1	4	9	16	23	27	74	29	20	46	22	2	19	31
18:	35	49	8	4	0	3	1	15	6	4	7	4	4	7	1	6	4		0	1	1	12	4	2	1	4	15	10	1	7	90	1
19:	23	12	5	6	8	34	9	30	2	2	6	3	40	0	5	12	1	0		6	0	13	15	8	30	15	2	23	4	22	12	8
20:	37	27	15	1	2	2	0	20	11	4	2	0	0	1	2	1	4	1	6		6	7	14	11	3	9	1	23	8	2	2	36
21:	11	12	9	6	5	4	2	11	22	3	24	5	6	6	9	2	9	1	0	6		21	10	14	10	1	4	10	1	10	0	7
22:	1	44	14	11	12	5	16	29	56	12	4	16	0	14	16	4	16	12	13	7	21		16	22	10	9	8	7	2	71	12	19
23:	54	20	12	8	9	4	15	26	26	0	10	12	20	4	2	14	23	4	15	14	10	16		6	14	3	6	8	11	55	4	2
24:	27	2	4	0	5	21	8	39	19	23	3	9	22	2	2	3	27	2	8	11	14	22	6		25	2	2	8	10	13	10	83
25:	22	17	16	34	37	62	31	7	8	5	9	20	78	36	27	51	74	1	30	3	10	10	14	25		12	21	20	18	1	20	22
26:	1	28	46	19	65	12	1	5	22	2	27	4	15	51	71	14	29	4	15	9	1	9	3	2	12		1	60	21	9	6	1
27:	9	3	9	47	20	5	15	5	13	2	7	11	25	14	24	7	20	15	2	1	4	8	6	2	21	1		3	12	6	3	3
28:	21	42	48	11	50	7	13	25	4	0	1	17	15	46	65	12	46	10	23	23	10	7	8	8	20	60	3		27	12	11	17
29:	12	11	43	40	38	3	6	6	15	5	3	0	11	42	57	23	22	1	4	8	1	2	11	10	18	21	12	27		11	3	15
30:	41	23	18	14	3	1	25	43	64	10	11	22	13	14	10	4	2	7	22	2	10	71	55	13	1	9	6	12	11		12	13
31:	43	50	8	0	8	10	1	22	9	7	1	0	18	2	7	0	19	90	12	2	0	12	4	10	20	6	3	11	3	12		8
32:	42	16	2	4	4	22	5	45	14	19	9	2	19	10	2	2	31	1	8	36	7	19	2	83	22	1	3	17	15	13	8	

Figure 7. Full correlation matrix of r *values (in unsigned percent) between the PEOE−VSA descriptors (rows/columns 1–14), SlogP−VSA descriptors (rows/cols 15–24), and SMR−VSA descriptors (rows/columns 25–32).*

PEOE−VSA descriptors were calculated with MOE. A multiple linear PCA regression model was calculated. Thirty-one principal components were retained, and the resulting r^2 was 0.88 with an RMSE of 2.1. The leave-one-out cross-validated r^2 was 0.87 with an RMSE of 2.2. Figure 13 presents a scatter plot of the predicted vapor pressure and the experimental vapor pressure.

Receptor Class Discrimination

We chose the database of 455 compounds, each active against one of seven receptors, described in Xue et al.[28] (used to develop a set of substructure keys for clustering applications). The database consisted of seven fairly congeneric classes:

Class 1: Serotonin receptor ligands
Class 2: Benzodiazepine receptor ligands
Class 3: Carbonic anhydrase II inhibitors
Class 4: Cyclooxygenase-2 (Cox-2) inhibitors
Class 5: H3 antagonists
Class 6: HIV protease inhibitors
Class 7: Tyrosine kinase inhibitors.

We tested the utility of the VSA descriptors for compound discrimination using the Binary QSAR[29] methodology (a Bayesian

Figure 8. The r^2 correlation coefficients for linear models of traditional descriptors as a function of the 32 VSA descriptors. [a]Connectivity and kappa shape indices.[4] Codes ending in C refer to connectivity indices restricted to the carbon skeleton. [b]van der Waals surface area, volume, and density: molecular weight divided by volume. [c]vsa−hyd, vsa−don, vsa−acc, vsa−pol, and vsa−other refer to van der Waals surface areas of hydrophobic, h-bond donor, h-bond acceptor, polar (donor and acceptor), and other atoms not in those categories, respectively (rule-based atom typing). [d]a−hyd, a−don, and a−acc refer to the number of hydrophobic, h-bond donor, and h-bond acceptor atoms (rule-based atom typing). [e]a−count, a−heavy, a−nC, a−nH, a−nO, a−aro, a−nN, a−nF, a−nP, a−nS, and a−nCl refer to the number of atoms, heavy atom, carbon, hydrogens, oxygen, aromatic, nitrogen, fluorine, phosphorus, sulfur, and chlorine atoms, respectively. [f]Element and graph adjacency matrix entropy measures. [g]Sum of CRC Handbook atomic polarizabilities and sum absolute polarizability differences across each bond. [h]b−count, b−heavy, b−ar, b−single, b−double, and b−triple refer to the number of bonds, heavy aromatic, single, double, and triple bonds, respectively. [i]Total and fractional rotatable and rotatable bonds. [j]Wildman and Crippen logP (octanol/water) and molar refractivity. [k]Molecular weight. [l]Balaban's J index.[6] [m]Graph radius and diameter.[7] [n]Wiener indices.[8] [o]Zagreb index: the sum of the squares of the heavy valences of all heavy atoms.

Name	r^2	Name	r^2	Name	r^2	Name	r^2
chi0 [a]	0.99	chi0v_C [a]	0.97	b_ar [h]	0.89	b_1rotN [i]	0.78
Kier1 [a]	0.99	KierA1 [a]	0.97	Kier2 [a]	0.89	b_double [h]	0.77
vdw_area [b]	0.99	a_hyd [d]	0.96	vsa_pol [c]	0.89	b_rotN [i]	0.77
vdw_vol [b]	0.99	a_nC [e]	0.96	vsa_acc [c]	0.88	a_ICM [f]	0.73
vsa_hyd [c]	0.99	a_nH [e]	0.96	diameter [m]	0.87	vsa_don [c]	0.73
a_count [e]	0.98	a_nO [e]	0.95	VadjEq [f]	0.87	KierFlex [a]	0.69
a_heavy [e]	0.98	b_heavy [h]	0.95	a_nN [e]	0.86	balabanJ [l]	0.61
a_IC [f]	0.98	chi1_C [a]	0.95	KierA2 [a]	0.86	a_nP [e]	0.60
apol [g]	0.98	chi1v_C [a]	0.95	radius [m]	0.86	Kier3 [a]	0.57
b_count [h]	0.98	SlogP [j]	0.95	VdistMa [f]	0.86	a_nCl [e]	0.56
chi0v [a]	0.98	a_acc [d]	0.94	weinerPath [n]	0.85	KierA3 [a]	0.55
chi1 [a]	0.98	chi1v [a]	0.94	weinerPol [n]	0.84	a_nS [e]	0.53
SMR [j]	0.98	Weight [k]	0.93	VadjMa [f]	0.82	b_1rotR [i]	0.50
b_single [h]	0.97	a_aro [e]	0.91	VdistEq [f]	0.82	density [b]	0.49
bpol[g]	0.97	a_don [d]	0.91	vsa_other [c]	0.82	b_rotR [i]	0.48
chi0_C [a]	0.97	zagreb [o]	0.91	a_nF [e]	0.80	b_triple [h]	0.46

inference method of classification). A total of seven Binary QSAR models were made as follows. Model *i* was trained on a data set consisting of "active" molecules (those that were active against receptor *i*) and "inactive" molecules (those that were not active against receptor *i*). MOE was used to calculate the SlogP−VSA and SMR−VSA descriptors for each of the molecules in the database. A Binary QSAR model was constructed from these descriptors in an effort to predict membership in class *i*. The accuracy of prediction and the *p* value (probability of a chance occurrence) for each model was found to be:

Class 1: 98.7%, $p = 0.003$
Class 2: 96.7%, $p = 0.043$
Class 3: 96.5%, $p = 0.290$
Class 4: 98.7%, $p = 0.001$
Class 5: 98.7%, $p = 0.014$
Class 6: 98.7%, $p = 0.012$
Class 7: 99.1%, $p = 0.002$.

Each of the models exhibited high accuracy and all but one (class 3: carbonic anhydrase II inhibitors) exhibited high significance in

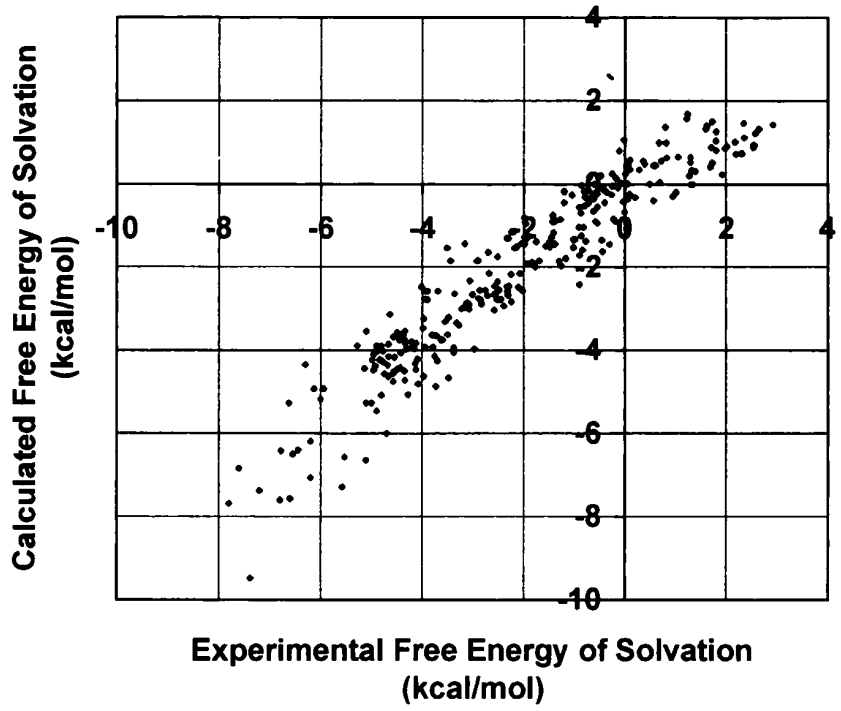


Figure 9. Scatter plot showing calculated (y-axis) and experimental (x-axis) free energy of solvation in kcal/mol for 291 small organic molecules.

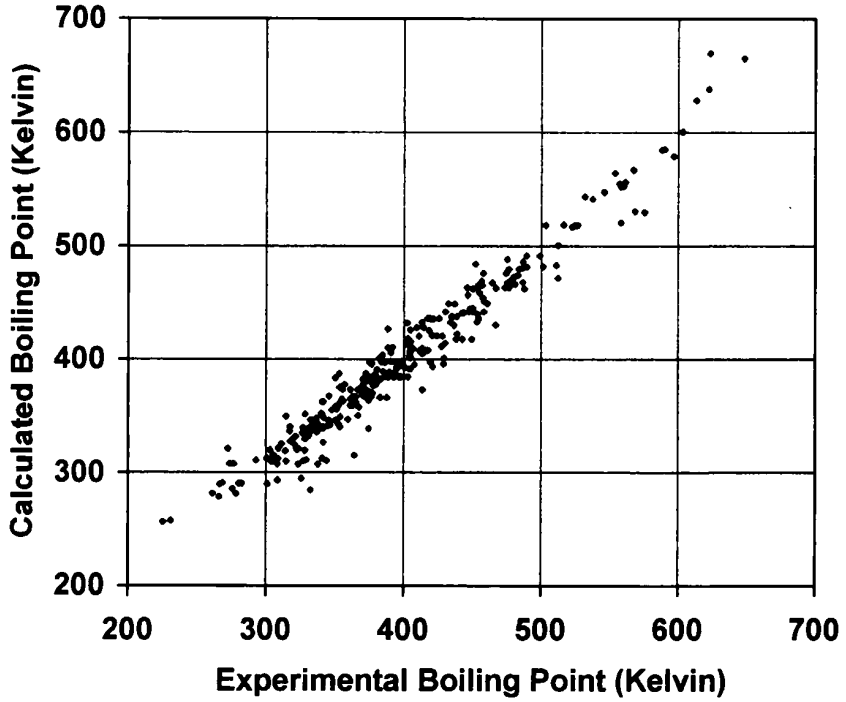


Figure 10. Scatter plot showing calculated (y-axis) and experimental (x-axis) boiling points (Kelvin) of 298 small organic molecules.

the χ-squared significance test. A similar classification model was built using the CART methodology with similar but lower accuracy (results not shown).

Activity Against Thrombin, Trypsin, and Factor Xa

To test the VSA descriptors in linear QSAR modeling, we used a set of 72 analogues described in Bohm et al.[30] with pK_i data for each of thrombin, trypsin, and factor Xa. One such structure is depicted in Figure 14. The PEOE−VSA, SlogP−VSA, and SMR−VSA descriptors were calculated for each structure, and a principal components regression was calculated for each receptor using MOE. For each activity model, descriptors with small (normalized) coefficients were discarded. Using the remaining descriptors, a principal components regression was

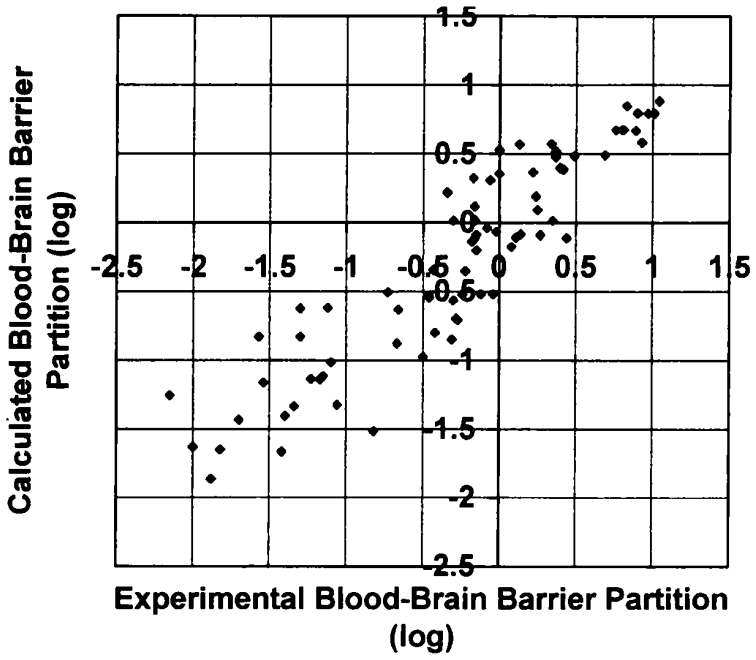


Figure 11. Scatter plot showing calculated (y-axis) and experimental (x-axis) values for 75 compounds of the log concentration ratio between the blood and brain.

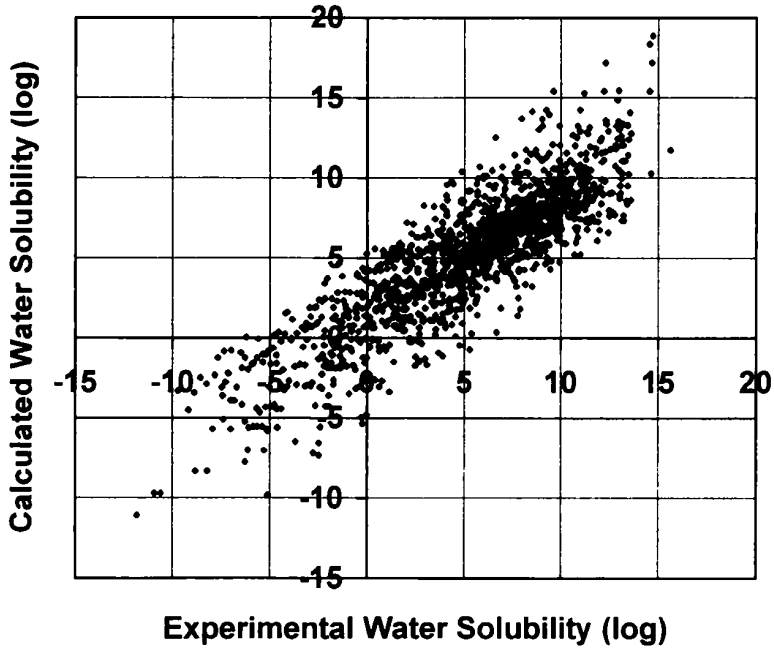


Figure 12. Experimental (x-axis) and calculated (y-axis) solubilities in water of 1,438 small organic compounds (units are log concentration ratios).

calculated. For thrombin, a 10-descriptor model using PEOE−VSA$_{1,2,5,8,10,11,12}$ and SlogP−VSA$_{1,5,9}$ resulted in an r^2 of 0.65 with an RMSE of 0.61 pK_i (Figure 15); the leave-one-out cross-validated r^2 was 0.54 with an RMSE of 0.70 pK_i. For trypsin, a 9-descriptor model using PEOE−VSA$_{1,8,11,12}$, SlogP−VSA$_{0,3,4,8}$, and SMR−VSA$_5$ resulted in an r^2 of 0.72 with an RMSE of 0.47 pK_i (Figure 16); the leave-one-out cross-validated r^2 was 0.62 with an RMSE of 0.54 pK_i. For factor Xa, a 15-descriptor model using PEOE−VSA$_{1,2,8,9,12,14}$, SlogP−VSA$_{5,7,8,10}$ and SMR−VSA$_{3,4,5,6,8}$ resulted in an r^2 of 0.69 with an RMSE of 0.35 pK_i (Figure 17); the leave-one-out cross-validated r^2 was 0.52 with an RMSE of 0.45 pK_i.

DISCUSSION

Rather than discuss the individual results in detail, we will concentrate the discussion on the notion of the 32-dimensional chemistry space defined by the entire collection. The presented linear correlations are not extraordinarily high, and, perhaps, it is too much to expect that a relatively small set of non-3D

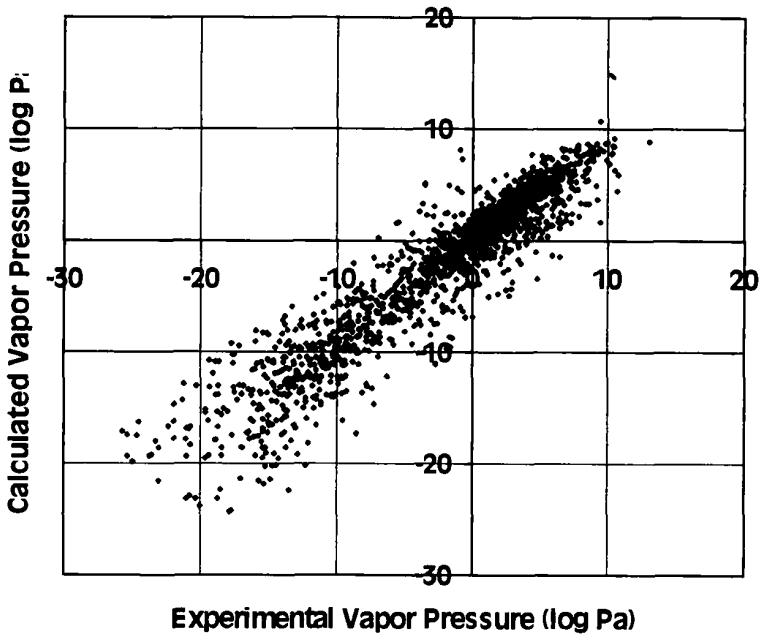


Figure 13. Experimental (x-axis) and calculated (y-axis) vapor pressure of 1,771 small organic compounds (units are log Pa).

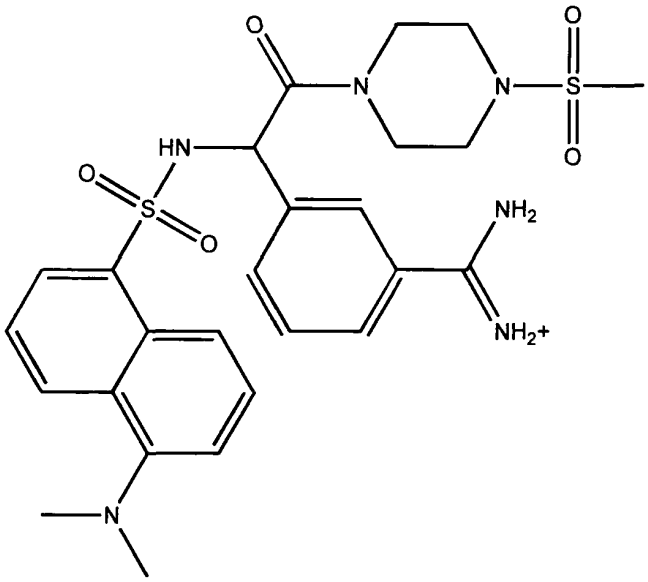


Figure 14. Representative structure from a set of 72 analogues assayed on each of thrombin, trypsin, and factor Xa (described by Bohm et al.[30])

descriptors will be capable of describing a given molecule in great detail in relation to a specific property. What is interesting is the fact that the *same set of descriptors was used throughout.* In this respect, the results suggest that the collection of VSA descriptors could be put to good use in a cheminformatics context, e.g., for chemical diversity, combinatorial library design, and high-throughput screening (HTS) data analysis.

Pearlman and Smith[31] pointed out three problems with the use of "traditional" descriptors in the definition of a chemistry space:

1. *Orthogonality.* Many "traditional" descriptors are highly correlated. A good set of descriptors should be as orthogonal as possible (indicating that each descriptor is encoding different properties from the others).
2. *Relevance.* It can be argued that the "traditional" descriptors, such as logP, pK_a, etc., are more relevant to drug transport or pharmacokinetics than receptor affinity.
3. *Wholism.* One might fear that the "traditional" descriptors are "whole molecule" properties that cannot distinguish the details of important substructural differences.

The orthogonality of a set of molecular descriptors is a very desirable property. Classification methodologies such as CART (or other decision-tree methods) are not invariant to rotations of the chemistry space. Such methods may encounter difficulties with correlated descriptors (e.g., production of

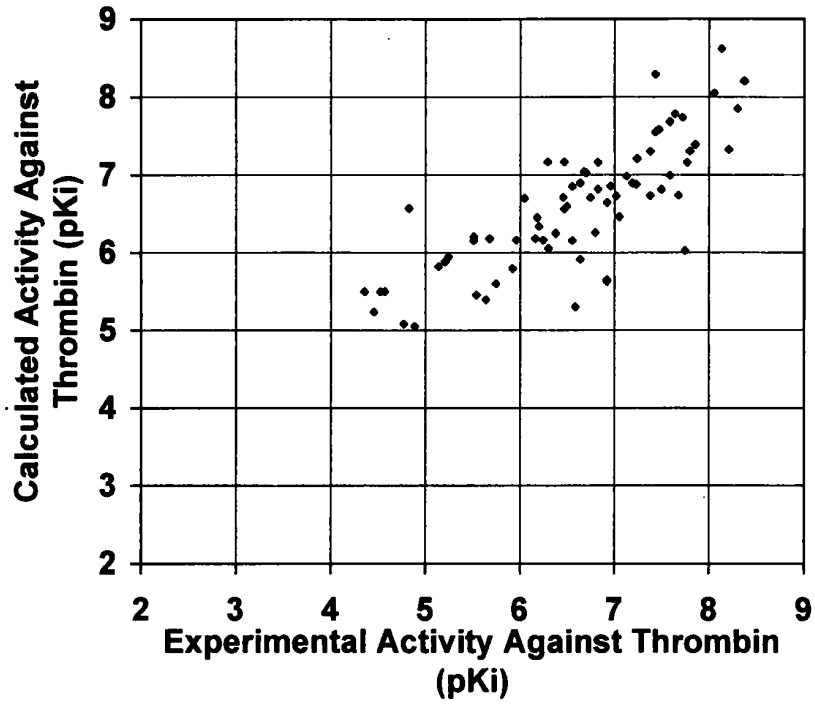


Figure 15. Scatter plot showing predicted (y-axis) and experimental (x-axis) activity of 75 ligands against thrombin (units are pKi).

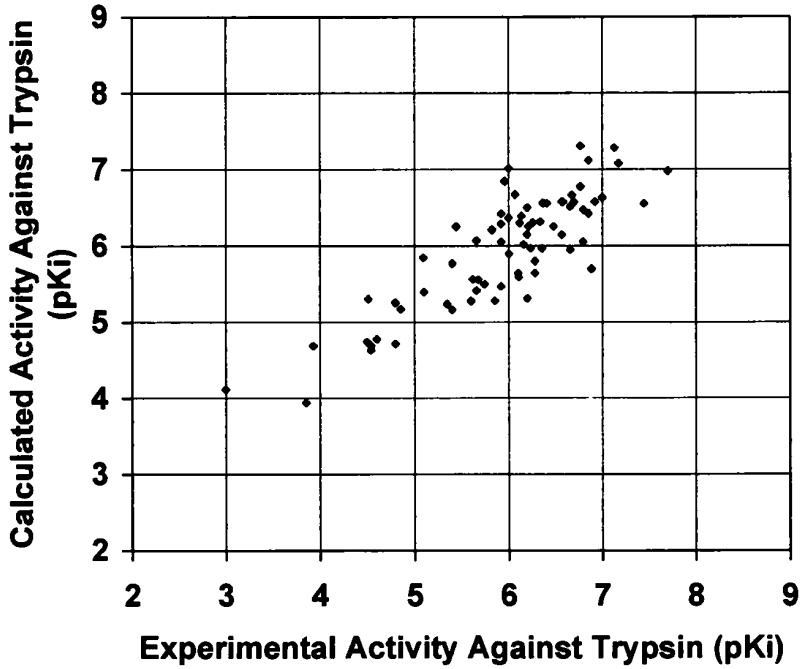


Figure 16. Scatter plot showing predicted (y-axis) and experimental (x-axis) activity of 75 ligands against trypsin (units are pKi).

larger decision trees). Often, correlated descriptors necessitate the use of principal components transforms, which requires a set of reference data for their estimation (at worst the transforms depend only on the data at hand and at best they are trained once from some larger collection of compounds). In probabilistic methodologies, such as Binary QSAR, approximation of statistical independence is simplified when uncorrelated descriptors are used. In addition, descriptor transformations can lead to difficulties in model interpretation. Our results strongly suggest that the VSA descriptors are weakly correlated with each other. As a consequence, we expect that methodologies such as Binary QSAR, CART, principal components analysis, principal components clustering, neural networks, and *k*- means clustering to be more effective (when measured over many problem instances).

The notion of relevance to receptor affinity of a collection of

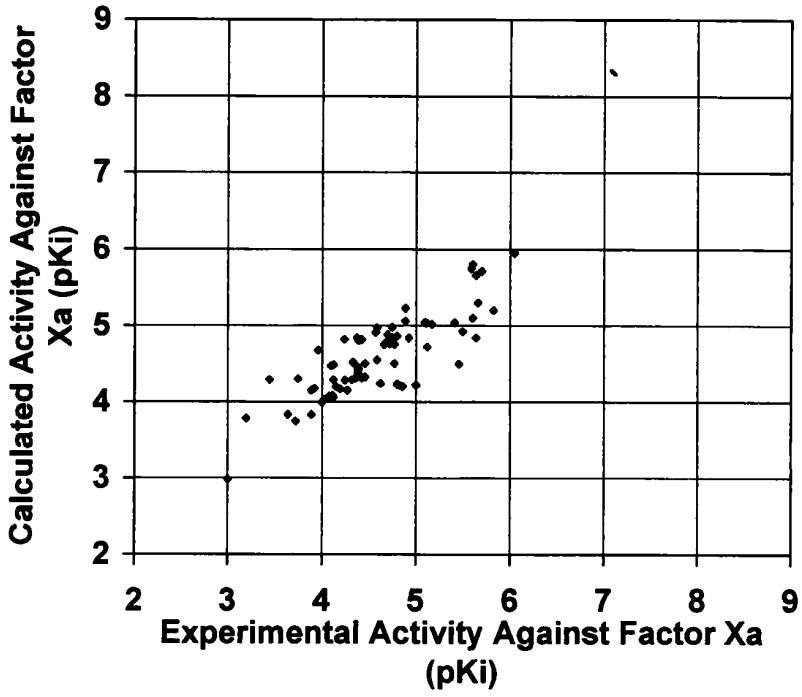


Figure 17. Scatter plot showing predicted (y-axis) and experimental (x-axis) activity of 75 ligands against factor Xa (units are pKi).

descriptors is difficult to quantify. The assertion that "traditional descriptors (e.g., logP and pK_a) are strongly related to drug transport or pharmacokinetics but are very weakly related to receptor affinity or activity..." has to be considered with some care. One says that a descriptor is strongly related to a particular property when effective QSPR models of the property have been made using the descriptor. The failure to produce a QSAR/QSPR model using a descriptor is *not*, in general, evidence of a lack of relevance (for example, the fault could lie with the mathematical model and not the descriptors). The relevance of descriptors must be evaluated either from theoretical considerations or long-term empirical success. Recent work has suggested that the underlying atomic contributions to partial charge, molar refractivity, and logP *are* relevant to receptor affinity.[32] Our results suggest that the presented VSA descriptors are useful not only for physical property modeling but also in receptor affinity modeling. The fact that the same set of descriptors were used with different methodologies (Binary QSAR, linear regression, CART) suggests that it is the descriptors themselves that are encoding relevant information for both classification (compound distinction) and binding affinity QSAR. Our results seem to bear out the intuition that contact surfaces describing hydrophobicity, refractivity (polarizability), and charge localization are relevant to many molecular properties, including receptor affinity. It is an added advantage that the underlying properties used in the definition of the VSA descriptors are relevant to drug transport or pharmacokinetics.

The idea that descriptor wholism is undesirable is a subtle one. It is difficult to quantify the wholism of a descriptor. A qualitative definition might be that a "whole molecule" property is one in which small bioisosteric modifications to the structure lead to large changes in the descriptor value. It is interesting to note that BCUT values, extensions of Burden[33] numbers derived from graph adjacency or distance matrix eigenvalues, are likely exhibit far more wholism than more group-additive properties (such as logP and free energy of solvation). Nevertheless, BCUT values have shown utility in QSAR/QSPR studies[34] and diversity work. Descriptors such as HOMO and LUMO energies are very wholistic, and even these have been used successfully in QSAR work. Intuitively, it seems that group-additive descriptors should be better for compound classification; however, it is difficult to be sure. Notwithstanding these considerations, the VSA descriptors we described are derived from what are widely believed to be "whole molecule" properties. It is hoped that the surface area subdivision will effect a reductionist conversion suitable for QSAR work and compound classification. The atomic VSA contributions are sensitive to connectivity and the properties considered (logP, MR, and charge) are sensitive to the chemical context of each atom. Moreover, each of the VSA descriptors is fundamentally additive in nature, which suggests a more reductionist than wholist character. The high correlations seen when modeling other descriptors, such as number of nitrogens, number of oxygens, and number of aromatic atoms, support this reductionist assertion.

If we take it as true that the presented VSA descriptors form a (relatively) low-dimensional chemistry space encoding trend information for many properties of interest, we can consider cheminformatics applications. Compound selection methods based on chemical diversity are likely to benefit from the VSA descriptors: not only are physical properties taken into account, but also properties relevant to binding, transport, and pharmacokinetics. Setting aside diversity-based methodologies, we now consider the application of the VSA descriptors to HTS QSAR and focused combinatorial library design.

The automation of physical experiments through robotics to effectively perform hundreds of thousands or millions of ex-

periments in a short time has opened the door to a large-scale approach to drug discovery. HTS and combinatorial chemistry offer access to a huge set of candidate structures; however, time and economic considerations require a selection of only a subset of this vast space for physical testing. Unfortunately, most (if not all) people find it very difficult to interpret all of the HTS data when effecting a focused combinatorial library design. HTS QSAR is an alternative to human inspection of HTS data. In this alternative, a set of HTS results is considered to be "understood" if an effective QSAR model can be constructed (by effective, we mean statistically significant). The activity of new compounds (for example, in a proposed library) can be predicted with the model. The PEOE−VSA descriptors have been used successfully in several HTS QSAR attempts[35] using the Binary QSAR method. Accuracy levels of 40%–70% have been routinely observed on active compounds on data sets with hit rates well below 1% (inactives usually exhibit >90% accuracy). It is hoped that the SlogP−VSA and SMR−VSA descriptors will improve the accuracy levels (although the PEOE−VSA accuracy levels still resulted in significant enrichment when compared to the hit rate). Such a study will be presented in a future publication.

Suppose, now, that a statistically significant HTS QSAR model has been constructed. Further suppose that a proposed virtual combinatorial library L is made up of m substituent libraries $R_1,\ldots,R_m$. Thus, the number of compounds in L is $|\ R_1\ |\ \mathrm{X}\ldots\mathrm{X}\&\mathrm{z.fnc};R_m\ |$. Even moderately sized substituent libraries can result in extremely large product libraries. In such a case, some method is required to select a subset of each R_i for physical synthesis. An obvious criterion is to select those members of the substituent libraries that are most likely to result in products that are active against some target. Let r_{ij} be the j-th member of the i-th substituent library (R_i). We can use a first-order ranking of the members of r_{ij} to select the most promising members. One such ranking is the probability of observing the substituent in an active product, namely:

$$\Pr(R_i = r_{ij}|\text{active}) = \frac{\Pr(\text{active}|R_i = r_{ij})\Pr(R_i = r_{ij})}{\sum_k \Pr(\text{active}|R_i = r_{ij})\Pr(R_i = r_{ik})}$$

(after an application of Bayes theorem). We can assume that the members of R_i are equally likely so that this last formula simplifies to:

$$\Pr(R_i = r_{ij}|\text{active}) = \frac{\Pr(\text{active}|\mathrm{R}_i = r_{ij})}{\sum_k \Pr(\text{active}|R_i = r_{ij})}$$

The term Pr(active $|\ R_i = r_{ij}$) is precisely the output of the Binary QSAR methodology; that is, one can use the VSA descriptors and the HTS results as input to Binary QSAR and obtain an estimate for Pr(active $|\ R_i = r_{ij}$) for each member r_{ij} of R_i. After the indicated division by the sum of these estimates, one would obtain a probability distribution over R_i. A simple design methodology would be to retain the top scoring members of each substituent library. One advantage of this sort of probabilistic ranking scheme is that in the event that L is so large that enumeration of all structures is prohibitive, random sampling techniques can be used to estimate the required probabilities without loss of theoretical soundness.

In summary, our results suggest the utility of the presented VSA descriptors for QSAR/QSPR studies and cheminformatics applications based on either chemical diversity or bias. The success of the descriptors in modeling various properties does not seem tied to particular numerical methodology. In particular, the use of the Binary QSAR method used with the VSA descriptors provides a statistically sound method of focused combinatorial library design.

CONCLUSIONS

We defined three sets of (easily calculated) molecular descriptors based on atomic contributions to logP, molar refractivity, and atomic partial charge. The individual descriptors were found to be weakly correlated with each other (over a suitably large collection of compounds). Moreover, the chemistry space determined by the new descriptors was capable of expressing (as linear combinations) traditional QSAR/QSPR descriptors. Reasonably good QSAR/QSPR models of boiling point, vapor pressure, free energy of solvation in water, water solubility, receptor class, and activity against thrombin, trypsin, and factor Xa were built using only the new descriptors.

We conclude that (1) the new descriptors are likely to be a very good starting point for QSAR/QSPR work; and (2) the collection of new descriptors may be a meaningful low-dimensional chemistry space for chemical diversity, HTS data analysis, and combinatorial library design.

The software to calculate the new descriptors is available (under license) in the Molecular Operating Environment software from Chemical Computing Group Inc.

ACKNOWLEDGMENTS

The author thanks Ana Lin, William Long, and Chris Williams for their help in collecting the data for this paper, as well as their valuable suggestions regarding the presentation of the method and results.

REFERENCES

1 Hansch, C., and Fujita, T. ρ-σ-π Analysis: A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* 1964, **86,** 1616–1626
2 Leo, A., Hansch, C., and Church. C. Comparison of parameters currently used in the study of structure-activity relationships. *J. Med. Chem.* 1969, **12,** 766–771
3 Hogg, R.V., and Tanis, E.A. *Probability and statistical inference.* MacMillan Publishing Company, New York, 1993
4 Hall, L.H., and Kier, L.B. The molecular connectivity chi indices and kappa shape indices in structure-property modeling. In: *Reviews of Computational Chemistry, Volume 2,* Boyd, D.B., and Lipkowitz, K., Eds., 1991
5 Hall, L.H., and Kier, L.B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* 1995, **35**
6 Balaban, A.T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* 1979, **53,** 355–375
7 Petitjean, M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* 1992, **32,** 331–337

8 Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* 1947, **69,** 17–20
9 Rogers, D., and Hopfinger, A.J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* 1994, **34**
10 Breiman, L., Friedman, J., Olshen, R.A., and Stone, C.J. *Classification and regression trees.* Wadsworth Inc., 1984
11 Pearlman, R.S., and Smith, K.M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Design* 1998, **9,** 339–353
12 Todeschini, R., Lasagni, R., and Marengo, E. New molecular descriptors for 2D and 3D structures: Theory. *J. Chemometrics* 1994, **8,** 263–272
13 Stanton, D.T. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 11–20
14 Bayada, D.M., Hamersma, H., and van Geerestein, V.J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 1–10
15 Jones, G., Willett, P., and Glen, R.C. A genetic algorithm for flexibal molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Design* 1995, **9**
16 Wodak, S.J., and Janin, J. Analytical approximation to the solvent accessible surface area of proteins. *Proc. Natl. Acad. Sci. U.S.A.* 1980, **77,** 1736–1740
17 Halgren, T.A. MMFF94 The Merck Force Field. *J. Comp. Chem.* 1996, **17**
18 MOE: The Molecular Operating Environment from Chemical Computing Group Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada H3B 3X3
19 The expression $[a,b)$ denotes the half closed interval $\{x: a < x < b\}$.
20 Wildman, S. A., and Crippen, G.M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 868–873
21 Maybridge Chemical Company Ltd., Cornwall, PL34 OHW England. http://www.maybridge.co.uk
22 Gasteiger, J., and Marsali, M. Iterative partial equalization of orbital electronegativity: A rapid access to atomic carges. *Tetrahedron* 1980, **36,** 3219
23 Compounds were selected by molecular weight from the mref.mdb database on the MOE 1999.05 CD-ROM
24 Viswanadhan, V.N., Ghose, A.K., Singh, U.C., and Wendoloski, J.J. Prediction of solvation free energies of small organic moleucles: Additive-constitutive models based on molecular fingerprints and atomic constants. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 405–412
25 The boiling point data can be made available upon request of the author.
26 Luco, J.M. Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least squares (PLS) methodology. *J. Chem. Info. Comput. Sci.* 1999, **36,** 396–404
27 Syracuse Research Corporation, 6225 Running Ridge Road, North Syracuse, NY 13212. http://www.syyres.com
28 Xue,L., Godden,J., Gao,H., and Bajorath,J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 699
29 Labute,P. Binary QSAR: A new method for quantitative structure activity relationships. *Proceedings of the 1999 Pacific Symposium.* World Scientific Publishing, Singapore, 1999
30 Bohm, M., Sturzebecher, J., and Klebe, G. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin and factor Xa. *J. Med. Chem.* 1999, **42,** 458–477
31 Pearlman, R.S., and Smith, K.M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 28–35
32 Crippen, G.M. VRI: 3D QSAR at variable resolution. *J. Chem. Inf. Comput. Sci.* 1999, **20,** 1577–1585
33 Burden, F.R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* 1989, **29,** 225–227
34 Stanton, D.T. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 11–20