

Development and Use of Hydrophobic Surface Area (HSA) Descriptors for Computer-Assisted Quantitative Structure–Activity and Structure–Property Relationship Studies

David T. Stanton,^{*,§} Brian E. Mattioni,[†] James J. Knittel,^{‡,§} and Peter C. Jurs[†]

Procter & Gamble Pharmaceuticals, Health Care Research Center, 8700 Mason Montgomery Road, Mason, Ohio 45040-9462, and Chemistry Department, 152 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 7, 2003

A new series of 25 whole-molecule molecular structure descriptors are proposed. The new descriptors are termed *Hydrophobic Surface Area*, or *HSA* descriptors, and are designed to capture information regarding the structural features responsible for hydrophobic and hydrophilic intermolecular interactions. The utility of the HSAs in capturing this type of information is demonstrated using two properties that have a known hydrophobic component. The first study involves the modeling of the inhibition of Gram-positive bacteria cell growth of a series of biaryl amides. The second application involves the study of the blood-brain barrier penetration of a diverse series of drug molecules. In both cases, the HSAs are shown to effectively capture information related to the hydrophobic components of these two properties. Additional evaluation of the new class of descriptors shows them to be unique in their ability to measure hydrophobic features among a diverse set of conventional structural descriptors. The HSAs are evaluated regarding their sensitivity to conformational changes and are found to be similar in that regard to other widely used molecular descriptors.

INTRODUCTION

To develop sound, useful, and physically meaningful Quantitative Structure–Activity Relationship (QSAR) and Quantitative Structure–Property Relationship (QSPR) models, one must pay special attention to how one characterizes the structure of the molecules under investigation. One method for characterizing structure is to compute direct measures of structure that capture very specific features of the molecule. Such measures are typically referred to as *descriptors*. A very wide variety of descriptors has been described over the years.^{1,2} For example, the molecular connectivity indices³ are representatives of the larger family of parameters called topological descriptors.⁴ These descriptors are based only on the connectivity of the atoms in the structure and are independent of conformation. Historically, these descriptors have been found to be useful for providing measures of molecular size, shape, and branching. Another class of parameters is the geometric descriptors. These also capture information concerning the size and shape of molecules but are sensitive to changes in conformation. Examples of this type of descriptor are the principal moments of inertia and various measures of surface area and volume.⁵ A third general class of parameters is the electronic descriptors. These provide measures of the distribution of charge within the molecule and are exemplified by descriptors such

as dipole moment,⁶ greatest positive or negative charge,⁷ and the submolecular polarity parameter.⁸

There are a number of ways to determine the utility or importance of a given descriptor. One is to find that its use increases one's ability to develop an improved model for a property of interest. Another is to find that particular descriptors are often included in many models for different types of properties. A third way is to determine the specific purpose of a descriptor in a model. Recently, PLS analyses have been used to extract the physical interpretation for QSAR regression models.⁹ Using that methodology, it is possible to determine the specific purpose of each descriptor in the model. This not only provides a way of generating a detailed picture of the structure–activity or structure–property relationship encoded in the model but also provides a physical meaning behind the inclusion of a particular descriptor in the model. Using that methodology, it was determined that hydrophobic features of a set of biaryl amides were being identified using descriptors originally developed to capture information concerning structural features responsible for polar intermolecular interactions.¹⁰ While that model still performed well for the intended purpose, and it was still possible to determine the physical reasons why the descriptors were included in the model, the evidence suggested that none of the descriptors evaluated were designed to capture information regarding structural features responsible for hydrophobic intermolecular interactions. This suggested two problems. The first was that it was likely that the hydrophobic features were not being well characterized. The second problem involves the physical interpretation of the descriptors in the model. It is possible that a descriptor designed for one purpose actually turns out to be the best available

* Corresponding author e-mail: stanton.dt@pg.com. Current address: Procter & Gamble Co., Corporate Research – Chemical Technology Division, Miami Valley Laboratories, 11810 East Miami River Road, Cincinnati, OH 45252.

[†] The Pennsylvania State University.

[‡] Current address: Division of Pharmaceutical Sciences, College of Pharmacy, University of Cincinnati, P.O. Box 670004, 3223 Eden Ave, Cincinnati, OH 45267-0004.

[§] Procter & Gamble Pharmaceuticals.

descriptor for measuring another and unrelated type of structural feature. For example, during the development of a model for surface tension, it was found that a descriptor that expresses the relative distribution of atomic partial charges in a molecule was the best available descriptor for capturing differences in branching among a set of hydrocarbons,¹¹ even though a wide range of topological descriptors had been evaluated. The inclusion of an electronic descriptor in the model might naturally lead one to develop a physical interpretation that involves an electrostatic interaction, even though the correct interpretation was based on the descriptor's ability to express differences in molecular shape. Such a situation can slow the proper physical interpretation of the model in the best case and, in the worst case, can lead to a misinterpretation of the physical meaning of the model. For these reasons, we began development of a set of descriptors that would be designed specifically to capture information regarding features responsible for hydrophobic intermolecular interactions.

Hydrophobicity has been defined as "the association of nonpolar groups or molecules in an aqueous environment which arises from a tendency of water to exclude nonpolar molecules",¹² and it is often represented in QSAR models through the calculation of the logarithm of the *n*-octanol/water partition coefficient ($\log P$).^{13,14} The hydrophilicity index (*Hy*) of Todeschini is another proposed way of capturing molecular hydrophobicity.¹⁵ While being reliable ways to provide a one-dimensional, holistic view of hydrophobicity, little progress has been made on expanding and developing new descriptors which capture regional or localized hydrophobic effects on a molecular surface. In several QSAR models for $\log P$ prediction, it was shown that surface area is an important descriptor in determining accurate partition coefficients of organic compounds.^{16–26} More specifically for two recent atom-additive approaches to $\log P$ estimation, it was shown that the combination of partial atomic surface areas and atomic hydrophobicity constants improved model quality and predictive ability.^{27,28} Furthermore, several researchers have developed various kinds of surface area descriptors which could be used to capture hydrophobic/hydrophilic effects. In most cases, either all or only specific atoms of a molecule are classified into regions of varying degrees of hydrophobic character with the resultant descriptors being computed by simply summing the respective partial atomic surface areas for each region.^{29–31} Due to the significant role that surface area played in the aforementioned studies, the initial focus was to include surface area information in our new suite of hydrophobic descriptors in order to fully capture information about hydrophobic intermolecular interactions.

As a starting point, we began with the same concepts that form the basis of the charged partial surface area (CPSA) descriptors.^{32,33} These descriptors are hybrids of two general classes of descriptors: geometric and electronic. In the case of the CPSA descriptors, it was proposed that polar interactions between molecules would be driven primarily by the charge distribution within the molecule and that the degree of influence of individual atoms on that interaction would be a function of their exposure at the surface of the molecule. Similarly, we propose that hydrophobic interactions between molecules will be driven by hydrophobicity of the individual atoms in a molecule and that the influence of any particular

atom on this interaction will be modulated by the degree of its exposure on the molecular surface. In past work, we have found that the solvent-accessible surface of the molecule works well for characterizing the surface exposure of individual atoms. What was needed next was a measure of the atomic contributions to hydrophobicity. Recently, Wildman and Crippen described a series of hydrophobicity parameters for a wide variety of atom types.³⁴ These fragment values could be used to compute other properties such as *n*-octanol/water partition coefficients.

In this work, we describe a new series of descriptors that we term *hydrophobic surface area* or *HSA* descriptors that combine the solvent-accessible surface and hydrophobic contributions of atoms in a variety of ways. As previously referenced, some of the HSA descriptors are comparable to the VSA descriptors developed by Labute.³⁰ He used the hydrophobicity constants of Wildman and Crippen to place atoms into one of 10 hydrophobic activity bins. After binning, 10 surface area descriptors (one descriptor per bin) were computed by summing the individual atomic solvent accessible surface areas for each bin. The applicability and quality of the VSA descriptors have been demonstrated in the literature.^{35–40} We decided to investigate this type of structural descriptor in more detail in an effort to better describe molecular features that are responsible for hydrophobic interactions. However, instead of separating atoms into a number of bins, our implementation classifies atoms as either hydrophilic or hydrophobic dependent upon the value of their atomic hydrophobicity constant. Furthermore, we go beyond a simple summation of solvent accessible surface area and use a more elaborate scheme to combine information about partial surface areas and the individual hydrophobic contributions for each atom in a molecule.

We have found that these descriptors function well in characterizing features responsible for hydrophobic intermolecular interactions. The utility of the new HSA descriptors is illustrated below using two separate QSAR studies that have an established hydrophobic physical interpretation. The first study involves the development of a QSAR model of the efficacy of a series of biaryl amides as inhibitors of bacterial cell growth expressed as the minimum inhibition concentration (MIC) for a series of Gram-positive organisms that included different strains of *Staphylococcus aureus*, *Enterococcus faecium*, *E. faecali*, and *Streptococcus pneumoniae*. The second study describes the development of a QSPR model of blood-brain barrier permeation for a diverse series of commercial drug molecules. In both cases, the HSAs are found to be involved in explaining the hydrophobic portion of the SAR for each model.

EXPERIMENTAL SECTION

Calculation of the HSAs. A series of 25 individual descriptors was created that combine the atomic contributions to hydrophobicity and solvent-accessible surface area in a variety of ways, similar to the approach taken with the original CPSA descriptors. These were implemented within the ADAPT software system.^{41,42} The descriptor name and formulas used to compute the HSAs are provided in Table 1. The atomic $\log P$ contributions were used as described by Wildman and Crippen.³⁴ A PERL script was written that examined HyperChem structure files (.hin) and assigned each

Table 1. Equations Used for Calculation of the HSA Descriptors

descriptor	label	equation ^a
hydrophobic surface area	PPHS-1	$\sum(+SA_i)$
hydrophilic surface area	PNHS-1	$\sum(-SA_i)$
total hydrophobic weighted PPHS	PPHS-2	$(\sum(+SA_i))(\log P_T)$
total hydrophilic weighted PNHS	PNHS-2	$(\sum(-SA_i))(-\log P_T)$
atomic constant weighted PPHS	PPHS-3	$\sum(+A_i)(\log P_i)$
atomic constant weighted PNHS	PNHS-3	$\sum(-A_i)(-\log P_i)$
difference between hydrophobic and hydrophilic surface areas	DHS-1	PPHS-1 - PNHS-1
	DHS-2	PPHS-2 - PNHS-2
	DHS-3	PPHS-3 - PNHS-3
fractional hydrophobic/hydrophilic surface areas	FPHS-1 FNHS-1	HSA/total molecular surface area
	FPHS-2 FNHS-2	
	FPHS-3 FNHS-3	
surface weighted hydrophobic/hydrophilic surface areas	WPHS-1 WNHS-1	HSA \times total molecular surface area/1000
	WPHS-2 WNHS-2	
	WPHS-3 WNHS-3	
relative hydrophobicity	RPH	most hydrophobic atom constant/sum total hydrophobic constants
relative hydrophilicity	RNH	most hydrophilic atom constant/sum total hydrophilic constants
relative hydrophobic surface area	RPHS	max. $SA_{MPOS} \times RPH-1$
relative hydrophilic surface area	RNHS	max. $SA_{MNEG} \times RNH-1$

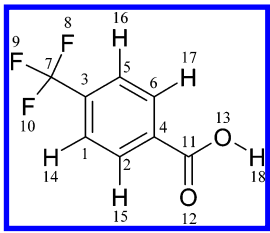
^a $(+SA_i)$ and $(-SA_i)$ are the surface area contributions of the *i*th hydrophobic or hydrophilic atom in a molecule. $(\log P_i)$ and $(-\log P_i)$ are the atomic constants for the *i*th hydrophobic or hydrophilic atom in a molecule, while $(\log P_T)$ and $(-\log P_T)$ are the sum total hydrophobic and hydrophilic constants for each molecule.

atom to one of the 68 atom types described in their work. Due to software limitations with ADAPT, only 59 of the atom types were used in our implementation of the atom typing scheme. The atomic contributions to the solvent accessible surface area were computed using the SAVOL program of Pearlman⁵ using a probe radius of 1.5 Å.

In Table 1, there are three partial hydrophobic surface area descriptors (PPHS) and three partial hydrophilic surface area descriptors (PNHS). There are also three descriptors which assess the difference in the partial surface area descriptors, six fractional HSA descriptors (FPHS/FNHS), and a similar set of six total surface area weighted partial surface area descriptors (WPHS/WNHS). In addition to the partial surface area descriptors, it was desired to examine the relative influence of the most hydrophobic and most hydrophilic atom on the overall lipophilicity of the molecule. These descriptors are named relative hydrophobicity (RPH) and relative hydrophilicity (RNH). This information was then combined with the solvent accessible surface area of the most hydrophobic and most hydrophilic atoms to obtain the relative hydrophobic surface area (RPHS) and relative hydrophilic surface area (RNHS) descriptors. An example of values which would be needed to calculate the HSA descriptors for *p*-trifluoromethylbenzoic acid are given in Table 2. Table 3 shows the calculated values for the HSA descriptors when using the example molecule.

Biarylamine MIC Model Development and Validation.

The training set for this model consisted of a series of 47 compounds that were the focus of a recent study.¹⁰ The biological data are expressed as the geometric mean minimum growth inhibition concentration (MIC) for 6 g-positive microorganisms. The new set of 25 HSAs was computed for these structures and was combined with the other 144 descriptors computed for the original study. The process of objective feature analysis was then applied as described elsewhere,⁴³ yielding a reduced pool of information-rich descriptors. Model development was carried out in ADAPT using both the generalized simulated annealing approach⁴⁴

Table 2. Solvent-Accessible Surface Area and Hydrophobicity Constants Used To Calculate the HSA Descriptors for a Molecule of *p*-Trifluoromethylbenzoic Acid^a


atom number	SASA (Å ²)	hydrophobic constant
1	12.3	0.1581
2	14.2	0.1581
3	3.39	0.1360
4	6.29	0.1360
5	12.5	0.1581
6	14.5	0.1581
7	9.41×10^{-14}	-0.0967
8	42.8	0.4202
9	37.8	0.4202
10	37.9	0.4202
11	9.29	-0.2783
12	40.9	0.1129
13	26.0	-0.2893
14	18.2	0.123
15	17.7	0.123
16	18.3	0.123
17	18.7	0.123
18	27.5	0.298

^a Total molecular surface area = 358 Å². Maximum surface area of the most hydrophobic atom = 42.8 Å². Maximum surface area of the most hydrophilic atom = 26.0 Å².

and the genetic algorithm methodology.⁴⁵ Additional statistical evaluation of the model was performed using linear regression methods in the Minitab package.⁴⁶ Prior to acceptance of a final model, partial least squares (PLS) analysis⁴⁷ was performed using the Minitab program to ensure that the model was not overfitted. A model was considered to be overfitted if the PLS analysis showed the

Table 3. HSA Descriptor Values for *p*-Trifluoromethylbenzoic Acid

descriptor	value	descriptor	value
PPHS-1	323	FNHS-2	-0.0654
PPHS-2	991	FNHS-3	-0.0282
PPHS-3	81.4	WPHS-1	116
PNHS-1	35.3	WPHS-2	355
PNHS-2	-23.4	WPHS-3	29.2
PNHS-3	-10.1	WNHS-1	12.6
DHS-1	288	WNHS-2	-8.39
DHS-2	1020	WNHS-3	-3.62
DHS-3	91.5	RPH	0.137
FPHS-1	0.902	RNH	0.436
FPHS-2	2.77	RPHS	5.87
FPHS-3	0.227	RNHS	11.3
FNHS-1	0.0984		

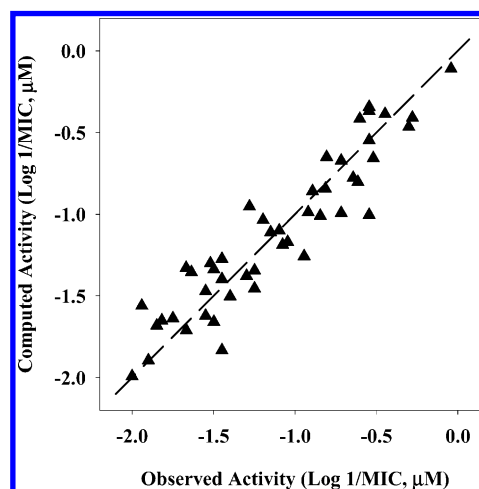
number of validated components to be less than the number of original descriptors in the model (e.g., a seven-variable model yielded six or fewer validated PLS components). A test for chance correlation was then performed that involved a 10-fold randomization of the dependent variable.⁴⁸ The set of the original descriptors from the model were then regressed against each of the 10 instances of the randomized dependent variable. The average R^2 value for the 10-fold randomized regressions was computed and compared to that of the original model. The correlation of the original MICs and the randomized values was also determined. The likelihood of having obtained a good model based strictly on chance was ruled out if the original model's R^2 was substantially higher than the average R^2 obtained from randomization experiments and if the correlation between the original MICs and the randomized values was also small. Once a final model was obtained, PLS analysis was repeated to obtain the score plots and X-variable weights for the components that explained the majority of the variance in the observed property values (Y-variable). This information is used to develop the final physical interpretation of the model using a procedure described previously.⁹

Blood-Brain Barrier Penetration Model Development and Validation. The training set for this model consisted of a series of 97 compounds that were part of a set of 106 compounds in a recent study by Rose et al.⁴⁹ The smaller training set resulted from the removal of structures of very small molecules such as methane and nitrogen or structurally unusual (e.g. sulfur hexafluoride). Additionally, compounds that were previously determined to be statistical outliers were not considered here. The blood-brain barrier penetration was modeled as the logarithm of the blood-brain partition coefficient (Log BB). All log BB values were checked with the original literature citations, and corrections to the data set were made when discrepancies were found. The set of 25 HSAs was computed for these structures and combined with a set of 133 diverse descriptors generated with the ADAPT program. In addition, the descriptors that comprised the final model described by Rose et al. were included. Objective feature analysis was used as described above to produce a reduced pool of information-rich descriptors. Two separate models for this data set were sought. The first case excluded the HSA descriptors in order to establish a structure-property model using established molecular descriptors. The second model development exercise was done with the HSA included in order to determine whether the HSAs would be selected to generate high-quality models if

Table 4. Summary of the New Biarylamine MIC Model Incorporating HSA Descriptors

descriptor	$R^2 = 0.864, s = 0.201, N = 47, F\text{-value} = 42.3$			variance inflation factor
	regression coefficient	SD of coefficient	t-value	
FNSA-2 ^a	-1.03	0.144	-7.14	3.4
ACGD ^b	8.35	1.92	4.36	1.6
NRA ^c	0.264	0.0255	10.4	2.0
1SP3 ^d	0.359	0.0546	6.58	1.8
FPHS-2 ^e	-0.237	0.0745	-3.19	4.5
FPHS-3 ^e	7.02	1.07	6.59	2.6
y-intercept	-10.3	0.903	-11.4	N/A

^a A CPSA descriptor. ^b The average difference in charge between all pairs of H-bond donors.⁵³ ^c Simple count of ring atoms. ^d Count of occurrences of an sp³ hybridized carbon atom bonded to only one other carbon atom. ^e An HSA descriptor

**Figure 1.** Plot of the computed and observed activity values for the 47-compound biarylamine data set model incorporating the HSA descriptors.

other conventional descriptors were available and, when included, if the HSAs would help explain the hydrophobic component of blood-brain barrier penetration. In both cases, model development and subsequent model validation was carried out using the same methods described for the biarylamine data set above. The physical interpretation of the final models was performed using the procedure described previously.

RESULTS AND DISCUSSION

Biarylamine MIC Model. Results obtained during the model development steps suggested an equation involving 6 terms (descriptors) was optimal. It was interesting to find the optimal model contained two of the new HSAs, selected from among the original set of 169 descriptors. The summary of the model is provided in Table 4. A scatter plot illustrating the correlation between the fitted and observed activity values is shown in Figure 1. The model exhibited a good fit to the observed activity values, yielding an R^2 value of 0.864 and a cross-validated R^2 (Q^2) of 0.820 based on the leave-one-out (LOO) method. The model is statistically sound, with an overall- F value⁵⁰ of 42.3 (compared to a critical- F value of 2.34 with 6 and 40 degrees of freedom and an alpha of 0.05) and the lowest partial- F value⁵¹ of 10.2 (compared to a critical- F of 4.28 with 1 and 40 degrees of freedom and an alpha of 0.05). There is little collinearity between the

Table 5.

a. Summary of the Results of the PLS Analysis of the Biarylarnides MIC Model Incorporating the HSA Descriptors

component	residual sum of squared error	R^2 (cumulative)	PRESS	Q^2 (cumulative)
1	4.76	0.597	5.51	0.533
2	3.28	0.722	4.15	0.649
3	2.40	0.797	3.31	0.720
4	1.67	0.858	2.34	0.802
5	1.62	0.863	2.17	0.816
6	1.61	0.864	2.13	0.820

b. X-Weights for Each Descriptor in the First Four Components of the PLS Analysis of the Biarylarnides MIC Model that Incorporates HSAs

descriptor	component-1 X-weight	component-2 X-weight	component-3 X-weight	component-4 X-weight
FNSA-2	-0.556	0.158	-0.331	-0.081
ACGD	0.219	-0.218	0.012	0.802
NRA	0.325	0.835	0.038	-0.106
ISP3	0.067	0.417	0.061	0.474
FPHS-2	0.620	-0.153	-0.711	-0.168
FPHS-3	0.385	-0.180	0.617	-0.293

model descriptors as is evidenced by a maximum variance inflation factor, VIF,⁵² of 4.5 and an average VIF of 2.7. Ideally, VIF values should be less than 10 and have an average close to 1. Since the model descriptors were selected from a large initial pool of descriptors, special attention was paid to the possibility of obtaining a good fit strictly by chance. The 10-fold randomization experiment yielded an average R^2 of 0.144, with a maximum observed R^2 of 0.315. The maximum correlation coefficient obtained between any of the 10-fold randomized sets of activity values and the original activity values was -0.204. Taken together, these results suggest the likelihood is low that the model described above is a result of chance correlations. Finally, the PLS analysis of the model described above validated all six components, suggesting the model is not overfitted. Based on these results, we concluded that the model was statistically valid and that attention could be turned to physical interpretation.

An examination of the final optimized biarylarnide QSAR model was performed using PLS as previously described. A similar analysis of a previous model for the biarylarnides has been reported elsewhere.¹⁰ A summary of the PLS analysis of the new 6-variable model is provided in Table 5. The first four PLS components account for nearly all the variance in the observed MIC values (85.8% of 86.4% total), so attention was focused on understanding the underlying structure-activity relationship (SAR) captured in them. The first PLS component accounts for 59.7% of the variance in the observed MICs values. The score plot for component-1 is shown in Figure 2. Points representing structures of interest are identified on the plot. This first component is dominated by just two descriptors, FPHS-2 and FNSA-2. The most highly weighted descriptor in component-1 is FPHS-2, one of the HSA descriptors. In this component, FPHS-2 has a positive weight, while the weight for FNSA-2, one of the CPSP descriptors, has a negative sign. Since the sign of the values of FNSA-2 are also negative, increases in the magnitude of both these descriptors implies increased activity. The key structural features for activity in this SAR trend

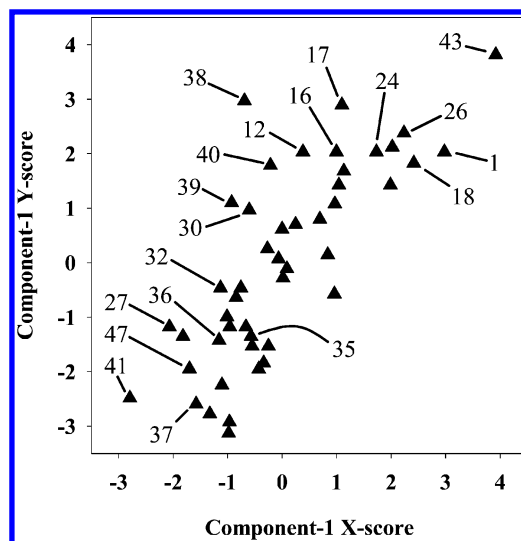


Figure 2. PLS score plot for component-1 of the biarylarnide MIC model incorporating the HSA descriptors. Points representing structures of interest for interpreting the structure-property relationship in component-1 and subsequent components are identified in the plot using the biarylarnide compound numbers.

are the groups at the end of the molecules. The nature of these features can be seen by comparing the four active and four less active compounds shown in Figure 3. The active structures are those with many exposed hydrophobic groups, primarily trifluoromethyl groups, on the end of the molecules. The less active compounds in this trend have fewer hydrophobic groups or features that are actually hydrophilic (e.g., compound 41).

A similar analysis was made of PLS component-2. This component accounts for an additional 12.5% of the variance in the observed MIC (72.2% cumulative). The score-plot for component-2 is shown in Figure 4. The structures that are the subject of this SAR trend were primarily those that were underpredicted by component-1. The model corrects for this using the descriptors NRA and ISP3. Both descriptors have positive weights indicating that increasing values of these descriptors correlates with increased activity. The structural features that are the focus of this component are also on the ends of the molecules. In each case, the molecule possesses either a larger fused ring system or other hydrophobic group. The ISP3 descriptor is responsible for capturing the presence of the methyl ring substituents.

The third PLS component accounts for an additional 7.5% of the variance of the observed MICs (79.7% cumulative) and is dominated by information provided by the two HSAs, FPHS-2 and FPHS-3. The score plot for this component is shown in Figure 5. It is interesting to note that in this trend, the PLS weight for FPHS-2 has a negative sign, while FPHS-3 has a positive sign. Where FPHS-2 uses the sum of all the positive atomic hydrophobic contributions as a weighting factor, FPHS-3 weights the surface area of each atom according to its individual hydrophobic contribution. In this way, the model can now differentiate between the types of groups that contribute to the overall hydrophobicity of the molecule. The purpose of this trend in the model is to further correct for mispredictions in the previous components. For example, compounds 12, 24, and 40 were underpredicted in component-1. Since component-2 corrected for other structural differences, these compounds could not be cor-

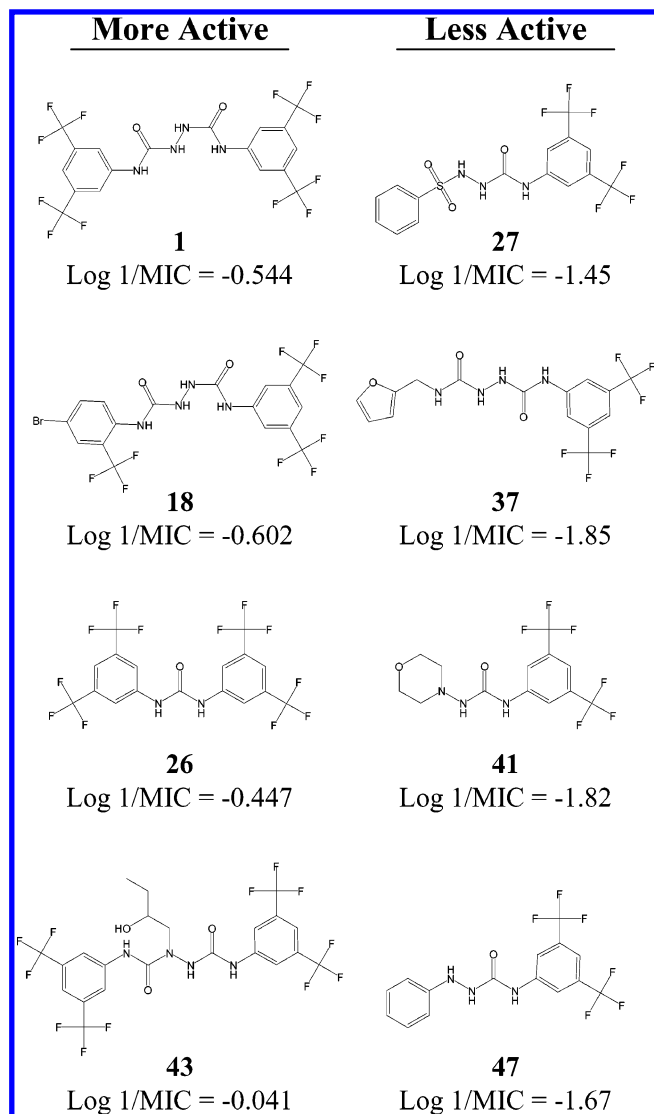


Figure 3. Structures used to evaluate PLS component-1 of the biaryl amide MIC model incorporating an HSA descriptors. The focus of component-1 is the hydrophobicity of the ends of the molecule.

rected until this trend. Other compounds, **35** and **36**, were slightly overpredicted by component-2 and are corrected here. This trend recognizes that other types of groups other than trifluoromethyl and large aromatic systems are hydrophobic. This provided corrections for ring substituents such as bromine and chlorine. Other groups, such as the methoxy substituents on compound **15** were overpredicted in component-1 because the heteroatom takes a negative partial charge and has a relatively large solvent-accessible surface. In component-1, this was primarily related to hydrophobic groups such as trifluoromethyl. However, the methoxy substituents on compound **15** are more hydrophilic than predicted in component-1, so the model corrects for that in component-3.

Component-4 accounts for an additional 6.0% of the variance in the observed MICs (85.8% cumulative) and is the first component that focuses on structural features that are not on the end of the molecule. In this component, two descriptors are highly weighted, ACGD and 1SP3. Both descriptors have positive weights, so increasing values for these descriptors correlate with increasing antibacterial activity. The score plot for component-4 is shown in Figure

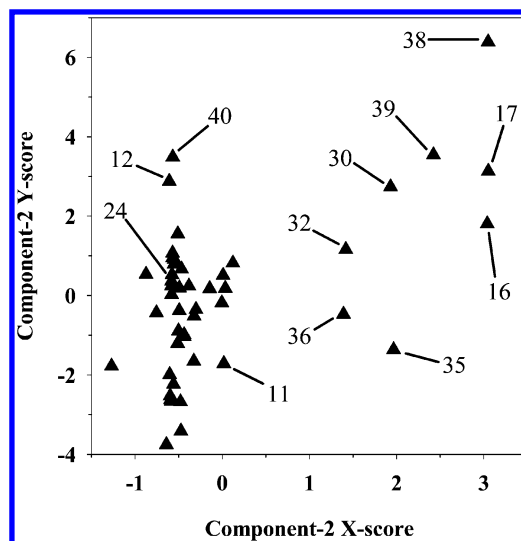


Figure 4. PLS score plot for component-2 of the biaryl amide MIC model incorporating the HSA descriptors. Points representing structures of interest for interpreting the structure–property relationship in component-2 and also component-3 are identified in the plot using the biaryl amide compound numbers.

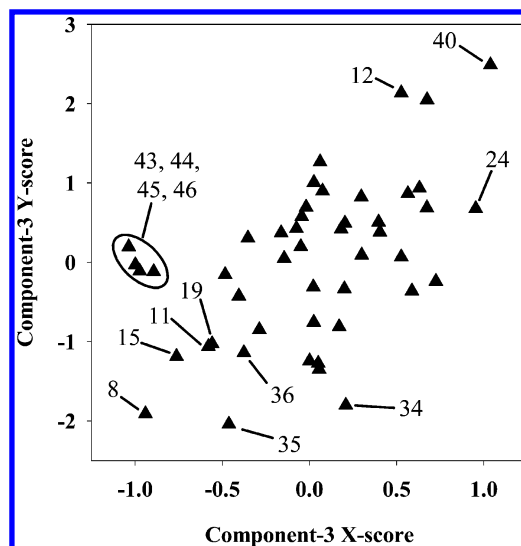


Figure 5. PLS score plot for component-3 of the biaryl amide MIC model incorporating the HSA descriptors. Points representing structures of interest for interpreting the structure–property relationship in component-3 and also component-4 are identified in the plot using the biaryl amide compound numbers.

6. In component-3, several compounds, **43**, **44**, **45**, and **46**, were severely underpredicted (see Figure 5) because they possessed features that were similar to other features responsible for decreased activity. Where the placement of methoxy groups at the ends of the molecule is related to decreased activity, the 2-hydroxybutyl groups attached at the middle amide nitrogen of compounds **43**, **44**, **45**, and **46** are associated with increased activity. Thus, these are corrected in component-4. The ACGD descriptor detects the presence of the hydroxyl, and 1SP3 detects the presence of the terminal methyl of the butyl chain. Together they detect the presence of the 2-hydroxybutyl group at the center of the molecule and allow the model to account for the activity of these compounds.

In the original MIC model, hydrophobicity of the ends of the molecule was identified as a key factor in determining the antibacterial activity of the biaryl amides. The new HSA

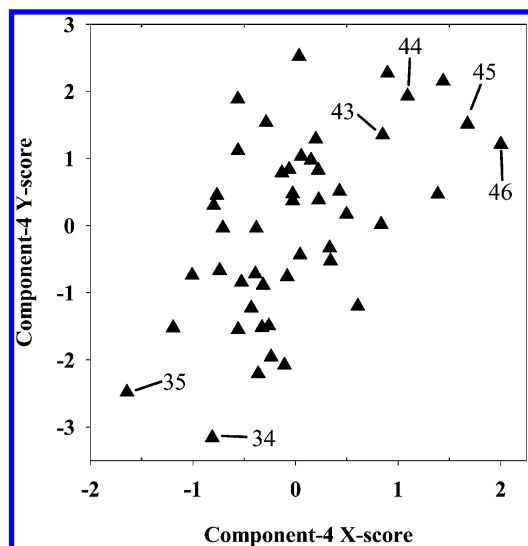


Figure 6. PLS score plot for component-4 of the biarylamine MIC model incorporating the HSA descriptors. Points representing structures of interest for interpreting the structure–property relationship are identified in the plot using the biarylamine compound numbers.

Table 6. Summary of the 5-Variable Blood-Brain Barrier Penetration Model that Excludes the HSAs

descriptor	$R^2 = 0.781, s = 0.375, N = 47, F\text{-value} = 64.9$			variance inflation factor
	regression coefficient	SD of coefficient	<i>t</i> -value	
PPSA-3 ^a	−0.0374	0.00642	−5.82	2.9
WNSA-3 ^a	0.0529	0.00679	7.80	2.3
RNCG ^a	−1.09	0.245	−4.44	1.3
V6P ^b	0.631	0.0799	7.90	2.0
NDB ^c	−0.107	0.0288	−3.70	1.9
Y-intercept	1.26	0.135	9.35	N/A

^a A CPSA descriptor. ^b Valence corrected sixth-order path molecular connectivity index.³ ^c Simple count of double bonds.

descriptors have helped to reinforce that hypothesis. The inclusion of the HSA descriptors has allowed the model to capture that information more succinctly, requiring one less descriptor, while maintaining nearly the same degree of fit to the experimental data.

Blood-Brain Barrier Model (Excluding HSAs). Initial results suggested an equation involving five terms (descriptors) was optimal. This model contained two CPSA descriptors and is summarized in Table 6. A scatterplot illustrating the correlation between the fitted and observed activity values is shown in Figure 7. The model exhibited a good fit to the observed activity values, yielding an R^2 value of 0.781 and a cross-validated R^2 (Q^2) of 0.769 based on the leave-one-out (LOO) method. The model is statistically sound with an overall- F value of 64.9 (compared to a critical- F value of 2.31 with 5 and 91 degrees of freedom and an alpha of 0.05) and the lowest partial- F values of 13.7 (compared to a critical- F of 3.95 with 1 and 91 degrees of freedom and an alpha of 0.05). There is little collinearity between the model descriptors as is evidenced by a maximum VIF of 2.9 and an average VIF of 2.1. To determine if the fit was obtained strictly by chance, the 10-fold randomization experiment yielded an average R^2 of 0.049, with a maximum observed R^2 of 0.091. The maximum correlation coefficient between any of the 10-fold randomized sets of activity values and the original activity values was −0.218. Thus the likelihood

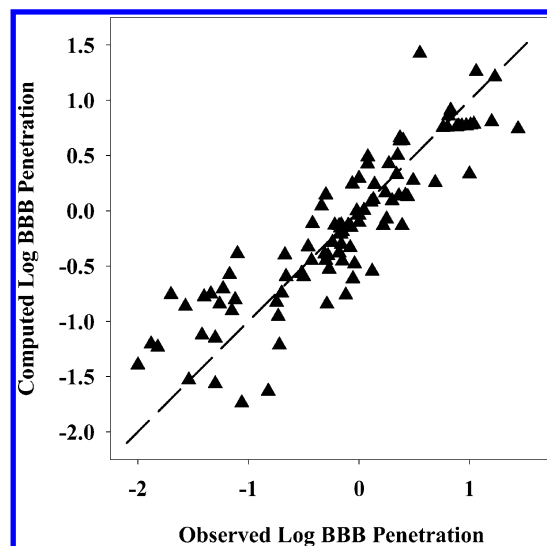


Figure 7. Plot of the calculated and observed BBB partition values based on the 5-variable model that excluded HSA descriptors.

Table 7.

a. Summary of the Results of the PLS Analysis of the 5-Variable Blood-Brain Penetration Model that Excludes the HSAs

component	residual sum of squared error	R^2 (cumulative)	PRESS	Q^2 (cumulative)
1	28.7	0.510	30.6	0.477
2	15.2	0.740	16.8	0.711
3	12.9	0.778	15.5	0.736
4	12.8	0.780	15.3	0.738
5	12.8	0.781	15.3	0.739

b. X-Weights for Each Descriptor in the First Two Components of the PLS Analysis of the 5-Variable Blood-Brain Penetration Model that Excludes the HSAs

descriptor	component-1 X-weight	component-2 X-weight
PPSA-3	−0.536	0.103
WNSA-3	0.578	0.014
RNCG	0.048	−0.606
V6P	−0.014	0.776
NDB	−0.614	−0.142

is low that the model described above resulted from chance correlations. PLS analysis of the model validated all five components, suggesting the model is not overfitted. The model is therefore statistically valid.

A summary of the PLS analysis of this model is shown in Table 7a. All five components were validated by either the PRESS or Q^2 criteria. However, the majority of the variance is accounted for in the first two components, and the weights for the descriptors in these components are shown in Table 7b. The determination of the physical meaning of the model began with an examination of component-1, which accounts for nearly 51% of the total 78.1% variance accounted for by the overall model. The descriptor with the greatest weight in component-1 is the count of the number of double bonds in the molecule (NDB) and has a negative weight in this component. Thus, increases in the number of double bonds in the molecule are correlated with decreasing ability to cross the BBB. Since double bonds are more polar than single bonds, this suggests that the increased number of double

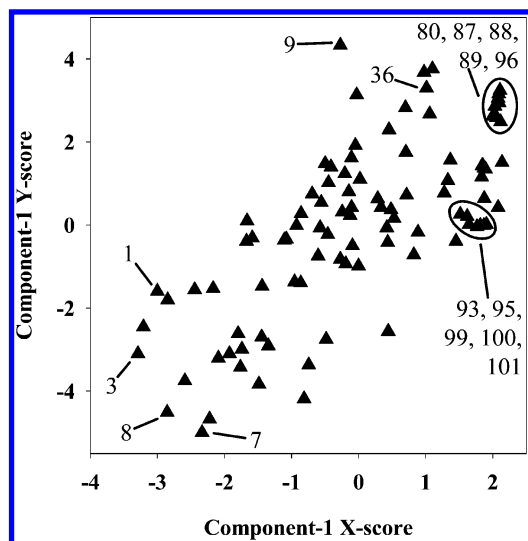


Figure 8. PLS score plot for component-1 of the 5-variable BBB model excluding HSAs. Points representing structures of interest for interpreting both component-1 and component-2 are identified in the graph. The compound numbers shown are the same as those preceded by the designation "BB-" in the discussion of the BBB permeation models.

bonds is increasing the overall hydrophilicity of the molecule and thereby decreasing its ability to cross the BBB. The other two descriptors in the first component measure features related to the solvent accessible surface area of positively charged atoms (PPSA-3) and the negatively charged weighted solvent accessible surface area (WNSA-3). Increases in the value of PPSA-3 are correlated negatively with increasing brain penetration due to the negative coefficient for the assigned weight in this component. The opposite is true for the WNSA-3 descriptor. This descriptor has a negative sign due to the way it is computed. When taken with the positively signed weight in this component, increases in weighted negative surface are correlated with decreasing ability to cross the BBB. The nature of the structural features that are the focus of these three descriptors can be seen by comparing the structures at either end of the structure-property relationship (SPR) trend captured by component-1. The score plot for component-1 is shown in Figure 8. Points representing the structures used to generate the SPR interpretation are identified on the plot, while the structure diagrams are provided in Figure 9. The small hydrocarbon compounds are devoid of double bonds, and the solvent accessible surface of these molecules is dominated by hydrogen atoms which take on small positive partial charges. Conversely, molecules at the other end of the SPR trend have numerous double bonds, and there is a much greater amount of surface area associated with atoms that possess partial negative charges (heteroatoms and carbon atoms at positions of unsaturation). The two CPSA descriptors and the count of double bonds appear to be working together to differentiate small hydrophobic compounds from large polar compounds. Both molecular size and hydrophobicity are being captured in this first component. This is consistent with the general observations that compounds which permeate the blood-brain barrier and that do not depend on any type of active transport mechanism do so by passive diffusion.⁵⁴⁻⁵⁶

Component-2 accounts for an additional 23.1% of the variance in the data (74.1% cumulative). The descriptors that are highly weighted in this component are the valence

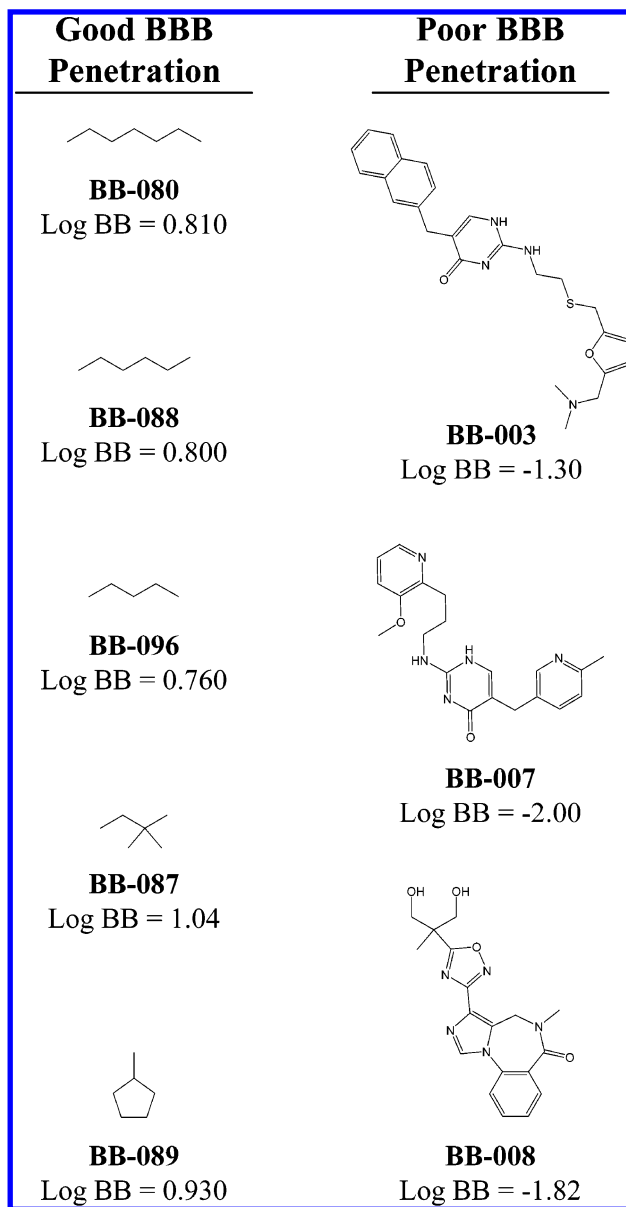


Figure 9. Structures used to evaluate PLS component-1 of the initial (non-HSA) blood-brain barrier penetration model. The focus of component-1 is the molecular size and hydrophobicity combined.

corrected sixth-order path molecular connectivity descriptor (V6P) and, to a lesser degree, the relative negative charge solvent accessible surface area (RNCG). The greater weight is given to V6P, and it is positive. This results in a positive correlation between increasing values for V6P and increasing Log BB values. The RNCG has a negative coefficient indicating that increasing values for the relative negative charge on the solvent accessible surface area is correlated with decreasing Log BB values. Component-2 corrects for both over- and underpredictions of the Log BB values for certain structures made by component-1. Examples of these structures are provided in Figure 10, and points representing these compounds are identified in the score plot shown in Figure 11. In component-1, compounds such as propan-1-ol, 2-methyl-propan-1-ol, and propan-2-ol (compound numbers **BB-99**, **BB-93**, and **BB-100**, respectively) were predicted to have higher permeability than is observed experimentally, based on the descriptors that were highly weighted in that component. These structures are small and

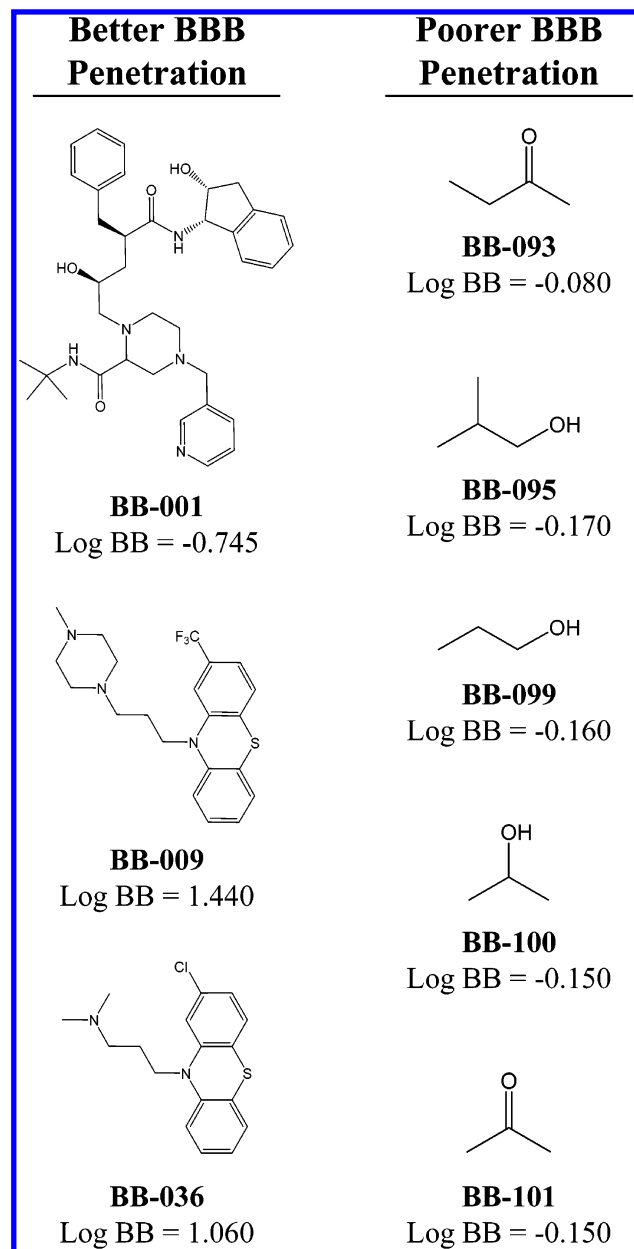


Figure 10. Structures used to evaluate PLS component-2 of the initial (non-HSA) blood-brain barrier penetration model. The focus of component-2 is on molecules that were poorly fit by component-1. The larger molecules are more hydrophobic and penetrate the BBB better than expected based on component-1. The small molecules are more hydrophilic and penetrate the BBB more poorly than expected by component-1.

contain no double bonds. However, it was necessary for the model to correct for the presence of a heteroatom. These structures take on a value of zero for the V6P descriptor, which identifies them as small molecules, and the presence of the oxygen heteroatom on these small molecules give them very large values for RNCG. Conversely, the model employs both V6P and RNCG to identify some large but hydrophobic molecules that were underpredicted by the SPR trend in component-1. These structures take on relatively large values for V6P and much smaller values for RNCG since the negative partial charges are more widely dispersed throughout the molecule. These particular compounds are more hydrophobic than predicted by component-1, and their greater blood-brain barrier penetration is accounted for in component-2. Since the remaining components account only for an

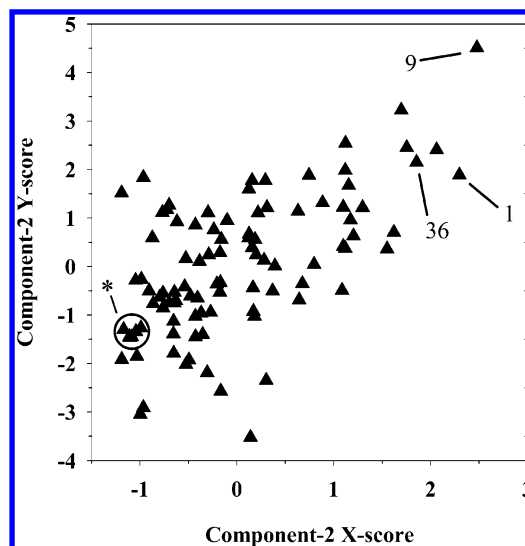


Figure 11. PLS score plot for component-2 of the 5-variable BBB model excluding HSAs. Points representing structures of interest for interpreting component-2 are identified in the graph. The compound numbers shown are the same as those preceded by the designation "BB-" in the discussion of the BBB permeation models.

Table 8. Summary of the 4-Variable Blood-Brain Penetration Model that Includes the HSA Descriptor PNHS-3

descriptor	$R^2 = 0.778, s = 0.375, N = 47, F\text{-value} = 80.7$			variance inflation factor
	regression coefficient	SD of coefficient	<i>t</i> -value	
WNSA-3	0.0443	0.00710	6.24	2.5
V4P ^a	0.244	0.0343	7.13	1.7
NDB	-0.132	0.0262	-5.05	1.5
PNHS-3 ^b	0.0302	0.00435	6.93	1.7
y-intercept	0.534	0.0733	7.28	N/A

^a Valence-corrected fourth-order path molecular connectivity index.³

^b An HSA descriptor.

additional 4% of the variance in the data, attention was turned instead to the model containing HSAs.

Blood-Brain Barrier Model (Including HSAs). When the HSA descriptors were added to the descriptor pool, a new equation resulted indicating only four descriptors were necessary. This model is summarized in Table 8. The fit plot for the new 4-variable model is shown in Figure 12. The model contains one of the new HSA descriptors and one CPSA. The other two descriptors are the valence-corrected fourth-order path molecular connectivity index³ (V4P) and a count of the number of double bonds in the molecule (NDB). The model exhibited a good fit to the observed activity values, yielding an R^2 value of 0.778 and a cross-validated R^2 (Q^2) of 0.748 based on the leave-one-out (LOO) method. The model is statistically sound with an overall- F value of 80.7 (compared to a critical- F value of 2.47 with 4 and 92 degrees of freedom and an alpha of 0.05) and the lowest partial- F values of 25.5 (compared to a critical- F of 3.94 with 1 and 92 degrees of freedom and an alpha of 0.05). There is little collinearity between the model descriptors as is evidenced by a VIF of 2.5 and an average VIF of 1.9. The test for chance correlation was carried out as before. The 10-fold randomization experiment yielded an average R^2 of 0.038, with maximum observed R^2 of 0.094. The maximum correlation coefficient obtained between any of the 10-fold randomized sets of activity values and the original

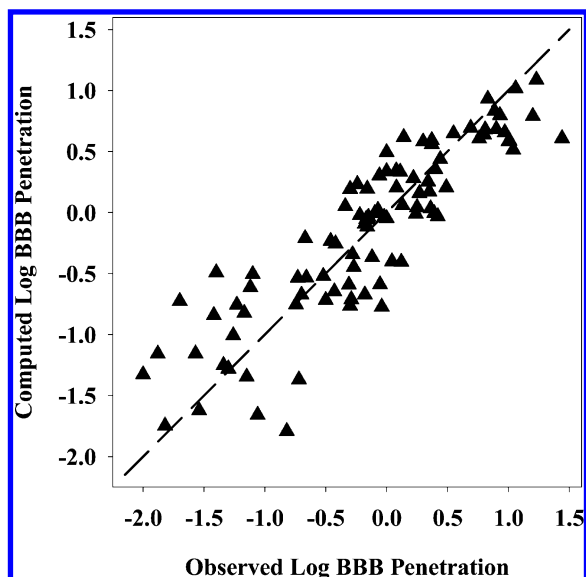


Figure 12. Plot of the calculated and observed BBB partition values based on the 4-variable model that included the HSA descriptor PNHS-3.

Table 9.

A. Summary of the Results of the PLS Analysis of the 4-Variable Blood-Brain Penetration Model that Includes the HSA Descriptor PNHS-3

component	residual sum of squared error	R^2 (cumulative)	PRESS	Q^2 (cumulative)
1	22.4	0.617	23.8	0.592
2	13.9	0.762	15.4	0.737
3	13.0	0.778	14.8	0.747
4	13.0	0.778	14.7	0.748

b. X-Weights for Each Descriptor in the First Two Components of the PLS Analysis of the 4-Variable Blood-Brain Penetration Model that Includes the HSA Descriptor

descriptor	component-1 X-weight	component-2 X-weight
WNSA-3	0.535	-0.128
V4P	-0.085	0.974
NDB	-0.567	-0.080
PNHS-3	0.621	0.170

activity values was 0.225. Thus, the likelihood is low that the model described above resulted from chance correlations. PLS analysis of the model validated all four components, suggesting the model is not overfitted. The model is therefore statistically valid.

As before, PLS was used to extract the physical interpretation of the new model. These results are summarized in Table 9. The first component of this model explains about 62% of the variance in the observed Log BB values. The score plot for this component is shown in Figure 13. Three descriptors are highly weighted in component-1, with the HSA descriptor having the largest weight. The other two descriptors are a CPSA (WNSA-3) and count of the number of double bonds in the molecule (NDB). The latter descriptor has a negative coefficient indicating increasing numbers of double bonds results in decreasing ability to penetrate the BBB as was also found in the 5-variable model described above. WNSA-3, as in the 5-component model, has a positive weight, and therefore increasing values of WNSA-3 are correlated with

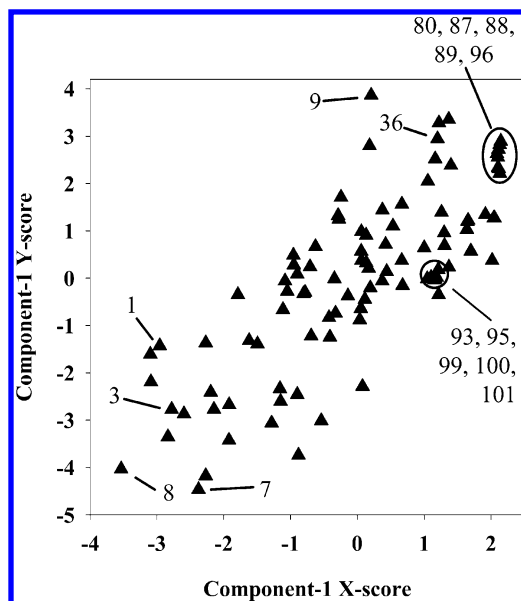


Figure 13. PLS score plot for component-1 of the 4-variable BBB model that includes the HSA descriptor PNHS-3. Points representing structures of interest for interpreting both component-1 and component-2 are identified in the graph. These are the same structures considered in the non-HSA model case. The compound numbers shown are the same as those preceded by the designation "BB-" in the discussion of the BBB permeation models.

increasing Log BB values, since the sign of the value of this descriptor is negative. This component shows the same pattern observed in the first component of the non-HSA model. Small hydrophobic molecules are favored for penetration over large hydrophilic molecules. Here, the HSA descriptor has improved the ability of the model to capture this information, resulting in a nearly 22% increase in the variance accounted for by the first component.

Component 2 accounts for an additional 14.5% (76.2% cumulative). The score plot for component-2 is shown in Figure 14. Again, nearly the same pattern is observed for component-2 in this model as was observed for component-2 of the non-HSA model described above. Some molecules, most notably the small hydrophilic ones (see Figure 13), are overestimated by the first component, and the model corrects for that in component-2. Likewise, large hydrophobic molecules are underestimated in component-1, and the model makes those corrections here. These are the same structure-property trends observed for the model that excluded the HSA descriptors. Thus, the two models provide the same SPR information. The important difference is that the HSA descriptor captures the key information concerning the hydrophobicity and size of the molecules better, yielding a more compact and easier to interpret model.

General Descriptor Characteristics. Since the HSA descriptors have been found to successfully capture structural information concerning molecular features responsible for hydrophobic intermolecular interactions, attention was turned to evaluation of their more general characteristics. Two characteristics that seemed most important to consider were the sensitivity of the descriptors to changing molecular conformation and the correlation of the HSAs with other conventional QSAR descriptors.

The use of the solvent-accessible surface area in the construction of the HSA descriptors implies that these

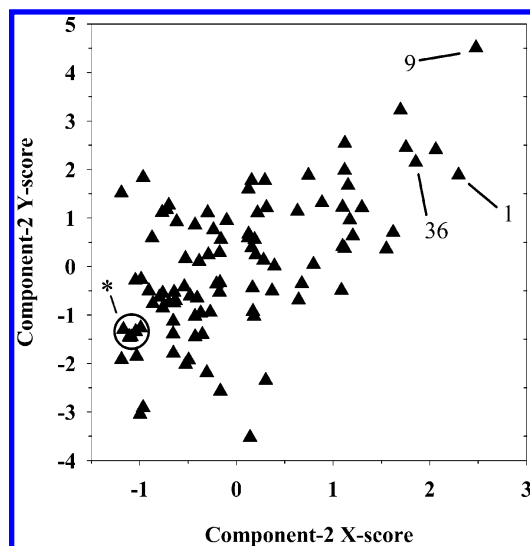


Figure 14. PLS score plot for component-2 of the 4-variable BBB model that includes the HSA descriptor. Points representing structures of interest for interpreting component-2 are identified in the graph. These are the same structures considered in the non-HSA model case. The compound numbers shown are the same as those preceded by the designation “BB-” in the discussion of the BBB permeation models.

descriptors will yield different values for a given compound depending on the conformation considered. There are both positive and negative aspects to this characteristic. The ability for the descriptor to be sensitive to a change in geometry may be an important factor in helping to generate a physically meaning model. In fact, the geometric sensitivity of the related CPSA descriptors has proven to be an advantage in some situations.³³ However, geometric sensitivity may make it more difficult to obtain accurate QSAR predictions unless special attention is paid to the generation of appropriate 3D coordinates of structures for which predictions are sought. With these issues in mind, we evaluated the set of HSA descriptors by selecting three structures from the blood-brain barrier penetration model training set. The whole data set was evaluated for flexibility as characterized by a count of rotatable bonds using the Dragon program.⁵⁷ The three structures chosen (compounds **BB-002**, **BB-031**, and **BB-056**, see Figure 15) possess 21, 10, and 2 rotatable bonds, respectively. They were selected to represent the range of flexibility of the structures in the data set. Each structure was submitted for conformational analysis using Spartan '02⁵⁸ and the MMFF molecular mechanics force field (all other settings set to default). This procedure performs a conformational search to find all unique, energy minimized conformers of a molecule that are within 10 kcal/mol of the lowest conformation found. A total of 65, 85, and 13 unique conformers were found for **BB-002**, **BB-031**, and **BB-056**, respectively. The suite of 25 HSA descriptors was calculated for each conformer with the resultant descriptor values being used to test the sensitivity to changes in molecular conformation. The results from the conformational analysis are presented in Table 10. It is interesting to note that the HSAs describing the hydrophilic character of the molecules exhibit larger deviations than the HSAs that contain hydrophobic information. This is one result of having fewer heteroatoms (thus fewer hydrophilic atoms) in the molecules. Heteroatoms only comprise 10–25% of the total number of atoms in the

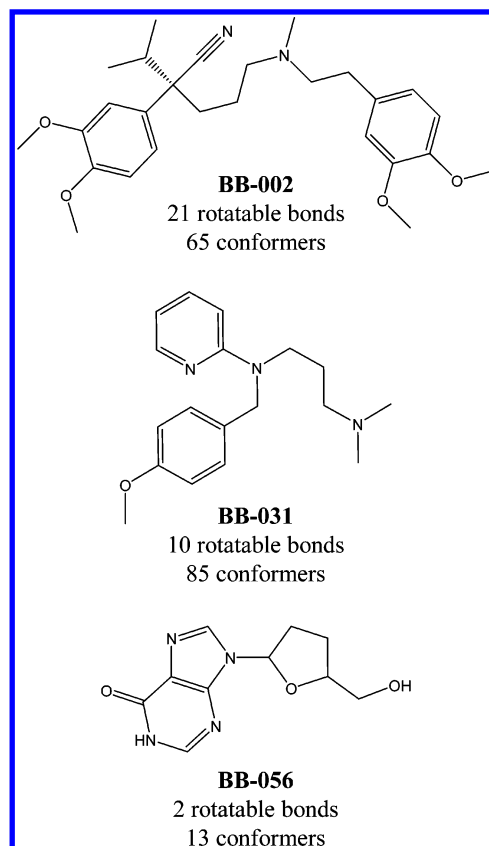


Figure 15. Diagrams of the structures evaluated in the HSA conformation sensitivity analysis experiment. The conformer count shown corresponds to the number of unique conformers identified within 10 kcal/mol of the lowest energy geometry found.

test molecules. Therefore, a change in the surface area for one heteroatom will have more of an influence on the overall value of the hydrophilic HSAs. The HSA descriptors which provide hydrophobic information are calculated using a larger majority of the atoms in the molecules. Thus, the changes in the atomic surface areas will be averaged out, thereby lessening the dependence of conformation on these descriptors.

The HSAs which show the largest variance with changing conformation are RPHS and RNHS. Since these descriptors are encoding information about one single atom type, they are more susceptible to changes in molecular conformation. As an example, RNHS has a 5-fold (533%) variation for conformers of BB-031. The specific atom that this descriptor encodes is the tertiary nitrogen atom situated next to the pyridine ring. Since the surface area for this atom can range from zero to 0.75 Å², it causes the RNHS descriptor to have values on the range of 0.00–0.13. Furthermore, since the magnitude of this descriptor is relatively small, a larger percent change is seen when compared to the other HSA descriptors.

As an overall comparison, the CPSA descriptors were also calculated for each conformer of BB-002, BB-031, and BB-056. The percent changes seen for the HSA descriptors were comparable to the percent changes observed for the CPSA descriptors. Since the degree of conformational sensitivity of the HSAs is similar to that of the CPSAs, the utility of the new descriptors should not be adversely affected. While some variation is expected for any set of descriptors which contain surface area information, these results demonstrate

Table 10. Conformational Analysis Showing the Change Observed in the HSA Descriptors as a Function of Changes in the Geometry of Three Molecules in the Blood Brain Barrier Data Set

HSA	% change ^a		
	BB-002 (65 conformers)	BB-031 (85 conformers)	BB-056 (13 conformers)
PPHS-1	14.9	8.8	7.4
PPHS-2	15.0	8.8	7.5
PPHS-3	24.0	10.6	7.0
PNHS-1	33.0	25.9	27.3
PNHS-2	33.2	25.9	27.4
PNHS-3	38.8	26.6	26.4
DHS-1	17.4	11.4	30.6
DHS-2	14.6	7.7	4.0
DHS-3	24.2	8.4	9.7
FPHS-1	3.0	2.8	9.9
FPHS-2	3.0	2.7	9.6
FPHS-3	13.0	4.9	9.3
FNHS-1	24.1	26.4	24.9
FNHS-2	23.9	26.8	24.9
FNHS-3	30.2	27.9	23.8
WPHS-1	28.5	15.4	6.7
WPHS-2	28.5	16.0	6.7
WPHS-3	36.6	17.4	7.0
WNHS-1	46.2	28.4	30.2
WNHS-2	46.0	28.4	29.7
WNHS-3	51.1	28.0	29.4
RPH	0.0	0.0	0.0
RNH	0.0	0.0	0.0
RPHS	88.1	48.2	23.3
RNHS	30.3	533.5	2.8

^a % change = (descriptor range)/(descriptor average) × 100.

that care must be taken with conformations of the molecules in a particular data set because they could have an impact on the calculated descriptor values.

Whenever a new QSAR descriptor is proposed, one needs to consider whether another descriptor already exists that performs the same function. While it is not practical to try to examine all the previously existing descriptors, it is important to evaluate those that are closely related on a conceptual basis. There also may be other descriptors that have an unexpected close relationship to the newly proposed descriptor simply due to a close relationship between the types of molecular features upon which they are based (e.g., branching and distribution of charge in a molecule). Since the HSA descriptors are conceptually related to the CPSA descriptors, a comparison of these classes was deemed important. Additionally, ADAPT provides a wide selection of other types of descriptors. We therefore chose to perform a correlation analysis using all 133 descriptors, in addition to the 25 HSAs, computed for the blood-brain barrier penetration data set. This data set seemed particularly appropriate since it represents a diverse collection of structures with a wide variety of different molecular features characteristic of drug-like compounds.

The correlation coefficient⁵⁹ for all pairs of descriptors were computed using the Minitab program.⁴⁶ These were then exported to a Microsoft Excel⁶⁰ spreadsheet for further analysis. All instances of high correlation between a given HSA descriptor and any other descriptors (including other HSAs) were noted. For the purpose of this analysis, a high correlation was one that yielded a correlation coefficient greater than or equal to a value of 0.80. While this value was chosen arbitrarily, it represents a level of collinearity that would begin to pose stability problems within QSAR

Table 11.

a. Summary of the Evaluation of the Degree of Correlation of the Class of HSA Descriptors with Other Well Known Classes of Descriptors Commonly Used in QSAR and QSPR Studies

descriptor class	count of instances of high correlation	count of number of descriptors in class	av number of correlated descriptors
electronic	6	1	6.00
topological	131	61	2.15
geometric	33	16	2.06
CPSA	46	25	1.84
HPSA	41	25	1.64
fragment	30	19	1.58
Hbond	1	11	0.09

b. Summary of the Evaluation of the Degree of Correlation of the Individual HSA Descriptors with the Set of 158 Descriptors, Including Other HSAs, Considered in This Study

HSA descriptor label	count of highly correlated descriptors	HSA descriptor label	count of highly correlated descriptors
PPHS-1	46	FNHS-2	3
PPHS-2	48	FNHS-3	5
PPHS-3	1	WPHS-1	0
PNHS-1	0	WPHS-2	47
PNHS-2	1	WPHS-3	34
PNHS-3	1	WNHS-1	6
DPHS-1	33	WNHS-2	8
DPHS-2	0	WNHS-3	10
DPHS-3	2	RPH	0
FPHS-1	1	RNH	0
FPHS-2	38	RPHS	0
FPHS-3	2	RNHS	0
FNHS-1	2		

models. In this respect, it seemed to be an appropriate critical value.

The results of the correlation analysis are summarized in Table 11. We chose to examine the correlations in a number of different ways. Overall, the HSA descriptor class is not highly correlated with other general classes of descriptors. Excluding the electronic class of descriptors (because there is only one representative), the HSAs are most highly correlated to the topological descriptors, followed by the geometric class, and then the CPSAs (see Table 11a). The average number of correlated descriptors for the topological class was 2.1. This value is quite low considering that 131 topological descriptors were evaluated. The low correlation with the CPSA class was interesting because of the similarity in the method of computation. The individual HSAs also seem to encode different information within their own class.

The correlation of individual HSA descriptors were also examined (see Table 11b). Only 6 of the 25 HSA descriptors were highly correlated with more than 10 other descriptors. These were PPHS-1, PPHS-2, DHS-1, FPHS-2, WPHS-2, and WPHS-3. In general, these six descriptors show the same pattern of correlation with other QSAR descriptors, although each to a different extent. A common theme is a correlation with descriptors that include a measure of molecular size. For example, molecular volume and surface area, molecular weight, moments of inertia, and radius of gyration are among the descriptors found to correlate most with these six HSAs. Several of the CPSA descriptors, not surprisingly, are included in this set. More unexpected were high correlation with the molecular connectivity (MC) indices. Both the simple and valence-corrected zeroth-order through the fifth-

order path MC indices were correlated above 0.8 with several of the six HSAs. Outside of these six, the other HSAs are correlated only with other HSAs. This result is also not surprising since some of the HSAs are derived from combinations of others.

Since the level of collinearity found between the HSAs and other descriptors is relatively low, it suggests that the HSAs, as a class, are capturing unique structural information. The correlation that does exist between some of the HSAs and other descriptors allows us to draw some interesting conclusions. One important conclusion is that a strict physical interpretation should not be assigned to any given molecular descriptor. For example, molecular size can be encoded a number of different ways. However, each descriptor may be measuring molecular size from a different perspective. The CPSAs were designed to measure molecular size from the point of view of solvent-accessible surface area of polar features. The HSAs were designed to measure molecular size related to solvent-accessible surface area of hydrophobic and hydrophilic features. These two perspectives are closely related, and so descriptors from either class may be included in a model strictly on a statistical basis. However, how one interprets the presence of a given descriptor in a model should be based on the context of the data set. This notion is further reinforced by the observation of the high correlation of the molecular connectivity descriptors with some of the HSAs. Since the HSAs have been shown to capture aspects of both molecular size and hydrophobicity, it is reasonable to assume that the highly correlated MC descriptors are encoding similar information. Thus, another conclusion that can be drawn is that some of the demonstrated utility of the MC descriptors may be related to their ability to capture structural information that is related to hydrophobicity, in addition to the more obvious ability to capture information regarding size, shape, and degree of flexibility.

CONCLUSIONS

The HSA descriptors have been found to be unique measures of the molecular structure features that are responsible for hydrophobic and hydrophilic intermolecular interactions. They provide this benefit better than other descriptors that we have evaluated. Their ability to capture the information intended has been demonstrated by examining the underlying structure–activity and structure–property relationship that have an established hydrophobic component. In the case of the biarylamine QSAR model, previous work generated a hypothesis that antibacterial activity was a function of the hydrophobicity of the ends of the molecule. The HSAs were shown to successfully capture that information and to do so in a more condensed fashion allowing for the development of a more concise model of similar quality. The property of blood-brain barrier penetration is known to have a very important hydrophobic component. Here the HSAs were also shown to successfully capture information related to the hydrophobic interactions between the molecules in question and the membrane components forming the blood-brain barrier.

Finally, the HSAs were shown to be unique descriptors that capture this type of hydrophobic information in a way that is different from and more effective than other descriptors examined. While certain members of the HSA descriptor

family have been shown to be sensitive to changes in conformation, the degree of sensitivity is not greater than other descriptors which have found widespread use. In some cases, such conformational sensitivity has even been shown to be an advantage. The HSA descriptors provide the ability to capture information regarding hydrophobic intermolecular interactions in a more complete and detailed fashion, and help in the identification of the true SAR relationship coded in QSAR models in which they are included.

Supporting Information Available: The structures and the observed and computed property values for the 47 biaryl- amides used to develop the MIC model and the 97 drugs used to develop the blood-brain barrier penetration models. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York NY, 2000; pp 141–383.
- (2) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: Weinheim, Federal Republic of Germany, Methods and Principles in Medicinal Chemistry, Vol. 11, 2000.
- (3) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, NY, 1976.
- (4) Ivanciuc, O.; Balaban, A. T. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach, The Netherlands, 1999.
- (5) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, NY, 1980; pp 321–347.
- (6) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.; AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (7) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure–Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492–504.
- (8) Kaliszan, R. *Quantitative Structure Chromatographic Retention Relationships*; Wiley & Sons: New York, NY, 1987; pp 118–119.
- (9) Stanton, D. T. On the Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423–1433.
- (10) Stanton, D. T.; Madhav, P. J.; Wilson, L. J.; Morris, T. W.; Hershberger, P. M.; Parker, C. N. Development of a Quantitative Structure–Activity Relationship Model for Inhibition of Gram-positive Bacterial Cell Growth by Biarylaminides. *J. Chem. Inf. Comput. Sci.*, in press.
- (11) Stanton, D. T.; Jurs, P. C. Computer-Assisted Study of the Relationship Between Molecular Structure and Surface Tension of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109–115.
- (12) International Union of Pure and Applied Chemistry, Chemistry and Human Health Division. Glossary of Terms Used in Computational Drug Design (IUPAC Recommendations 1997). <http://www.iupac.org/reports/1997/6905vandewaterbeemd/glossary.html> (accessed January 2004).
- (13) Leo, A. J. Calculating log P_{oct} from Structures. *Chem. Rev.* **1993**, *93*, 1281–1306.
- (14) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71*, 525–616.
- (15) Todeschini, R.; Vighi, M.; Finizio, A.; Gramatica, P. 3D-Modelling and Prediction by WHIM Descriptors. Part 8. Toxicity and Physico-Chemical Properties of Environmental Priority Chemicals by 2D-TI and 3D-WHIM Descriptors. *SAR QSAR Environ. Res.* **1997**, *7*, 173–193.
- (16) Gaillard, P.; Carrupt, P.; Testa, B. The Conformation-Dependent Lipophilicity of Morphine Glucuronides as Calculated From Their Molecular Lipophilicity Potential. *Bioorg. Med. Chem. Lett.* **1994**, *4*, 737–742.
- (17) Takacs-Novaki, K.; Nagy, P.; Jozani, M.; Orfi, L.; Dunn, W. J. III; Szasz, G. Relationship Between Partitioning Properties and (Calculated) Molecular Surface. SPR Investigation of Imidazoquinazoline Derivatives. *Acta Pharm. Hungarica* **1992**, *61*, 55–64.
- (18) Debruijn, J.; Hermens, J. Relationships Between Octanol Water Partition-Coefficients and Total Molecular-Surface Area and Total Molecular Volume of Hydrophobic Organic Chemicals. *Quant. Struct.-Act. Relat.* **1990**, *9*, 11–21.
- (19) Dunn, W. J. III Surface Area and Hydrophobicity of Small Molecules. *Prog. Clin. Biol. Res.* **1989**, *291*, 47–51.

- (20) Koehler, M. G.; Grigoras, S.; Dunn, W. J. III The Relationship Between Chemical Structure and the Logarithm of the Partition Coefficient. *Quant. Struct.-Act. Relat.* **1988**, *7*, 150–159.
- (21) Camilleri, P.; Watts, S. A.; Boraston, J. A. A Surface Area Approach to Determination of Partition Coefficients. *J. Chem. Soc., Perkin Trans.* **1988**, *2*, 1699–1707.
- (22) Dunn, W. J. III; Koehler, M. G.; Grigoras, S. The Role of Solvent-Accessible Surface Area in Determining Partition Coefficients. *J. Med. Chem.* **1987**, *30*, 1121–1126.
- (23) Doucette, W. J.; Andren, A. W. Correlation of Octanol/Water Partition Coefficients and Total Molecular Surface Area for Highly Hydrophobic Aromatic Compounds. *Environ. Sci. Technol.* **1987**, *21*, 821–824.
- (24) Iwase, K.; Komatsu, K.; Hirono, S.; Nakagawa, S.; Moriguchi, I. Estimation of Hydrophobicity Based on the Solvent-Accessible Surface Area of Molecules. *Chem. Pharm. Bull.* **1985**, *33*, 2114–2121.
- (25) Yalkowsky, S. H.; Valvani, S. C. Partition Coefficients and Surface Areas of Some Alkylbenzenes. *J. Med. Chem.* **1976**, *19*, 727–728.
- (26) Leo, A.; Hansch, C.; Jow, P. Y. C. Dependence of Hydrophobicity of Apolar Molecules on Their Molecular Volume. *J. Med. Chem.* **1976**, *19*, 611–615.
- (27) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 2. Prediction of Partition Coefficient by Atom-Additive Approach Based on Atom-Weighted Solvent Accessible Surface Areas. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1058–1067.
- (28) Masuda, T.; Jikihara, T.; Nakamura, K.; Kimura, A.; Takagi, T.; Fujiwara, H. Introduction of Solvent-Accessible Surface Area in the Calculation of the Hydrophobicity Parameter log *P* from an Atomistic Approach. *J. Pharm. Sci.* **1997**, *86*, 57–63.
- (29) Nicolau, D. V., Jr.; Nicolau, D. V. Database Comprising Biomolecular Descriptors Relevant to Protein Adsorption on Microarray Surfaces. *SPIE Proceedings* **2002**, *3*, 109–116.
- (30) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464–477.
- (31) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.
- (32) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Descriptors in Computer-Assisted Quantitative Structure–Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329.
- (33) Stanton, D. T.; Dimitrov, S.; Grancharov, V.; Mekenyan, O. G. Charged Partial Surface Area (CPSA) Descriptors. QSAR Applications. *SAR QSAR Environ. Res.* **2002**, *13*, 341–351.
- (34) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
- (35) Saxena, A. K.; Ram, S.; Saxena, M.; Singh, N.; Prathipati, P.; Jain, P. C.; Singh, H. K.; Anand, N. QSAR Studies in Substituted 1,2,3,4,6,7,12,12a-octa-hydropyrazino[2',1':6,1]pyrido[3,4-b]indoles – A Potent Class of Neuroleptics. *Bioorgan. Med. Chem.* **2003**, *11*, 2085–2090.
- (36) Deretey, E.; Feher, M.; Schmidt, J. M. Rapid Prediction of Human Intestinal Absorption. *Quant. Struct.-Act. Relat.* **2002**, *21*, 493–506.
- (37) Schmidt, T. J.; Heilmann, J. Quantitative Structure-Cytotoxicity Relationships of Sesquiterpene Lactones derived from partial charge (Q)-based fractional Accessible Surface Area Descriptors (Q_frASAs). *Quant. Struct.-Act. Relat.* **2002**, *21*, 276–287.
- (38) Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennett, K. P.; Cramer, S.; Tugcu, N. Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1347–1357.
- (39) Godden, J. W.; Xue, L.; Bajorath, J. Classification of Biologically Active Compounds by Median Partitioning. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1263–1269.
- (40) Labute, P.; Williams, C.; Feher, M.; Sourial, E.; Schmidt, J. M. Flexible Alignment of Small Molecules. *J. Med. Chem.* **2001**, *44*, 1483–1490.
- (41) Stuper, A. J.; P. C. Jurs. ADAPT: A Computer System for Automating Data Analysis using Pattern-Recognition Techniques. *J. Chem. Inf. Comput. Sci.* **1976**, *2*, 99–105.
- (42) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, R. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; pp 103–129.
- (43) Stanton, D. T. Development of a Quantitative Structure–Property Relationship Model for Estimating Normal Boiling Points of Small Multifunctional Organic Molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 81–90.
- (44) Sutter, J. M.; Jurs, P. C. Selection of Molecular Descriptors for Quantitative Structure–Activity Relationships. *Data Handl. Sci. Technol.* **1995**, *15*, 111–132.
- (45) Luke, B. T. An Overview of Genetic Methods. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: New York, NY, 1996; pp 35–66.
- (46) Minitab, Release 14, Beta version-3, Minitab, Inc., State College, PA, USA.
- (47) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
- (48) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometric Methods In Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, NY, 1995; pp 309–318.
- (49) Rose, K.; Hall, L. H.; Kier, L. B. Modelling Blood-Brain Barrier Partitioning Using the Electropotential State. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651–666, and references therein.
- (50) Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied Linear Statistical Models*, 4th ed.; McGraw-Hill: Boston, MA, 1996; pp 229–230.
- (51) Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied Linear Statistical Models*, 4th ed.; McGraw-Hill: Boston, MA, 1996; pp 268–268.
- (52) Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied Linear Statistical Models*, 4th ed.; McGraw-Hill: Boston, MA, 1996; pp 385–388.
- (53) Stanton, D. T.; Egolf, L. M.; Jurs, P. C. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306–316.
- (54) Audus, K. L.; Chikhale, P. J.; Miller, D. W.; Thompson, S. E.; Borchardt, R. T. Brain uptake of drugs: The influence of chemical and biological factors. *Adv. Drug. Res.* **1992**, *23*, 1–64.
- (55) Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffiths, R.; Jones, M.; Rana, K. K.; Saunders, D.; Smith, I. R.; Sore, N. E.; Wilks, T. J. Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H3 receptor histamine antagonists. *J. Med. Chem.* **1988**, *31*, 656–671.
- (56) Gratten, J. A.; Abraham, M. H.; Bradbury, M. W.; Chadha, H. S. Molecular factors influencing drug transfer across the blood-brain barrier. *J. Pharm. Pharmacol.* **1997**, *49*, 1211–1216.
- (57) Dragon, version 3.0, 2003, Todeschini, R.; Consonni, V.; Pavan, M., Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milano, Italy. <http://www.disat.unimib.it/chm/>.
- (58) Spartan '02, Build 119 X11 for IRIX 6.5, Wavefunction, Inc., Irvine, CA.
- (59) Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied Linear Statistical Models*, 4th ed.; McGraw-Hill: Boston, MA, 1996; pp 633–634.
- (60) Microsoft Excel-2000 for Windows, Microsoft Corp, Redmond, WA, U.S.A.

CI034284T