

Modeling Drug Albumin Binding Affinity with E-State Topological Structure Representation

L. Mark Hall

Hall Associates Consulting, 2 Davis Street, Quincy, Massachusetts 02170-2818

Lowell H. Hall*

Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170

Lemont B. Kier

Department of Medicinal Chemistry, School of Pharmacy, Virginia Commonwealth University,
Richmond, Virginia 23298

Received June 15, 2003

The binding affinity to human serum albumin for 94 drugs was modeled with topological descriptors of molecular structure, using as experimental data the HPLC chromatographic retention index [$\log k(\text{HSA})$] on immobilized albumin. The electrotopological state (E-State) along with the molecular connectivity chi indices provided the basis for a satisfactory model: $r^2 = 0.77$, $s = 0.29$, $q^2 = 0.70$, $s_{\text{press}} = 0.33$. The 10% leave-group-out (LGO) cross-validation method yielded $q^2 (= r^2_{\text{press}}) = 0.69$. Further, the model was tested on a 10 compound external validation set, yielding a mean absolute error, $\text{MAE} = 0.31$; $q^2 (= r^2_{\text{press}}) = 0.74$. MDL QSAR software was used for setting up the data set, creation of combination descriptors, modeling, and database management. All the statistical tests indicate that the topological model is useful for property estimation. Internal and external validation methods were used, and the results indicate that the model is useful for prediction. Randomizations of the activity values also indicate statistically sound models are very different from random statistics. The model indicates that positive factors for binding affinity include electron accessibility and the number of aromatic rings, aliphatic CH groups ($-\text{CH}_3$, $-\text{CH}_2-$, $>\text{CH}-$), halogens (fluorine and chlorine), and $-\text{OH}$ groups. Five-membered heteroatomic rings present a negative factor, whereas six-membered heteroatomic rings present a positive factor. The specific information described can be used as an aid to the drug design process.

INTRODUCTION

The binding of a drug to proteins in plasma has a strong influence on its pharmacodynamic behavior. Oral bioavailability is directly affected by the extent to which a drug binds to plasma proteins because the bound drug is not available to the mechanisms that govern first pass metabolism. The free concentration of the drug, hence its bioavailability, is at stake when a drug binds to serum proteins in this process. The reversible drug–protein complex circulates through the system and serves as a depot, making available an unbound drug when the elimination processes have depleted the concentration of a free drug. By this mechanism, a bound drug can replenish the free concentration of the drug in vivo. The consequences of this process may be prolonged activity which may be desirable or may lead to the emergence of undesirable side effects.

Estimation of this significant ADME property (absorption, distribution, metabolism, and excretion) is one of the important activities carried out in the early stages of drug design. Some understanding of the possible binding characteristics of candidate molecules is valuable information in the strategies of the design process. Recently available

information suggests that success in drug development occurs at a rate of less than 20%, and poor pharmacokinetic properties in the candidate compounds account for almost half of the reported failures.^{1,2} A goal of these present studies is the creation of models that are predictive of the extent of serum protein binding. Hence, a predictive model of serum protein binding is of value to optimize these events in the design of new drugs.

The specific area of our interest is an alternative experimental measure of albumin binding, obtained as a chromatographic retention index from a column of immobilized albumin. These measurements lead to a retention constant k that may be called the albumin binding affinity of the drug.

OBJECTIVES

The principle objective of this investigation is the development of a quantitative model that will establish a relation between structure and binding affinity as well as predict the extent to which drug and drug-like compounds exhibit reversible binding to human serum albumin (HSA). The study investigates QSAR model development for a diverse, heterogeneous group of commercially available drugs whose activity has been determined by a chromatographic (HPLC) experimental method. Our effort is directed toward the

* Corresponding author phone: (617)745-3549; e-mail: hall@enc.edu.

production of a model that will exploit fully the limits of the data available, that is develop the best model possible given the limitations of the experimental data on which the model is based.

This study makes use of the topological structure representation approach, the main focus of which is the development of QSAR models based on structure information that can be related to a property such as protein binding data in a reasonable manner. Rather than attempt to simulate the binding process or develop equations for interactions in an assumed mechanism, we model information that directly represents the structure in the data set of molecules. In this approach, the information resides in topological descriptors that represent molecular structure. Statistical methods yield a model that captures the parallel between variation in structure features and the corresponding variation in property values. These methods do not require 3D-based geometry information, which often requires time-consuming quantum mechanical calculations, nor is it necessary to make assumptions about the mechanism of the process.

The experiences with this topological approach indicate their usefulness in drug design and property estimation. Structure information from the model leads to direct statements about the role of structure features and is thus an aid to design. The model is readily developed by standard, widely used statistical methods and is implemented in this case with MDL QSAR software. The use of the model to compute protein binding for a virtual library is very fast, much faster than methods based on detailed 3D information. Of perhaps even greater significance is the fact that the nature of the structure information contained in a model created by the topological method can assist in the creation of such a library by indicating the significant structure features that should be included in the library. In this way, both of the necessary steps, library construction and library screening, are rapidly advanced by the development of topological QSAR models. These considerations are very important to the drug design process.

In this study of the investigation into modeling HSA binding [logk], the topological structure representation method is applied to a diverse set of drugs whose albumin binding affinity has been measured by high performance liquid chromatography. A set of 10 compounds was set aside as an external validation test set. In a recent study the topological structure representation method was applied to a set of 115 beta-lactams, a combination of penicillins and cephalosporins.³ The QSAR model obtained in this present study is of similar quality to the beta-lactam study and includes a similar but not identical set of structure descriptors.

DATA AND METHODS

The conventional method for the reporting of serum protein binding is given as the percentage of a drug bound to plasma proteins at clinically achievable concentrations. These values are usually in vitro measurements using several concentrations, resulting in a calculated mean value. A number of other methods have been explored which produce results not much different from those described above.^{4,5} An alternative method, designed for more rapid screening, involves the use of immobilized HSA as the stationary phase in an HPLC procedure. The experimental quantity has been

demonstrated to parallel the binding of drugs to free HSA.⁶⁻⁸ In this present study we have drawn upon albumin binding affinity values published by Colmenarejo.⁹

The data set consists of the binding affinity constant k for 94 diverse drugs. The experimental data was obtained by HPLC chromatographic methods. The affinity binding constant k was computed from retention time (t) as follows: $k = (t - t_0)/t_0$, where t_0 is the time for passage of a nonretained material. In this case, structures were drawn with ChemDraw Pro and saved as mol files.¹⁰ The compound name, along with logk(HSA) values and mol file names, representing a structure for each compound, were entered into an Excel spreadsheet. The resulting spreadsheet was then imported, along with the mol files, into MDL QSAR to create a database.¹¹ In all investigations of this type, extensive checking is required in order to ensure the accuracy of the structures and activity values that have been entered into the data sets that are used. The MDL QSAR system relies extensively on nonempirical topological and electrotopological parameters including electrotopological state indices (E-State),¹² molecular connectivity chi indices,^{13,14} and kappa shape indices.¹⁵

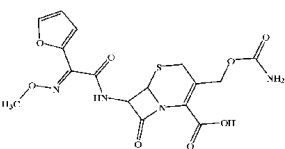
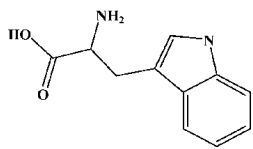
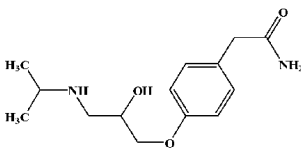
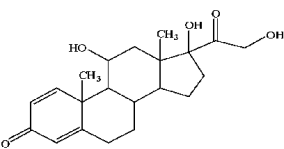
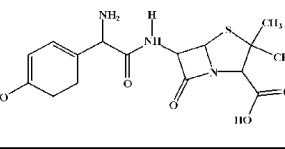
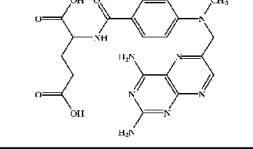
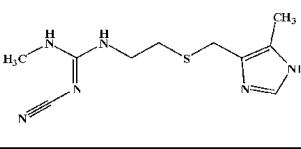
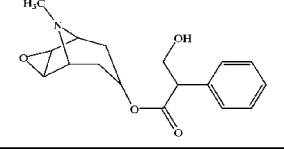
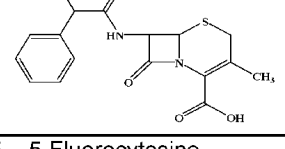
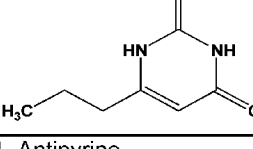
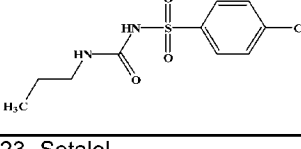
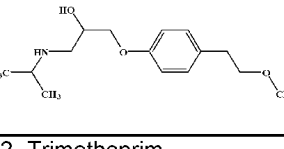
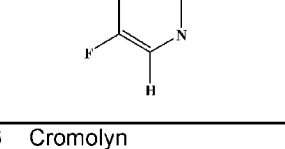
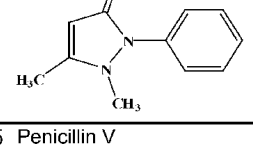
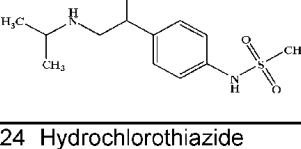
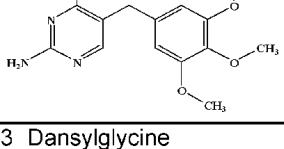
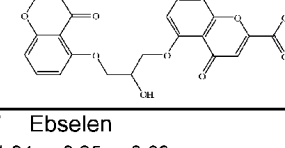
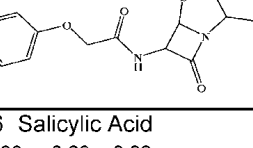
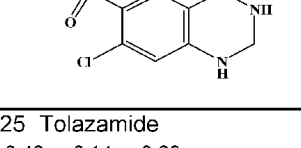
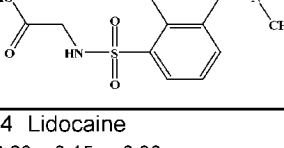
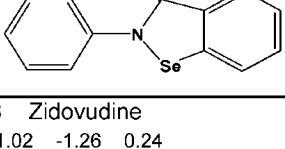
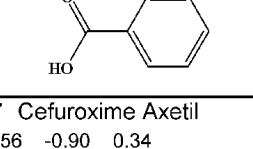
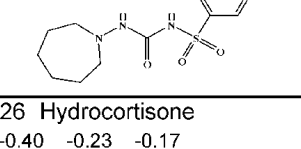
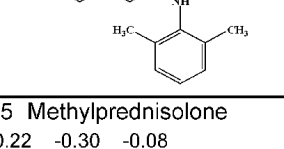
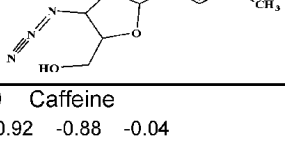
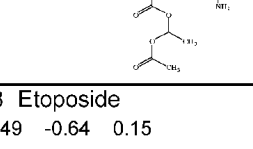
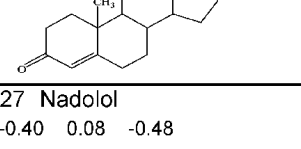
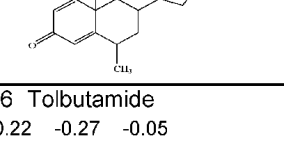
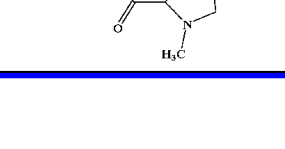
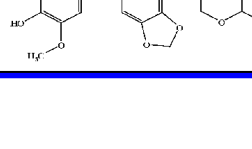
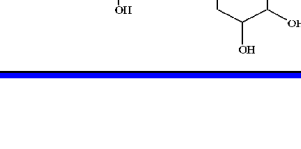
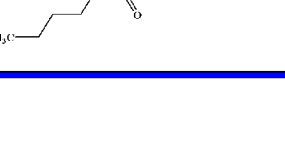
For this data set, the log k values range from -1.39 (acetylsalicylic acid) to $+1.34$ (clotrimazole). Structures for training set compounds in the study are given in Figures 1 and 2 for the test set, along with reported, calculated, and residual values of HSA column affinity [logk(HSA)].

Classification of Descriptors Considered in the Modeling Process. In order for property modeling to be useful, the model should lead to information about structure features that relate significantly to the property being modeled; the model should be statistically sound for prediction; and use of the model should be fast and economical. The choice of molecular descriptors is important to ensure that these ends are achieved. The model presented here makes use of a combination of E-State and molecular connectivity chi indices, which have been used to develop models for many activities and properties.

The E-state indices have been used in both their atom-level^{12,16-23} and atom-type forms.^{12,16,23-26} E-State QSAR models reveal structure features significantly related to properties. Further, development of hydrogen E-State values (and hydrogen atom-type E-State indices^{12,24,27}) has extended the capability of the E-State as a powerful set of structure descriptors. Several QSAR models of binding have been reported,^{12,16,24-27} indicating their ability to represent both hydrogen bonding groups as well as nonpolar regions of molecules.

Validation of E-State models is further supported by cross-validation experiments using the leave-group-out (LGO) method.²⁴⁻²⁶ The atom type E-State structure descriptors have also been shown to be very useful in searching a chemical database for structures similar to a desired target structure,^{28,29} indicating that an E-State QSAR model can assist in similarity search of a database, experimental or virtual.

Recent work has revealed the relation of the molecular connectivity chi indices to intermolecular accessibility.^{30,31} Molecular connectivity chi indices have also been reported in published models, indicating their ability to encode structure features that have a significant relationship to properties.^{32,33}

ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res
2 Cefuroxime -1.33 -0.90 -0.43 	11 L-tryptophan -0.78 -0.56 -0.22 	19 Atenolol -0.48 -0.22 -0.26 	28 Prednisolone -0.40 -0.40 0.00 
3 Amoxicillin -1.21 -1.08 -0.13 	12 Methotrexate -0.77 -0.35 -0.42 	21 Cimetidine -0.44 -0.59 0.15 	29 Scopolamine -0.34 -0.17 -0.17 
4 Cephalexin -1.11 -0.40 -0.71 	13 Propylthiouracil -0.75 -0.83 0.08 	22 Chlorpropamide -0.44 -0.42 -0.02 	31 Metoprolol -0.29 -0.10 -0.39 
5 5-Fluorocytosine -1.11 -0.79 -0.32 	14 Antipyrine -0.69 -0.37 -0.32 	23 Sotalol -0.44 -0.22 -0.22 	32 Trimethoprim -0.26 -0.22 -0.04 
6 Cromolyn -1.07 -0.45 -0.62 	15 Penicillin V -0.69 -0.71 0.02 	24 Hydrochlorothiazide -0.42 -0.76 0.34 	33 Dansylglycine -0.26 0.12 -0.14 
7 Ebselen -1.04 -0.35 -0.69 	16 Salicylic Acid -0.66 -0.69 0.03 	25 Tolazamide -0.42 -0.14 -0.28 	34 Lidocaine -0.23 0.15 -0.38 
8 Zidovudine -1.02 -1.26 0.24 	17 Cefuroxime Axetil -0.56 -0.90 0.34 	26 Hydrocortisone -0.40 -0.23 -0.17 	35 Methylprednisolone -0.22 -0.30 -0.08 
9 Caffeine -0.92 -0.88 -0.04 	18 Etoposide -0.49 -0.64 0.15 	27 Nadolol -0.40 0.08 -0.48 	36 Tolbutamide -0.22 -0.27 -0.05 

ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res
37 Sulfaphenazole -0.21 -0.13 -0.08 	46 Ranitidine -0.10 -0.30 0.20 	55 Acrivastine -0.02 0.20 -0.22 	64 Labetalol 0.14 0.24 -0.10
38 Acebutolol -0.21 -0.04 -0.17 	47 Carbamazepine -0.10 -0.10 0.00 	56 Phenytoin 0.00 -0.12 0.12 	65 Norfloxacin 0.14 0.12 0.02
39 Procaine -0.19 -0.15 -0.04 	48 Camptothecin -0.08 -0.18 0.10 	57 Doxycycline 0.01 -0.38 0.36 	66 Phenylbutazone 0.19 0.20 -0.01
41 Oxprenolol -0.15 -0.20 0.05 	49 Tetracycline -0.08 -0.24 0.16 	58 Ketoprofen 0.03 -0.01 0.04 	67 Sancycline 0.21 -0.24 0.45
42 Lamotrigine -0.13 -0.40 0.27 	51 Sumatriptan -0.05 0.19 0.14 	59 Alprenolol 0.04 -0.10 0.14 	68 Minocycline 0.21 -0.01 0.22
43 Clonidine -0.13 -0.47 0.34 	52 Warfarin -0.04 0.05 -0.09 	61 Digitoxin 0.13 0.49 -0.36 	69 Naproxen 0.25 -0.01 0.26
44 Pindolol -0.13 -0.15 0.02 	53 Bumetanide -0.03 -0.09 0.06 	62 Ofloxacin 0.14 0.21 -0.07 	71 Propranolol 0.28 0.26 0.02
45 Furosemide -0.13 -0.64 0.51 	54 Oxyphenbutazone -0.02 0.09 -0.11 	63 Ciprofloxacin 0.14 0.10 0.04 	72 Tetracaine 0.32 0.16 0.16

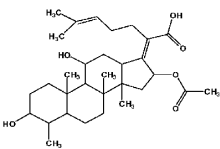
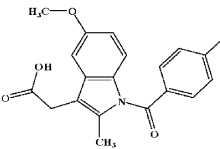
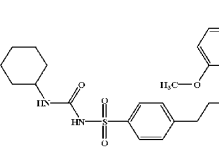
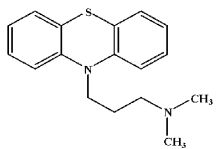
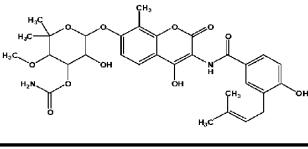
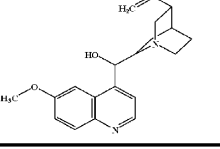
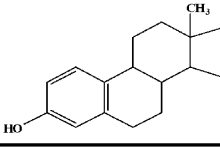
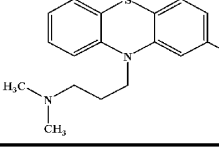
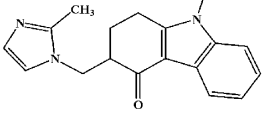
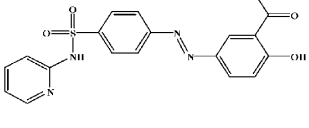
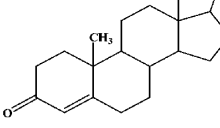
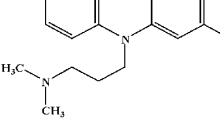
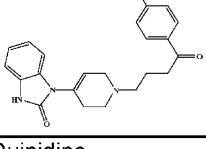
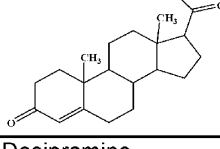
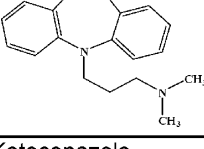
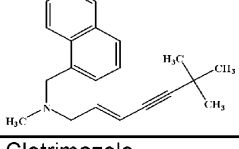
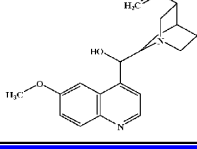
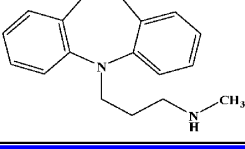
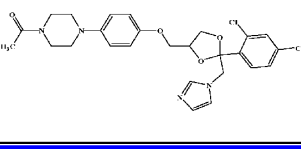
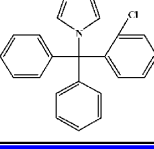
ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res	ID Name logkHSA / Calc / Res
73 Fusidic acid 0.33 0.72 -0.39 	78 Indomethacin 0.47 0.16 0.31 	84 Glibenclamide 0.68 0.58 0.10 	89 Promazine 0.92 0.77 0.15 
74 Novobiocin 0.35 0.13 0.22 	79 Quinine 0.49 0.57 -0.08 	85 Estradiol 0.68 0.36 0.32 	91 Triflupromazine 1.05 1.42 -0.37 
75 Ondansetron 0.37 0.18 0.19 	81 Sulfasalazine 0.56 -0.04 0.60 	86 Testosterone 0.74 0.20 0.54 	92 Clorpromazine 1.10 0.83 0.27 
76 Droperidol 0.43 0.47 -0.04 	82 Progesterone 0.59 0.30 0.29 	87 Imipramine 0.75 0.91 -0.16 	93 Terbinafine 1.17 0.71 0.46 
77 Quinidine 0.44 0.57 -0.13 	83 Desipramine 0.61 0.72 -0.11 	88 Ketoconazole 0.84 0.76 0.08 	94 Clotrimazole 1.34 1.05 0.29 

Figure 1. Structures for the diverse drugs included in the training set for the albumin binding affinity study. Structures are given along with observed logk(HSA) value. Calculated logk(HSA) value and residual are derived from eq 1. Compound numbering is the same as in the original ref 9, beginning with compound number 2. The rest of the data is in Figure 2 as the validation test set.

In the initial appraisal of the data, a qualitative analysis of the relation between albumin binding affinity and structure descriptors suggests that several diverse structure features appear to be related to binding activity in the data set. In the topological structure representation method employed in this investigation, many of the features in question are encoded as atom type E-State and hydrogen E-State descriptors. It is evident from this preliminary analysis that, from a statistical point of view, several of the structural characteristics that appear to be related to binding occur in too few instances for their atom type E-States to be used individually.

Two objects of concern are brought to light by this observation of the significance yet low population of certain descriptors. The first involves the need to include as many of the relevant structure information (atom-types) in the QSAR equation as possible, while at the same time keeping the number of variables within statistically acceptable limits. The second is to ensure the use of structure descriptors that are sufficiently populated so as to be included in a valid model. In answer to these two concerns, new descriptors were

created from chemically sensible combinations of some atom types whose parameter space was underpopulated. This course of action is justified by the assumption that the chemical interactions of similar structural features will likely be related to binding in a similar way.

Combination descriptors were created only when the sign of the regression coefficient for each member in the combination was identical, and the structural features could be reasonably assumed to interact in a similar way. These combination descriptors are created in a straightforward manner by the MDL QSAR expression-generator. Specifically, three combination descriptors were created in MDL QSAR: the sum of aromatic carbon atom types, $S^T(\text{arom})$; the sum of aliphatic carbons that bear hydrogens, $S^T(\text{CHsat})$; and the sum of the atom type E-State for fluorine and chlorine, $S^T(-\text{F}, \text{Cl})$. The combination descriptors appear in Figure 3 and are defined along with all others appearing in the final model.

The preliminary collection of descriptors in the database was screened, resulting in the removal of all descriptors that

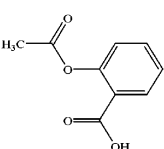
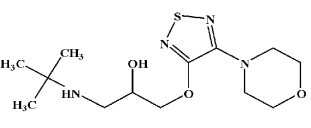
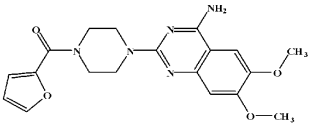
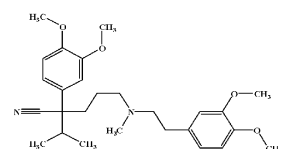
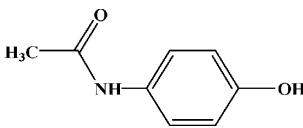
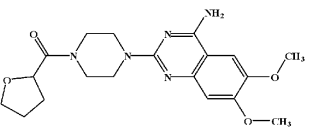
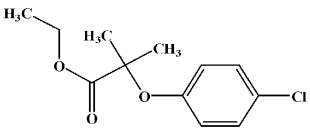
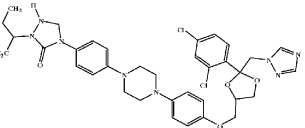
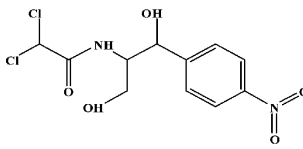
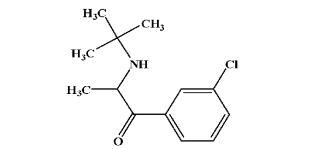
ID Name logkHSA / Pred / Res	ID Name logkHSA / Pred / Res	ID Name logkHSA / Pred / Res	ID Name logkHSA / Pred / Res
1 Acetylsalicylic acid -1.39 -0.64 -0.75 	30 Timolol -0.33 -0.38 0.05 	60 Prazosin 0.06 -0.06 0.12 	80 Verapamil 0.52 1.16 -0.64 
10 Acetaminophen -0.81 -0.57 -0.24 	40 Terazosin -0.16 -0.26 0.10 	70 Clofibrate 0.27 -0.12 0.35 	90 Itraconazole 1.04 1.50 -0.46 
20 Chloramphenicol -0.46 -0.70 0.24 	50 Bupropion -0.05 0.08 -0.13 		

Figure 2. Structure for the external validation test set of diverse drugs used to evaluate the model given in eq 1. Structures are given along with observed logk(HSA) value. Predicted logk(HSA) value and residual are derived from eq 1. Compound numbering is the same as in the original ref 9, beginning with compound number 1.

Symbol	Definition	Description
$S^T(\text{arom})$	$S^T(=\text{CH}=\text{}) + S^T(=\text{C}=\text{C}=\text{}) + S^T(=\text{C}=\text{C}=\text{C}=\text{})$	atom type E-State for aromatic carbon atoms
$S^T(\text{CHsat})$	$S^T(-\text{CH}_3) + S^T(-\text{CH}_2-) + S^T(>\text{CH}-)$	atom type E-State for saturated carbon atoms (with hydrogens)
$S^T(-\text{F}, -\text{Cl})$	$S^T(-\text{F}) + S^T(-\text{Cl})$	atom type E-State for halogens fluorine and chlorine
$S^T(-\text{OH})$		atom type E-State for -OH
${}^6\chi_{\text{CH}}$		sixth order valence molecular connectivity chi chain index
${}^5\chi_{\text{CH}}$		fifth order valence molecular connectivity chi chain index

Figure 3. Structure descriptor symbols that appear in the model discussed in this investigation given along with their definitions.

had the same value (including zero) for more than 90% of the compounds in the training set. Intercorrelated pairs of descriptors were left in the descriptor pool for consideration, but generally no two descriptors with pairwise correlation of greater than 0.80 were used in a final model (0.46 for the present model). Because of the nature of the topological structure representation method, a value for every descriptor does exist for each compound. A value of zero for an atom type E-State index indicates the absence of that atom type from the molecule. Such presence/absence of information is valuable in the design process as an aid in selection of groups for synthesis of new candidate structures.

The final model was developed using the all-possible-subsets (APS) regression feature of MDL QSAR. Models for all possible combinations of descriptors are computed and rank-ordered on r^2 , starting from one variable at a time, up to some prescribed limit. The APS procedure is not a stepwise method; every combination of descriptors is computed in a model and evaluated statistically. The final model

was selected on the basis of direct statistics, leave-one-out and leave-group-out cross-validation, and then tested on a ten-compound external validation test set.⁹

RESULTS

We modeled the training set of 84 compounds (Figure 1) using the APS regression feature of MDL QSAR and then predicted the 10 compounds comprising the external validation test set (Figure 2). A six variable model was found with the following statistics:

$$\begin{aligned} \log k(\text{HSA}) = & 0.0503 (\pm 0.0050) * S^T(\text{arom}) + \\ & 0.0787 (\pm 0.0083) * S^T(\text{CHsat}) + \\ & 0.0291 (\pm 0.0059) * S^T(-\text{F}, \text{Cl}) + \\ & 0.00871 (\pm 0.0032) * S^T(-\text{OH}) + \\ & 2.38 (\pm 0.95) * {}^6\chi_{\text{CH}} - \\ & 2.96 (\pm 0.99) * {}^5\chi_{\text{CH}} - 1.18 \quad (1) \end{aligned}$$

$$r^2 = 0.77, s = 0.29, F = 43, n = 84, q^2 = 0.70,$$

$$s_{\text{press}} = 0.33$$

The definitions of the structure descriptors in eq 1 are given in Figure 3. Quantities in parentheses are the standard deviation of the coefficients. The statistical quantities q^2 and s_{press} are based on the leave-one-out (LOO) method. Only three residuals exceed two-sigma, and no compound may be classified as an outlier (residual exceeding three-sigma). The descriptors in the model are essentially uncorrelated. The largest intercorrelation exists for $S^T(-\text{OH})$ and ${}^6\chi_{\text{CH}}$; $r^2 = 0.21$. The second largest intercorrelation exists for $S^T(-\text{OH})$ and $S^T(\text{arom})$; $r^2 = 0.14$. A plot of observed and calculated logk(HSA) for the albumin binding affinity training set is given in Figure 4.

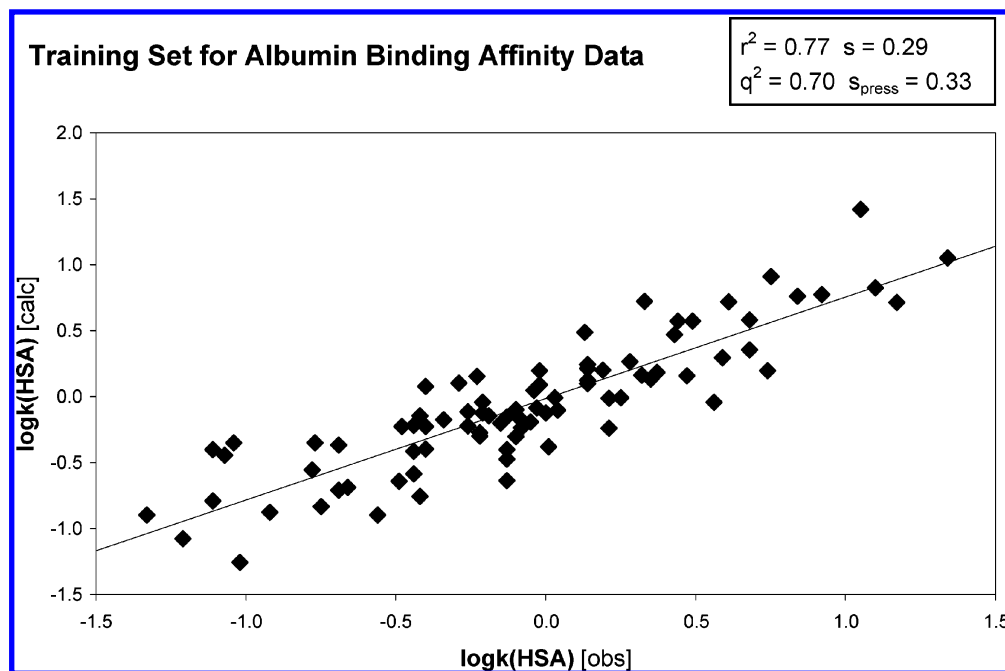


Figure 4. Plot of calculated $\log k(\text{HSA})$ versus experimental values for the albumin binding affinity training set, according to the topological model, eq 1.

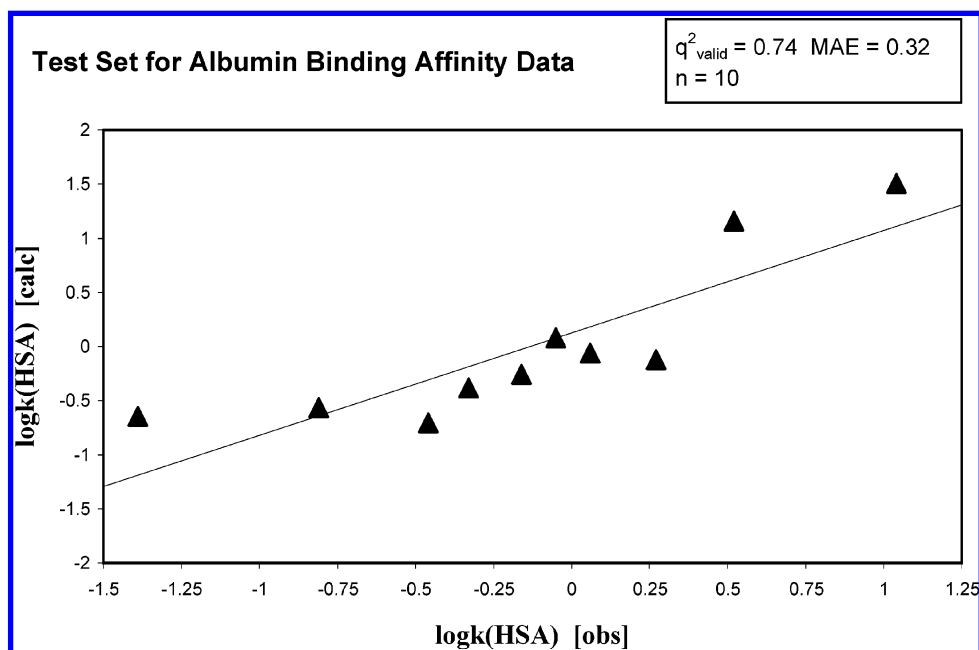


Figure 5. Plot of predicted $\log k(\text{HSA})$ versus experimental values for the albumin binding affinity external validation test set, according to the topological model, eq 1.

Three tests of the model were performed. First, the $\log k(\text{HSA})$ values were randomized, and the statistics were recomputed by MDL QSAR. This process was carried out 100 times.^{35,36} The r^2 values obtained in this randomization process ranged from 0.014 to 0.23, with an average of 0.011. Based on this information, we conclude that the model obtained (eq 1) is significantly different from that obtained with random numbers; that is, the model encodes significant information rather than random numbers. The second test process was based on the leave-group-out (LGO) (and predict) approach.^{24–26,35,36} In this case, a 10% group of the data was deleted; a new model was obtained for the remaining 90%; and then the deleted compounds were predicted. This process was repeated until all 84 compounds

had been left out once. Then that whole process was repeated 10 times. For all of the residuals obtained in this manner (84×10), the $q^2 = r^2_{\text{press}} = 0.68$.

The third test of this model is the use of an external validation test set, the same test set as used by Colmenarejo.⁹ For the 10 compounds left out and used as a test set, the correlation coefficient $r^2 = 0.74$ and the mean absolute error of prediction is $\text{MAE} = 0.31$ ($\text{rms} = 0.32$). These results strongly support the predictive potential of the model (eq 1). A plot of observed and calculated $\log k$ values for the albumin binding affinity external validation test set is given in Figure 5. The definition of the variables is given in Figure 3, and the structure interpretation of each variable is given in the Discussion. It should be emphasized that no residual

in either the training or test sets exceeds three standard deviations.

DISCUSSION

The significant structure information resident in the model will be presented through discussion of each structure descriptor in the model. Further, a brief comparison is made to an earlier model for these data published by Colmenarejo.⁹

Interpretation of the Model. $S^T(\text{arom})$. The $S^T(\text{arom})$ descriptor encodes the collective E-State values for aromatic carbons and CH fragments in a molecule, including phenyl and heteroaromatic rings. This descriptor has been found useful in earlier studies in which contributions from the three aromatic carbon atom types were all important,¹⁷ including the recent study of beta-lactam human serum protein binding.³ The magnitude of this index depends on the presence and nature of the substituents; it is not a mere count of atoms but varies with the bonding environment of each atom. For example, electronegative groups decrease the value of $S^T(\text{arom})$. Substituted aromatic carbon atoms also tend to have smaller E-State values than the unsubstituted carbon atoms. In this model, for the 72 compounds with aromatic rings, $S^T(\text{arom})$ accounts for 45.2% of the calculated $\log k(\text{HSA})$ value and ranges from 6.4 to 87.8%. $S^T(\text{arom})$ makes a large contribution for carbamazepine (87.8%), sulfenazole (80.8%), and clotrimazole (78.6%). Because of the positive coefficient in the model, increasing the electron accessibility and/or the count of aromatic rings increases the calculated $\log k(\text{HSA})$ value. The same descriptor is found in the QSAR model for beta-lactam protein binding.³

$S^T(-\text{F}, \text{Cl})$. Sixteen of the molecules contain fluorine and/or chlorine. This descriptor accounts for 29.0% of the calculated $\log k(\text{HSA})$ value for these 16 compounds and ranges from 7.9 to 90.2%. For three compounds the contribution is large: 90.2% for 5-fluorocytosine, 44.2% for lamotrigine, and 43.9% for triflupromazine. When present, fluorine or chlorine makes a significant contribution to calculated $\log k$. The positive coefficient indicates that increasing the descriptor also increases the calculated $\log k$ values. This same descriptor was found to be important in the beta-lactam QSAR model.³

$S^T(\text{CHsat})$. This structure descriptor is the sum of the atom type E-State values for the aliphatic CH_n groups: $-\text{CH}_3$, $-\text{CH}_2-$, $>\text{CH}-$. The descriptor is nonzero for 75 compounds in the data set. For those 75 compounds, on average this index contributes 31.9% to the calculated $\log k$ value but ranges from 0.3% (hydrochlorothiazole, -0.42) to 72.3% (progesterone, 0.56) to 87.8% (propylthiouracil, -0.75). The descriptor can have negative values, especially when electronegative groups, such as $-\text{OH}$, $-\text{F}$ and $-\text{Cl}$, are bonded to a CH_n group. These negative values make significant percent contributions for a few compounds, such as cromolyn (13.9%), chloramphenicol (10.6%), and doxycycline (10.6%). Four of the five compounds with negative values for this descriptor also have negative $\log k$ values: $\log k < -0.35$. A similar structure descriptor, for methylene groups only, was found in the beta-lactam human serum protein binding model.³

$S^T(-\text{OH})$. This descriptor encodes the electron accessibility and number of $-\text{OH}$ groups in the molecule. For the 49 compounds with $-\text{OH}$ groups, $S^T(-\text{OH})$ contributes on

average 14.2% but ranges from 3.8 to 50.6%. The compound with the largest value for this descriptor is tetracycline (#49, 50.6%). Because of its positive coefficient, increasing this descriptor also increases calculated $\log k$ values.

$^6\chi^v_{\text{CH}}$. This molecular connectivity chi index, the sixth-order valence chain, encodes information for a six-membered ring, encoding the presence of heteroatoms. For the 82 compounds with nonzero values, the $^6\chi^v_{\text{CH}}$ index contributes 12.8% on average to calculated $\log k$ and ranges from 4.3 to 41.6%. A wide variety of six-membered rings are encountered in this data set including phenyl, pyridine, various di- and triazo rings, the sulfur-containing ring in cephalosporins, saturated and unsaturated rings, pyrans, etc. Its positive coefficient indicates that increasing value increases the calculated $\log k$ value.

$^5\chi^v_{\text{CH}}$. This chi index, the fifth-order chi valence chain, encodes information for a five-membered ring, encoding the presence of heteroatoms. Five-membered rings encountered in this data set include furan, indole, pyrrole, and various systems with more than one heteroatom. It is the only index with a negative coefficient in this model. This negative effect on the $\log k(\text{HSA})$ values is also reflected in the fact that 47% of the negative $\log k(\text{HSA})$ values have a nonzero value for this index, but only 36% of the positive $\log k(\text{HSA})$ values do. For the 36 compounds with nonzero values for this index, its contribution to the calculated value is 14.5% on average and ranges from 4.5 to 50.6%. The compound with the largest contribution is #8 zidovudine (-1.02); the second largest is #3, amoxicillin (41.6%). The negative coefficient indicates that five-membered rings make a negative contribution to calculated $\log k$; highly substituted rings make a smaller contribution than less substituted rings.

In their statistical study of this chromatographic data set, Colmenarejo et al. found a model with statistics of lower quality than eq 1.⁹ More importantly in their study, five compounds were found to be outliers: #55, doxycycline; #61, digitoxin; #67, sancicline; #73, fusidic acid; and #74, novobiocin, necessitating their removal from the training set by Colmenarejo. As can be seen in Figure 4, none of these five compounds is found to be an outlier in our study, and all 84 compounds were retained in this present study. The quality of our external validation test for eq 1, $\text{MAE} = 0.32$, indicates that our model is useful for prediction. Only acetylsalicylic acid has a somewhat large residual, -0.75 . Colmenarejo et al. found a larger residual for aspirin, approximately -1.0 . Finally, in this present study, an analysis of the structure features in the model is given so that the model may be used in the design of new compounds or a virtual library for screening. No similar analysis was given by Colmenarejo.

CONCLUSIONS

The model for this data set yields reasonable estimates of binding affinity, $\log k$, commensurate with the experimental quality of the data. It is important to recognize that for our model, no residual exceeds three sigma, suggesting a sound basis for estimation of the $\log k$ values for new candidate compounds. The equations may be used for quantitative estimates with the expectation that predictions are useful for the usual process of drug design. For the protein binding affinity, the model indicates that binding affinity ($\log k$) is

increased by the presence and electron accessibility of aromatic groups as well as aliphatic CH groups and the presence and electron accessibility of -F and -Cl atoms and -OH groups but is decreased by the presence of five-membered heteroatomic rings. The actual model, eq 1, may be used with confidence to make quantitative estimates of logk(HSA) for drugs.

Because this model involves topological descriptors of molecular structure, creating the model is straightforward and direct with the use of the MDL QSAR software. Further, the model may be used to predict albumin binding affinity at high speed for virtual libraries of structures. A significant conclusion from this study is that data that are typically available during early phases of the drug discovery process may be used to predict the properties of commercial drugs through the use of the topological method for QSAR modeling.

REFERENCES AND NOTES

- (1) Prentis, R. A.; Lis Y.; Walker, S. R. Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964–1985) *Br. J. Clin. Pharmacol.* **1988**, *25*, 387–396.
- (2) Kennedy, T. Managing the Drug Discovery/development Interface. *Drug Discovery Today* **1997**, *2*(10), 436–444.
- (3) Hall, L. M.; Hall, L. H.; Kier, L. B. QSAR Modeling of Beta-Lactam Binding to Serum Proteins. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 103–118.
- (4) Bird, A. E.; Marshall, A. C. Correlation of Serum Binding of Penicillins with Partition Coefficients. *Biochem. Pharmacol.* **1967**, *16*, 2275–2290.
- (5) Pacifici, G. M.; Viani, A. Methods of Determining Plasma and Tissue Binding of Drugs. *Clin. Pharmacokinet.* **1992**, *23*, 449–459.
- (6) Frostell-Karlsson, A.; Ramaeus, A.; Roos, H.; Andersson, K.; Borg, P.; Hamalainen, M.; Karlsson, R. Biosensor Analysis of the Interaction between Immobilized Human Serum Albumin Binding Levels. *J. Med. Chem.* **2000**, *43*, 1986–1992.
- (7) Domenici, E.; Bortucci, C.; Salvadori, P.; Motellier, S.; Wainer, I. W. Immobilized Serum Albumin: Rapid HPLC Probe of Stereoselective Protein Binding Interactions. *Chirality* **1990**, *2*, 263–268.
- (8) Domenici, E.; Bortucci, C.; Salvadori, P.; Wainer, I. W. Use of a Human Serum Albumin-Based Chiral Stationary Phase for High-Performance Liquid Chromatography for the Investigation of Protein Binding: Detection of the Allosteric Interaction Between Warfarin and Benzodiazepine Binding Sites. *J. Pharm. Sci.* **1991**, *80*, 164–169.
- (9) Colmenarejo, G.; Alvarez-Pedraglio, A.; Lavandera, J.-L. Cheminformatic Models to Predict Binding Affinities to Human Serum Albumin. *J. Med. Chem.* **2001**, *44*, 4370–4378.
- (10) CS ChemDraw PRO, ver 4.5, CambridgeSoft, 875 Massachusetts Avenue, Cambridge, MA 02139.
- (11) MDL QSAR, MDL Information Systems, 14600 Catalina Street, San Leandro, CA 94577.
- (12) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, 1999.
- (13) (a) Kier, L. B.; Hall, L. H. *Molecular Connectivity In Chemistry and Drug Research*; Academic Press: New York, 1976. (b) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; John Wiley Publications: London, 1986.
- (14) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure–Property Relations. In *Reviews of Computational Chemistry*; Boyd, D., Lipkowitz, K., Eds.; VCH Publishers: 1991; Chapter 9, pp 367–422.
- (15) Kier, L. B.; Hall, L. H. The Kappa Indices for Modeling Molecular Shape and Flexibility. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 455–490.
- (16) Kier, L. B.; Hall, L. H. Inhibition of Salicylamide Binding: An Electrotopological State Analysis. *Med. Chem. Res.* **1992**, *2*, 497–502.
- (17) Gough, J. D.; Hall, L. H. QSAR Models of the Antileukemic Potency of Carboquinones: Electrotopological State and Chi Indices. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 356–361.
- (18) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.
- (19) Huuskonen, J. QSAR Modeling with the Electrotopological State: TIBO Derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425–429.
- (20) Pantakar, S. J.; Jurs, P. C. Prediction of IC₅₀ Values for ACAT Inhibitors from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 706–723.
- (21) Gozalbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. Molecular Search of New Active Drugs Against *Toxoplasma Gondii*. *SAR QSAR Environ. Res.* **1999**, *10*, 47–60.
- (22) Hall, L. H.; Mohny, B. K.; Kier, L. B. Comparison of Electrotopological State Indexes with Molecular Orbital Parameters: Inhibition of MAO by Hydrazides. *Quant. Struct.-Act. Relat.* **1993**, *12*, 44–48.
- (23) Hall, L. H.; 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*, 1039–1045.
- (24) (a) Maw, H. H.; Hall, L. H. E-State Modeling of Dopamine Transporter Binding. Validation of Model for a Small Data Set. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1270–1275. (b) Maw, H. H.; Hall, L. H. E-State Modeling of Corticosteroid Binding Affinity. Validation of Model for a Small Data Set. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1248–1254.
- (25) Maw, H. H.; Hall, L. H. E-State Modeling of HIV-1 Protease Inhibitor Binding Independent of 3D Information. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 290–298.
- (26) (a) Rose, K.; Hall, L. H.; Kier, L. B. Modeling Blood-Brain Barrier Penetration Using the Electrotopological State. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651–666. (b) Rose, K.; Hall, L. H. E-State Modeling of Fish Toxicity Independent of 3D Structure Information. *SAR QSAR Environ. Res.* **2003**, *14*, 113–129.
- (27) Hall, L. H.; Kier, L. B. The Electrotopological State: Structure Modeling for QSAR and Database Analysis. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 491–562.
- (28) (a) Kier, L. B.; Hall, L. H. Database Organization and Similarity Searching with E-State Indices. In *Symposium on Computer Methods for Structure Representation*; Kluwer Academic Publishing Co.: Amsterdam, The Netherlands, 2001; pp 33–49. (b) Hall, L. H.; Kier, L. B. The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784–791.
- (29) (a) Kier, L. B.; Hall, L. H. Database Organization and Searching with E-State Indices. *SAR QSAR Environ. Sci.* **2001**, *12*, 55–74. (b) Hall, L. H.; Kier, L. B. Molecular Similarity Based on Novel Atom Type Electrotopological State Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074–1080.
- (30) Kier, L. B.; Hall, L. H. Intermolecular Accessibility: The Meaning of Molecular Connectivity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792–795.
- (31) Kier, L. B.; Hall, L. H. Molecular Connectivity: Intermolecular Accessibility and Encounter Simulation. *J. Mol. Graph. Model.* **2001**, *20*, 76–83.
- (32) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Research Studies Press: John Wiley and Sons: Chichester, U.K., 1986.
- (33) Hall, L. H.; Kier, L. B. Molecular connectivity chi Indices for database analysis and structure–property modeling. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 307–360.
- (34) Hall, L. H.; Kier, L. B. Molecular Connectivity and Substructure Analysis. *J. Pharm. Sci.* **1978**, *67*, 1743–1747.
- (35) Hall, L. H.; Kier, L. B. A Molecular Connectivity Study of the Muscarinic Receptor Affinity of Acetylcholine Antagonists. *J. Pharm. Sci.* **1978**, *67*, 1408–1412.
- (36) Kier, L. B.; Hall, L. H. Structure–Activity Studies on Hallucinogenic Amphetamines Using Molecular Connectivity. *J. Med. Chem.* **1977**, *20*, 1631–1636.

CI030019W