Modeling Blood-Brain Barrier Partitioning Using the Electrotopological State

Kimberly Rose and Lowell H. Hall*

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

Lemont B. Kier

Department of Medicinal Chemistry, School of Pharmacy and The Center for The Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia 23298

Received December 3, 2001

The challenging problem of modeling blood-brain barrier partitioning is approached through topological representation of molecular structure. A QSAR model is developed for in vivo blood-brain partitioning data treated as the logarithm of the blood-brain concentration ratio. The model consists of three structure descriptors: the hydrogen E-State index for hydrogen bond donors, HS^T(HBd); the hydrogen E-State index for aromatic CHs, HS^T(arom); and the second order difference valence molecular connectivity index, $d^2\chi^v$ ($q^2 = 0.62$.) The model for the set of 106 compounds is validated through use of an external validation test set (20 compounds of the 106, MAE = 0.33, rms = 0.38), 5-fold cross-validation (MAE = 0.38, rms = 0.47), prediction of +/− values for an external test set (27/28 correct), and estimation of logBB values for a large data set of 20 039 drugs and drug-like compounds. Because no 3D structure information is used, computation of logBB by the model is very fast. The quality of the validation statistics supports the claim that the model may be used for estimation of logBB values for drug and drug-like molecules. Detailed structure interpretation is given for the structure indices in the model. The model indicates that molecules that penetrate the blood-brain barrier have large HS^T(arom) values (presence of aromatic groups) but small values of HS^T(HBd) (fewer or weaker H−Bond donors) and smaller $d^2\chi^v$ values (less branched molecules with fewer electronegative atoms). These three structure descriptors encode influence of molecular context of groups as well as counts of those groups.

INTRODUCTION

An important aspect of drug design is the consideration of the potential for penetration of the blood-brain barrier by any new candidate drug molecule. Whether the design requires penetration or demands minimum penetration, the ability to estimate the blood-brain ratio is an essential part of the design process. Modeling blood-brain partitioning is a challenging problem both because of the paucity of data and the task of establishing a useful relation between molecular structure and measured blood-brain partitioning. Two aspects of the ability to predict are addressed in this paper. First, knowledge of the structure basis underlying a model can assist the medicinal chemist in the molecular design process. Quantitative knowledge of the effect of structure components assists chemists in design decisions. Second, speed of computing properties is a very important consideration as chemists attempt to assess properties such as blood-brain barrier penetration for virtual libraries of thousands or millions of structures. The topological method is found to be very fast.

Blood-brain barrier penetration is associated with a particular structure in the brain. The arrangement of endothelial cells in the brain's internal circulation and adjacent capillaries manage the concentration of substances in the brain. The blood-brain barrier is a complicated system formed of tight junctions between brain parenchymal capillary cells in the choroid plexus and external tissue capillary cells. This barrier prevents many types of molecules from entering the CNS and cerebrospinal fluid.

An important factor in design of a therapeutic compound is its potential for penetrating the blood-brain barrier. Water along with small polar molecules and lipophilic drugs are known to cross the barrier. Larger polar molecules and hydrophyllic organic molecules, including plasma proteins, do not penetrate well. The partitioning of a compound across the blood-brain barrier is measured experimentally as the ratio of the concentration of the compound in the brain, [brain], to that in the blood, [blood]:

$$BB = [\text{brain}]/[\text{blood}]$$

BB is thought to be related to local hydrophobicity, molecular size, lipophilicity, and molecular flexibility,[1] but no explicit mathematical relationship has been given. Attempts to delineate the relationship have resulted in models based on various 3D methods. In this paper, we will explore an alternative approach based on topological representation of molecular structure that lends itself to very rapid property estimation.

In recent years, advances in modeling the relationship between molecular structure and various drug properties have taken two distinct routes. In one approach, significant information from 3D molecular geometries is required. Detailed atom-by-atom interactions are used for drug design

* Corresponding author phone: (617)745-3550; fax: (617)745-3509; e-mail: Hall@enc.edu.

10.1021/ci010127n CCC: $22.00
Published on Web 04/17/2002

and property estimation. These methods require individual atom coordinates for each molecule based on computed minimum-energy conformations for models of solvent interaction and, for drug design, data for an active site (from crystal structure data) and individual docking of each inhibitor into the active site. As a result, these methods are time-consuming and expensive but have demonstrated potential for useful drug design.[2,3]

In the other line of approach, topological representation of molecular structure is used. Statistical methods are employed to obtain a model that captures the parallel between variation in structure features and the corresponding variation in measured property values. These methods do not require 3D-based geometry information on the active site nor on the drug molecules. In the topological approach, two general methods are used: (1) topological superposition of common-core skeleton atoms in a series, in an attempt to identify key atoms in the interactions and (2) whole molecule or atom type indices, in an attempt to identify relevant whole molecule attributes of structure. These computational methods are very fast, significantly faster than those methods that require 3D geometry information. As a result, these methods are considerably less time-consuming and less expensive. In addition to speed of computation, models produced in this topological approach also yield significant specific structure information for the design of new compounds. The electrotopological state indices (E-State descriptors) have been used to model binding data, providing excellent statistics.[3−10] The topological structure representation approach has also been successfully applied to heterogeneous data sets with diverse molecular structures for biological as well as physicochemical data.[7,9,10] The totality of experiences with these topological approaches clearly indicates their usefulness in drug design and property estimation.

In this investigation, we developed a model for blood-brain barrier partitioning and demonstrated its validity by use of validation tests. The data set consists of 106 observations on neutral compounds, including a wide range of molecule size and complexity, ranging from small nonpolar molecules such as CH_4 and N_2 to small organics such as theophylline and caffeine to the larger drugs indinavir and verapamil.[11] The logBB values range over three and one-half orders of magnitude, from −2.15 to +1.44; molecular weights range from 16.0 to 613.8.

In this present paper, topological structure descriptors serve as the basis for structure representation in the model. This paper shows that the topological method applied to blood-brain barrier partitioning leads to a model that is statistically equivalent to or better than models reported in the literature.[11] Further, the method based on topological structure representation is considerably less time-consuming and, hence, less costly to use for prediction. The topologically based method leads to computation of logBB in the range of 5000−6000 molecules per minute as compared to 6 s for one molecule (10 per minute) as stated for one 3D-based method.[12]

Topological Descriptors. An important objective of property modeling is obtaining useful information about the structure features that relate significantly to the property being modeled, in addition to speed and cost considerations. For this present case, we use the molecular structure descriptors known as electrotopological state indices (E-State)[3−10,14−19] along with molecular connectivity chi indices.[20−22] The E-State indices have been used to develop models for many activities and properties in both their atom-level[3−5] and atom-type forms.[3,15−19,24,25] E-State QSAR models yield structure information that reveal structure features significantly related to activity. Further, the more recent development of hydrogen E-State values (and hydrogen atom-type E-State indices[3,19,24]) has extended the capability of the E-State as a powerful set of structure descriptors. Several QSAR models of binding have been reported,[3,5,19,24,25] indicating their ability to represent both hydrogen bonding groups as well as nonpolar regions of molecules. E-State structure descriptors have also been used to establish models based on large databases (thousands of compounds) for properties of particular interest to drug design, including logP and logW (aqueous solubility).[13] Validation of E-State models is further supported by cross-validation experiments.[24,25] 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,[17,19,26] indicating that an E-State QSAR model can assist the similarity search of a database, experimental or virtual.

Atom-Level E-State Values. Many QSAR and related studies have described the E-State formalism.[3−8,14−19,24−26] A brief summary of the method is presented here. In this topological approach to structure representation, structure information is developed for each atom (such as >N-, =O, −Cl) and each hydride group (such as $-CH_3$, $-NH_2$, −OH) in the molecule. For reasons of simplicity, both atoms and hydride groups are often called “atoms”. The E-State index for an atom in a molecule, S(i), is composed of an intrinsic state, I_i, plus the sum of perturbations, ΔI_{ij}, from all other atoms. The E-State value for atom i in a molecule is computed as follows:

$$S(i) = I_i + \Sigma_j \Delta I_{ij} \quad \text{sum over all other atoms j} \qquad (1)$$

The intrinsic state value for atom i, I_i, is derived from the ratio of the valence state electronegativity (given by the Kier-Hall electronegativity[3]) to a measure of the local topology (given as the number of skeletal neighbors). The general expression for atom i in row N of the periodic table (principal quantum number of valence electrons, N) is given as

$$I_i = ((2/N_i)^2\, \delta_i^{\,v} + 1)/\delta_i \qquad (2)$$

The perturbation term is as follows:

$$\Delta I_{ij} = (I_i - I_j)/r_{ij}^{\,2} \qquad (3)$$

in which r_{ij} is the topological distance between atoms, given as the number of atoms in the shortest path between atoms i and j. In the manner given by eq 1, the E-State value of each atom contains electronic and topological structure information from all other atoms within the structure.[3] *The E-State value S(i) is computed for each atom in the molecule (eq 1) and is called the atom-level E-State value.*

The information encoded in the E-State value for an atom is the electron accessibility at that atom. In this sense, the E-State values encode the potential for noncovalent intermolecular interaction[3,19] such as drug-receptor encounters, partitioning, vaporization, and solubility.[3] The atoms closest to a given atom have the greatest influence on its E-State

[S(i)] value. Influence diminishes for atoms separated by a path of several bonds; the influence decreases as the square of the number of atoms in the path. A parallel development provides the basis for *atom-level hydrogen E-State* indices, Hs(i).[3,19,24,25]

Atom-Type E-State Values. In addition to an atom-level E-State value computed for each atom, an atom-type formalism has been developed. For all data sets, including those with a common molecular scaffold as well as those with very diverse structures such as the present BB data set, the atom-type E-State indices provide much useful information. In the atom-type E-State formalism, each atom (or hydride group) in the molecule is classified into an atom type, such as OH, =O, or aromatic CH. *The atom type E-State index is defined as the sum of the individual atom level E-State values for a particular atom type.*[3,15,19] The atom-type descriptors combine three important aspects of structure information: (1) electron accessibility for the atoms of the same type, (2) presence/absence of the atom type, and (3) count of the atoms in the atom type.

Hydrogen atom-type E-State descriptors form a parallel set except that accessibility refers to hydrogen atom accessibility. *The hydrogen atom type E-State index is defined as the sum of the individual atom level hydrogen E-State values for all atoms of a particular atom type.*[3,24,25]

In the present blood-brain barrier partition training data set, the structures are very diverse and possess no common atoms for all molecules in the training set. Atom-type E-State and hydrogen E-State indices were used in model development in addition to molecular connectivity chi indices and kappa shape indices.

Model Validation Methods. An important aspect of QSAR modeling is the development of means for validation of the model. Good statistical criteria for fit to the training set are not a guarantee that the model can be used for accurate predictions of compounds outside the data set. In recent years, the leave-one-out (LOO) press statistics (q^2, s_{press}) have been used as a means of indicating predictive ability. A more exhaustive cross-validation method can be used in which a fraction of the data (e.g. 10−20%) is left out and predicted from a model based on the remaining data. This process (Leave-Group-Out, LGO) is repeated until each observation has been left out at least once.[25,26] Alternatively, one may set aside a selected part of the data (validation or test set) that is not used in any way to develop the model. Mean absolute error (MAE) for the prediction test set will be used as the significant criterion for assessing model quality. For this present investigation, all of these types of model validation have been used.

METHODS

Data Entry. Blood-brain barrier partitioning data has been measured experimentally by several investigators. For this present investigation, we limited the data to in vivo measurements obtained from rats.[11] In these protocols, the compound was administered via an iv mode; the rat was subsequently sacrificed. The concentration of the compound was measured in both the brain tissue and in the blood. BB is calculated from these analytical measurements.

Data and molecular structures were taken from 11 sources as follows, with the number of observations available from each source in parentheses: 1. Young (30);[11a] 2. Abraham (29);[11b] 3. Salminem (21);[11c] 4. Clark (7);[11d] 5. Luco (4);[11e] 6. Yazdanian (4);[11f] 7. Grieg (3);[11g] 8. Lin (3);[11h] 9. Lombardo (2);[11i] 10. Van Belle (2);[11j] 11. Calder (1).[11k] Although all the experimental protocols are not identical, we deem them sufficiently similar to be included in the same data set. If a successful model is obtained, the inclusion of all the data appears justified. The resulting data set contains 106 neutral compounds with the corresponding BB values. These BB values were converted to the logarithm basis for analysis. See Figure 1 for the structures and the experimentally derived logBB values.

Compounds were entered as structure drawings with ChemDraw[27] and structure data saved as mol files. All structure indices used in this investigation were computed from Molconn-Z, version 3.50.[28] Structure input was validated by visual inspection of the ChemDraw drawings as well as the structure analysis provided by Molconn-Z output.

For QSAR analysis, we selected atom-type E-State and hydrogen E-State indices, along with molecular connectivity chi and difference chi indices, and kappa shape indices. The pairwise correlation matrix was examined for correlation coefficients greater than 0.80. For each such occurrence, one of the pair of correlated variables was eliminated. Selection of a variable to be retained is primarily experience-based. Preference for retention is given to variables thought to be more easily interpreted in terms of molecular structure. For example, in this data set, the valence molecular connectivity chi path indices are correlated with the first-order chi valence index ${}^1\chi^v$ (r > 0.80). We selected the ${}^1\chi^v$ index (deleting the others) to include in the modeling because it may be more easily interpretable. A similar analysis of the difference chi indices, $d^m\chi^v$, leads to selection of the $d^2\chi^v$ index. After data matrix reduction, 41 variables remained for statistical analysis in model development.

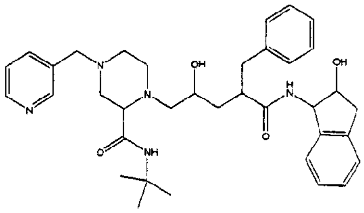
Statistical Analysis. The data matrix was submitted for statistical analysis using the SAS system.[29] The RSQUARE selection method in proc REG was used to examine every QSAR equation from one to five variables, listing the top 10 most statistically significant. The number of variables in a model was limited to five. The RSQUARE procedure is not a stepwise method; all possible sets of variables are considered and ranked on the r^2 values.

The three most prominent descriptors are E-State indices that emphasize hydrogen bond donating ability, $HS^T(HBd)$, along with certain nonpolar structure features, $HS^T(arom)$, and skeletal architecture, in the difference chi index, $d^2\chi^v$.

At this point in the model development, we introduced squares of the prominent variables to simulate a possible nonlinear relation between structure and logBB. Significant improvement was obtained in the regression statistics. Only the squares of the $HS^T(arom)$ and $d^2\chi^v$ variables were found to be statistically significant. In four- and five-variable equations, only the linear term in $d^2\chi^v$ was observed to be (marginally) statistically significant when the squares were also present. Based on these preliminary investigations, three structure descriptors were selected for further modeling and validation: $HS^T(HBd)$, $[HS^T(arom)]^2$, and $[d^2\chi^v]^2$.

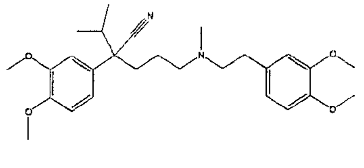
Development of Final Model. Based on the preliminary analysis and on validation studies (see below), the three selected structure descriptors were considered good choices

1.



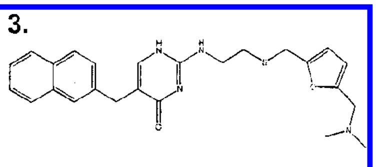
indinavir
logBB = -0.745

2.



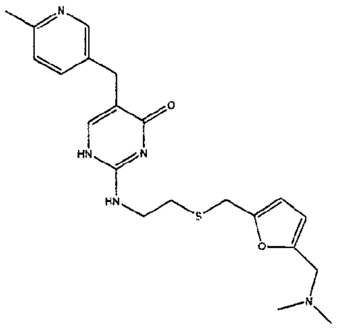
verapamil
logBB = -0.700

3.



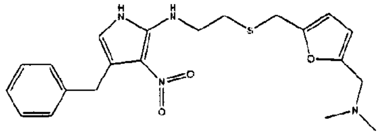
SK&F 93319
logBB = -1.300

4.



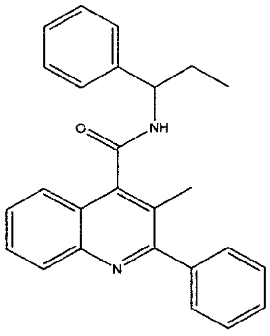
lupitidine
logBB = -1.060

5.



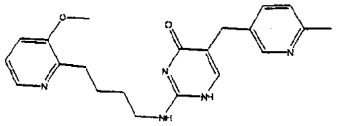
BBcpd60 (ranitidine analog)
logBB = -0.730

6.



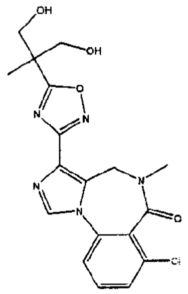
SB-222200
logBB = 0.300

7.



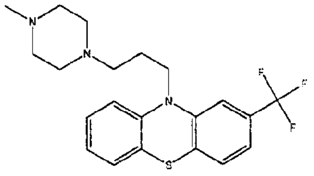
icotidine
logBB = -2.000

8.



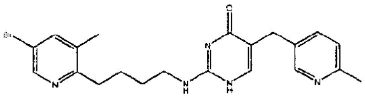
bis-hydroxylated L-663,581
logBB = -1.820

9.



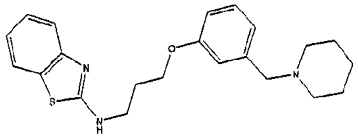
trifluoroperazine
logBB = 1.440

10.



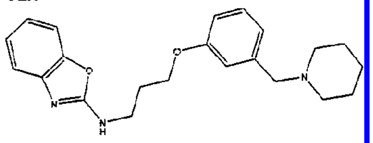
temelastine
logBB = -1.880

11.



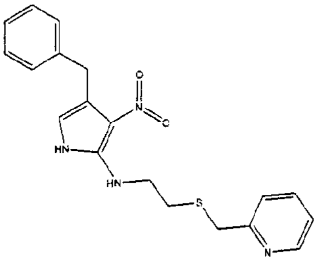
zolantidine (ranitidine analog)
logBB = 0.140

12.



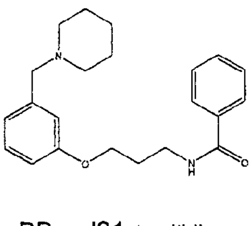
BBcpd26 (ranitidine analog)
logBB = 0.220

13.



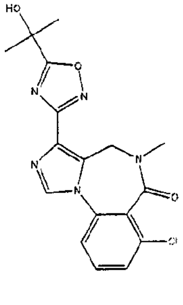
BBcpd14(cimetidine derivative)
logBB = -0.120

14.



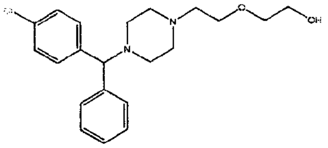
BBcpd21 (ranitidine analog)
logBB = -0.240

15.



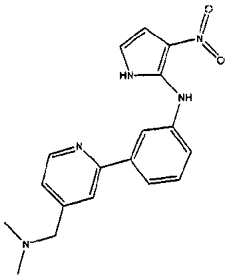
mono-hydroxylated L-663,581
logBB = -1.340

16.



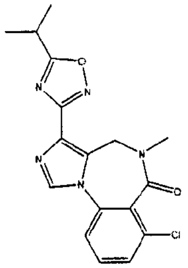
hydroxyzine
logBB = 0.390

17.



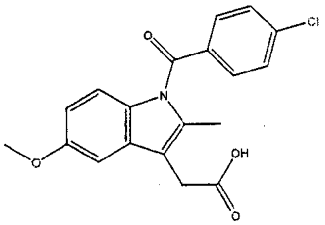
BBcpd19 (ranitidine analog)
logBB = -0.280

18.



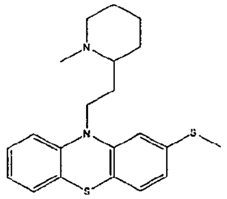
L-663,581
logBB = -0.300

19.



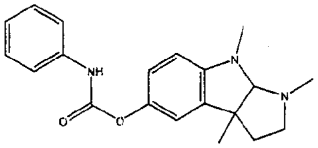
indomethacin
logBB = -1.260

20.



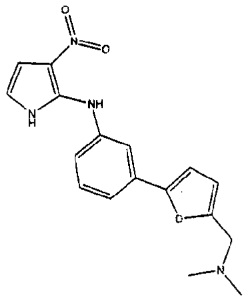
thioridazine
logBB = 0.240

21.



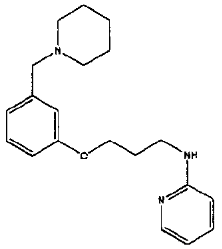
phenserine
logBB = 1.000

22.



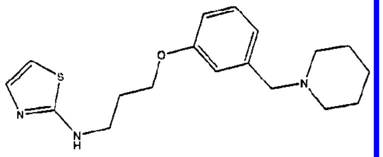
BBcpd18 (ranitidine analog)
logBB = -0.270

23.



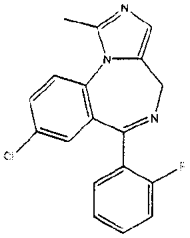
BBcpd23 (ranitidine analog)
logBB = 0.690

24.



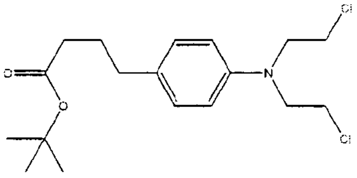
BBcpd24 (ranitidine analog)
logBB = 0.440

25.



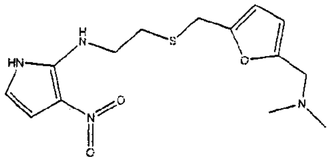
midazolam
logBB = 0.360

26.



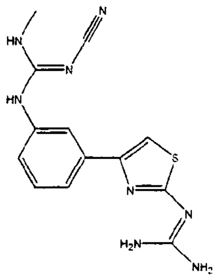
tertbutylchlorambucil
logBB = 1.000

27.



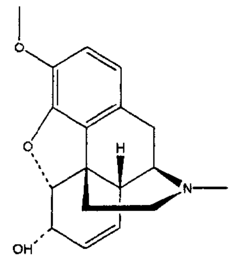
BBcpd17 (ranitidine analog)
logBB = -1.120

28.



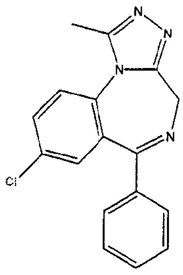
BBcpd58 (guanidinothiazole derivative)
logBB = -1.540

29.



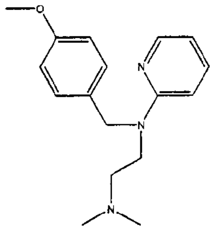
codeine
logBB = 0.550

30.



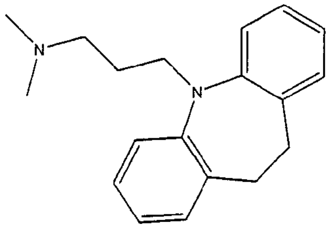
alprazolam
logBB = 0.044

31.



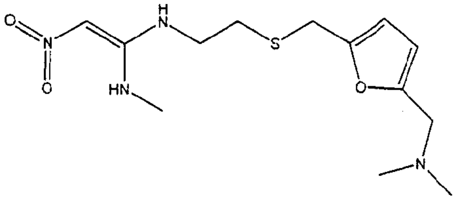
mepyramine
logBB = 0.490

32.



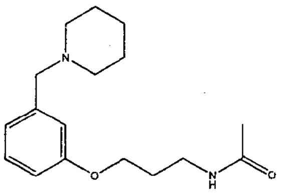
imipramine
logBB = 1.070

33.



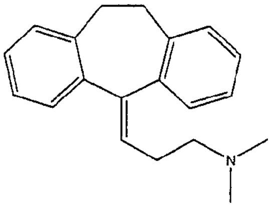
ranitidine
logBB = -1.230

34.



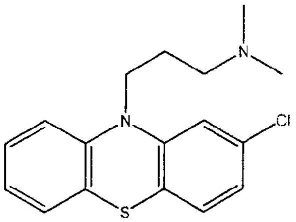
BBcpd20 (ranitidine analog)
logBB = -0.460

35.



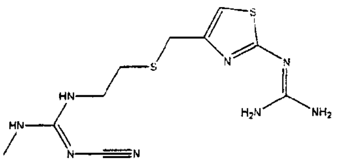
amitryptalline
logBB = 0.886

36.



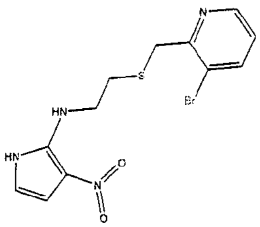
chlorpromazine
logBB = 1.060

37.



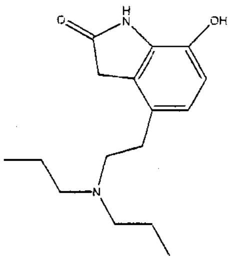
tiotidine
logBB = -0.820

38.



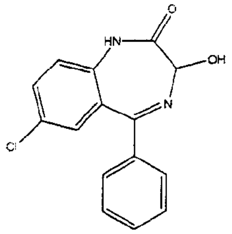
BBcpd12 (cimetidine derivative)
logBB = -0.670

39.



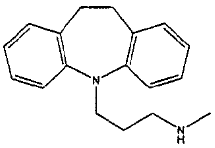
SKF89124
logBB = -0.060

40.



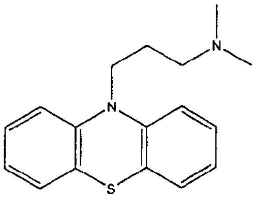
oxazepam
logBB = 0.610

41.



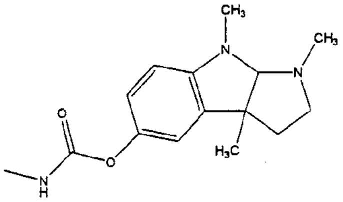
desipramine
logBB = 1.200

42.



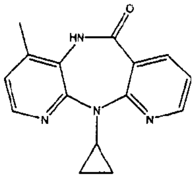
promazine
logBB = 1.230

43.



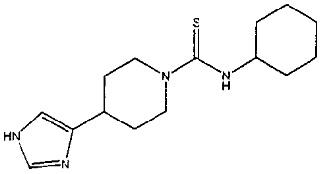
physostigmine
logBB = 0.079

44.



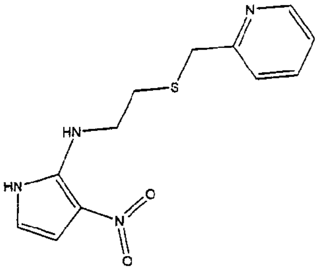
nevirapine
logBB = 0.000

45.



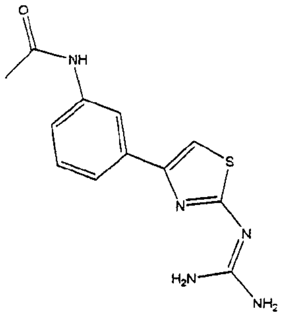
thioperamide
logBB = -0.160

47.



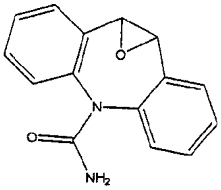
BBcpd13 (cimetidine derivative)
logBB = -0.660

48.



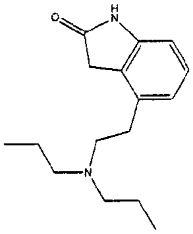
BBcpd16 (guanidinothiazole derivative)
logBB = -1.570

49.



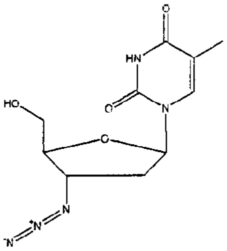
CBZ-EPO (carbamazepine-10,11-epoxide)
logBB = -0.350

50.



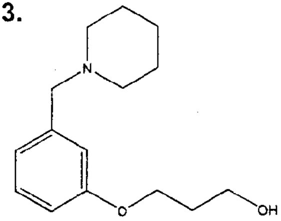
SKF101468
logBB = -0.300

51.



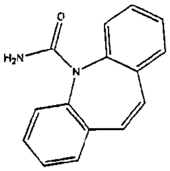
zidovudine
logBB = -0.720

53.



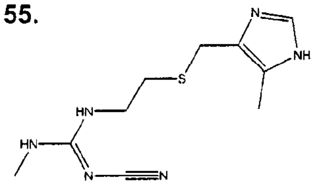
BBcpd22 (ranitidine analog)
logBB = -0.020

54.



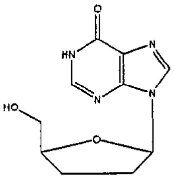
CBZ- carbamazepine
logBB = -0.140

55.



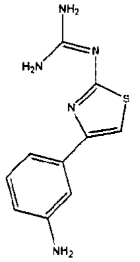
cimetidine
logBB = -1.420

56.



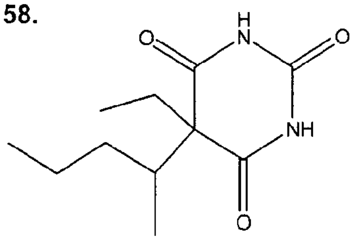
didanosine
logBB = -1.301

57.



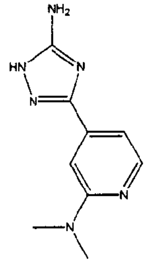
BBcpd57 (guanadinothiazole derivative)
logBB = -1.150

58.



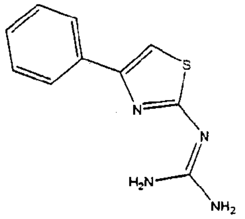
pentobarbital
logBB = -0.120

59.



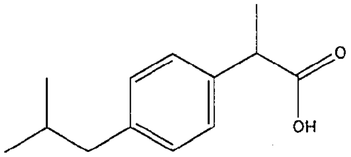
BBcpd10
logBB = -1.170

60.



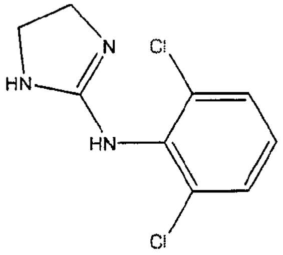
BBcpd15 (guanadinothiazole derivative)
logBB = -0.180

61.



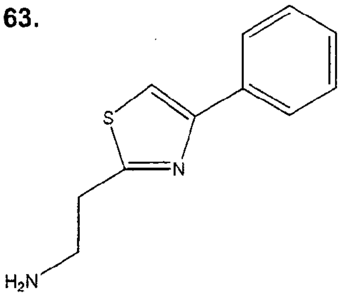
ibuprofen
logBB = -0.180

62.



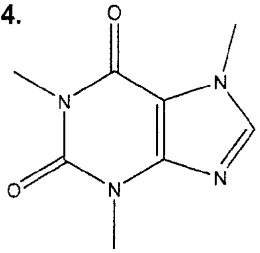
clonidine
logBB = 0.110

63.



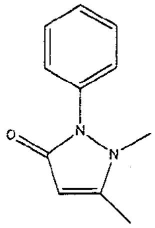
Y-G19
logBB = -0.430

64.



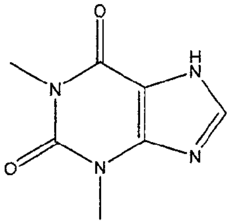
caffeine
logBB = -0.055

65.



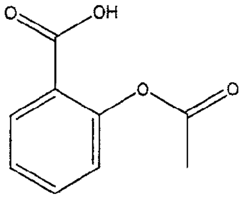
antipyrine
logBB = -0.097

66.



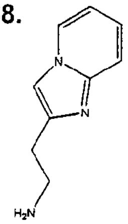
theophyline
logBB = -0.290

67.



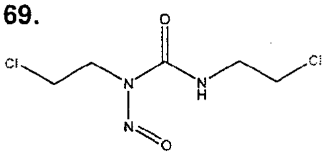
acetylsalicylic acid
logBB = -0.500

68.



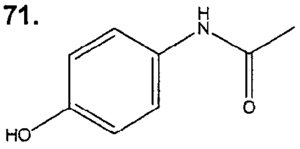
Y-G20
logBB = 0.250

69.



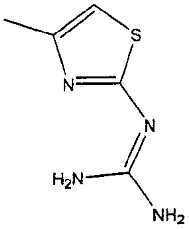
BCNU
logBB = -0.520

71.



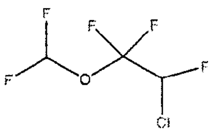
p-acetomidophenol
logBB = -0.310

72.



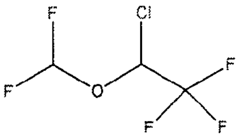
ICI17148
logBB = -0.040

73.



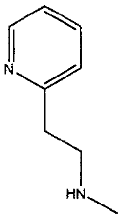
enflurane
logBB = 0.240

74.



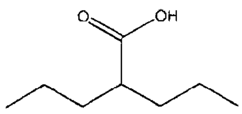
isoflurane
logBB = 0.420

75.



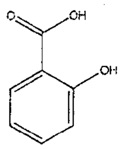
Y-G14
logBB = -0.420

76.



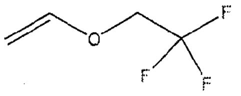
valproic acid
logBB = -0.220

77.



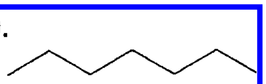
salicylic acid
logBB = -1.100

78.



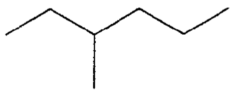
fluroxene
logBB = 0.130

80.



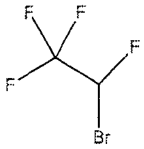
heptane
logBB = 0.810

81.



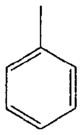
3-methylhexane
logBB = 0.900

82.



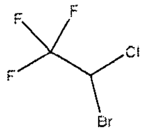
teflurane
logBB = 0.270

83.



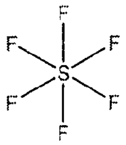
toluene
logBB = 0.370

84.



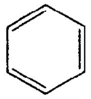
halothane
logBB = 0.350

85.



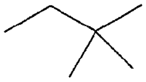
sulfur hexafluoride
logBB = 0.360

86.



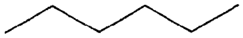
benzene
logBB = 0.370

87.



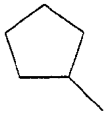
2,2-dimethylbutane
logBB = 1.040

88.



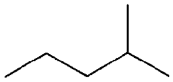
hexane
logBB = 0.800

89.



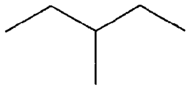
methylcyclopentane
logBB = 0.930

90.



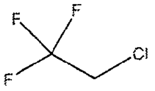
2-methylpentane
logBB = 0.970

91.



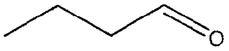
3-methylpentane
logBB = 1.010

92.



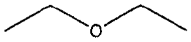
1,1,1-trifluro-2-chloroethane
logBB = 0.080

93.



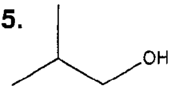
butanone
logBB = -0.080

94.



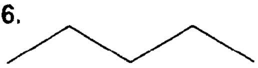
diethyl ether
logBB = 0.000

95.



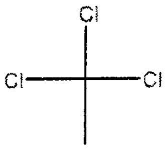
2-methylpropanol
logBB = -0.170

96.



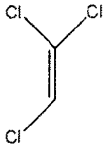
pentane
logBB = 0.760

97.



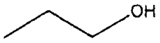
1,1,1-trichloroethane
logBB = 0.400

98.



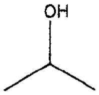
trichloroethene
logBB = 0.340

99.



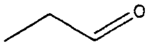
1-propanol
logBB = -0.160

100.



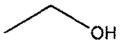
2-propanol
logBB = -0.150

101.



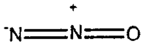
propanone
logBB = -0.150

102.



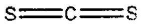
ethanol
logBB = -0.160

103.



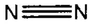
nitrous oxide
logBB = 0.030

104.



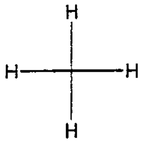
carbon disulfide
logBB = 0.600

105.



nitrogen
logBB = 0.030

106.



methane
logBB = 0.040

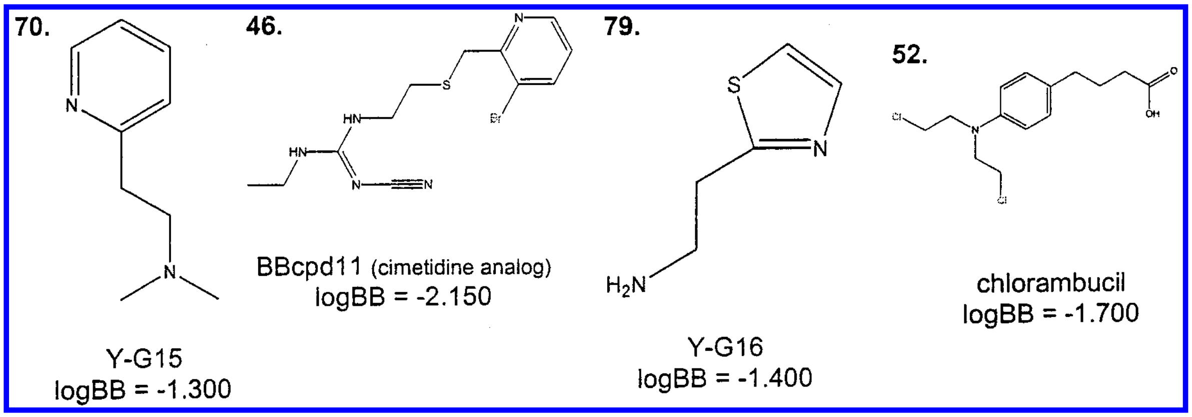


Figure 1. Structures of compounds used in this study, ranked by molecular weight. Name (when available) and logBB values provided. Structures deleted from modeling process are given at the bottom of the list.

for this data set. To attempt some measure of model validation, a set of 20 compounds was randomly selected as an initial validation test set. A model was developed on the remaining 86 compounds; the test set values were then predicted. This process is described in detail in the next section. The results of this test prediction were very good and provided impetus for support of the three structure descriptors.

To create a final model that could be used for prediction of compounds not in the current data set, we established a model for the whole set of 106 compounds and carefully examined the residuals. To optimize predictive ability, we decided to remove observations with large residuals, that is, residuals greater than 2.5 standard deviations. Four observations were found with residuals greater than ±1.2. These observations (nos. 46, 52, 70, and 79) were deleted, and the final model obtained and given as eq 4. The q^2 value improved from 0.52 to 0.62 upon removal of the four observations.

Validation Studies. Four approaches to validation of the model were adopted for this data set.

1. Validation Test Set. Twenty compounds were set aside from the full set as a validation test set, as indicated above. These compounds were selected randomly with the proviso that none of the compounds possesses the maximum or minimum logBB values among the 106 observations in the total set. In addition, the validation compounds were selected from those whose descriptor values were not on the boundaries of the structure parameter space. A model based on $HS^T(HBd)$, $[HS^T(arom)]^2$, and $[d^2\chi^v]^2$ was obtained from the 86 observations remaining after removal of the external validation set. The model was then used to predict logBB values for the validation test set. Because the compounds in the test set were subsequently placed back into the data set, we do not call this set an external validation set. Perhaps, a "preliminary validation test" set may be an appropriate term.

2. Cross-Validation. A full cross-validation test of the model was investigated. The data set of 102 compounds was divided randomly into five groups of (approximately) the same size (21 observations, 20%). Each group was left out (Leave-Group-Out, LGO) and that group predicted by a model developed from the remaining observations.[24,25] In this way, every observation was left out once, in groups of 20, and its value predicted.

3. External Prediction of CNS ± Data. A further test of the model (eq 4) was done by considering the blood-brain barrier penetration on a + or − scale. Testa examined experimental data for 108 compounds and assigned them as "+" if their logBB value > 0 and "−" if logBB < 0.[1] From this data set we were able to select 28 compounds whose BB values were obtained by methods similar to those for our data set and that are not already included in our data set. Each compound was drawn in ChemDraw, its mol file saved, and all the mol files run through Molconn-Z. Using the model in eq 4, these compounds were also predicted and ± values assigned. The results are given in Table 2.

4. Prediction of Large Database of Drug-like Molecules. In one final test, our model was used to predict logBB values for 20 039 drug and drug-like compounds taken from a modified form of the Pomona MedChem database.[30] Organometallic compounds and salts were removed; the structures were represented as SMILES strings. The compounds were selected from a somewhat larger data set by selecting only those whose structure descriptor values fall within the parameter values of our data set.

RESULTS AND DISCUSSION

The QSAR Model. The model based on three variables yielded statistical information as follows:

$$\begin{aligned} \text{logBB} = {} & -0.202(\pm 0.026)\ HS^T(HBd) \\ & + 0.00627(\pm 0.00095)\ [HS^T(arom)]^2 \\ & - 0.105(\pm 0.016)\ [d^2\chi^v]^2 - 0.425(\pm 0.069) \qquad (4) \\ & r^2 = 0.66,\ s = 0.45,\ F = 62.4,\ n = 102 \\ & q^2 = 0.62,\ s_{press} = 0.48 \end{aligned}$$

Quantities in parentheses are the standard deviations of the coefficients. The statistical quantities q^2 ($= r^2_{press}$) and s_{press} are based on the leave-one-out (LOO) method. A plot of the calculated logBB versus observed logBB values is given in Figure 2. An examination of the plot of residuals versus observed logBB (not shown) revealed no trends and appears randomly distributed. The observed logBB values, calculated (calc), residuals (res), and predicted residuals (pres from LOO) are given in Table 1. The variables in the model (eq 4) are essentially independent; the largest intercorrelation is $r^2 = 0.36$ for $[HS^T(arom)]^2$ and $[d^2\chi^v]^2$; the next largest is $r^2 = 0.25$ for $[d^2\chi^v]^2$ and $HS^T(HBd)$.

Table 1. Blood Brain Barrier Partitioning Data Observed, Calculated, and Residual Values

obsd[a]	name[b]	logBB[c]	calcd[d]	res[e]	pres[f]
1	indinavir	−0.745	−1.115	0.370	0.499
2	verapamil	−0.700	−0.775	0.075	0.099
3	SK&F 93319	−1.300	−1.000	−0.300	−0.311
4	lupitidine	−1.060	−0.891	−0.169	−0.174
5	BBcpd60 (ranitidine analog)	−0.730	−0.635	−0.095	−0.098
6	SB-222200	0.300	1.052	−0.752	−1.027
7	icotidine	−2.000	−1.519	−0.481	−0.558
8	bis-hydroxylated L-663,581 metabolite	−1.820	−1.766	−0.054	−0.060
9	trifluoroperazine	1.440	0.255	1.185	1.248
10	temelastine	−1.880	−0.936	−0.944	−0.973
11	zolantidine (ranitidine analog)	0.140	0.344	−0.204	−0.210
12	BBcpd26 (ranitidine analog)	0.220	−0.190	0.410	0.424
13	BBcpd14 (cimetidine derivative)	−0.120	−0.254	0.134	0.141
14	BBcpd21 (ranitidine analog)	−0.240	0.040	−0.280	−0.291
15	mono-hydroxylated L-663,581 metabolite	−1.340	−0.873	−0.467	−0.488
16	hydroxyzine	0.390	0.022	0.368	0.380
17	BBcpd19 (ranitidine analog)	−0.280	−0.536	0.256	0.272
18	L-663,581	−0.300	−0.232	−0.068	−0.073
19	indomethacin	−1.260	−0.766	−0.494	−0.531
20	thioridazine	0.240	0.886	−0.646	−0.676
21	phenserine	1.000	0.235	0.765	0.788
22	BBcpd18 (ranitidine analog)	−0.270	−0.770	0.500	0.526
23	BBcpd23 (ranitidine analog)	0.690	0.094	0.596	0.608
24	BBcpd24 (ranitidine analog)	0.440	0.119	0.321	0.325
25	midazolam	0.360	0.287	0.073	0.076
26	tertbutylchlorambucil	1.000	0.355	0.645	0.658
27	BBcpd17 (ranitidine analog)	−1.120	−0.726	−0.394	−0.402
28	BBcpd58 (guanidinothiazole derivative)	−1.540	−1.772	0.232	0.254
29	codeine	0.550	−0.285	0.835	0.850
30	alprazolam	0.044	0.638	−0.594	−0.619
31	mepyramine	0.490	0.517	−0.027	−0.028
32	imipramine	1.070	0.833	0.237	0.247
33	ranitidine	−1.230	−0.845	−0.385	−0.395
34	BBcpd20 (ranitidine analog)	−0.460	−0.261	−0.199	−0.202
35	amitryptalline	0.886	0.842	0.044	0.046
36	chlorpromazine	1.060	0.875	0.185	0.194
37	tiotidine	−0.820	−1.452	0.632	0.694
38	BBcpd12 (cimetidine derivative)	−0.670	−0.527	−0.143	−0.146
39	SKF89124	−0.060	−0.927	0.867	0.897
40	oxazepam	0.610	−0.526	1.136	1.184
41	desipramine	1.200	0.435	0.765	0.788
42	promazine	1.230	0.973	0.257	0.271
43	physostigmine	0.079	−0.117	0.196	0.198
44	nevirapine	0.000	−0.159	0.159	0.160
45	thioperamide	−0.160	−0.367	0.207	0.214
47	BBcpd13 (cimetidine derivative)	−0.660	−0.513	−0.147	−0.150
48	BBcpd16 (guanidinothiazole derivative)	−1.570	−0.964	−0.606	−0.630
49	CBZ-EPO− carbamazepine-10,11-epoxide	−0.350	0.231	−0.581	−0.596
50	SKF101468	−0.300	−0.177	−0.123	−0.125
51	zidovudine	−0.720	−1.375	0.655	0.692
53	BBcpd22 (ranitidine analog)	−0.020	−0.261	0.241	0.244
54	CBZ- carbamazepine	−0.140	0.174	−0.314	−0.321
55	cimetidine	−1.420	−1.011	−0.409	−0.431
56	didanosine	−1.301	−1.059	−0.242	−0.251
57	BBcpd57 (guanidinothiazole derivative)	−1.150	−0.666	−0.484	−0.507
58	pentobarbital	0.120	−0.757	0.877	0.904
59	BBcpd10	−1.170	−0.617	−0.553	−0.564
60	BBcpd15 (guanidinothiazole derivative)	−0.180	−0.196	0.016	0.016
61	ibuprofen	−0.180	−0.144	−0.036	−0.037
62	clonidine	0.110	−0.444	0.554	0.571
63	Y−G19	−0.430	0.327	−0.757	−0.774
64	caffeine	−0.055	0.056	−0.111	−0.114
65	antipyrine	−0.097	0.343	−0.440	−0.449
66	theophyline	−0.290	−0.323	0.033	0.034
67	acetylsalicylic acid	−0.500	−0.589	0.089	0.091
68	Y−G20	0.250	−0.111	0.361	0.366
69	BCNU	−0.520	−0.311	−0.209	−0.212
71	*p*-acetamidophenol	−0.310	−0.637	0.327	0.338
72	ICI 17148	−0.040	−0.367	0.327	0.337
73	enflurane	0.240	−0.009	0.249	0.257
74	isoflurane	0.420	0.000	0.420	0.433
75	Y−G14	−0.420	0.075	−0.495	−0.502
76	valproic acid	−0.220	−0.256	0.036	0.037
77	salicylic acid	−1.100	−0.856	−0.244	−0.257

Table 1 (Continued)

obsd[a]	name[b]	logBB[c]	calcd[d]	res[e]	pres[f]
78	fluroxene	0.130	0.116	0.014	0.014
80	heptane	0.810	0.338	0.472	0.483
81	3-methylhexane	0.900	0.333	0.567	0.580
82	teflurane	0.270	0.266	0.004	0.004
83	toluene	0.370	0.507	−0.137	−0.140
84	halothane	0.350	0.338	0.012	0.013
85	sulfur hexafluoride	0.360	0.282	0.078	0.079
86	benzene	0.370	0.567	−0.197	−0.202
87	2,2-dimethylbutane	1.040	0.197	0.843	0.862
88	hexane	0.800	0.338	0.462	0.472
89	methylcyclopentane	0.930	0.293	0.637	0.651
90	2-methylpentane	0.970	0.316	0.654	0.668
91	3-methylpentane	1.010	0.334	0.676	0.692
92	1,1,1-trifluoro-2-chloroethane	0.080	0.291	−0.211	−0.216
93	butanone	−0.080	0.326	−0.406	−0.415
94	diethyl ether	0.000	0.311	−0.311	−0.318
95	2-methylpropanol	−0.170	−0.148	−0.022	−0.023
96	pentane	0.760	0.338	0.422	0.431
97	1,1,1-trichloroethane	0.400	−0.127	0.527	0.551
98	trichloroethene	0.340	0.335	0.005	0.005
99	1-propanol	−0.160	−0.140	−0.020	−0.020
100	2-propanol	−0.150	−0.145	−0.005	−0.005
101	propanone	−0.150	0.330	−0.480	−0.491
102	ethanol	−0.160	−0.136	−0.024	−0.025
103	nitrous oxide	0.030	0.337	−0.307	−0.314
104	carbon disulfide	0.600	0.337	0.263	0.269
105	nitrogen	0.030	0.338	−0.308	−0.315
106	methane	0.040	0.338	−0.298	−0.305
46[g]	BBcpd11 (cimetidine derivative)	−2.150	−0.569	**−1.581**	−1.616
52[g]	chlorambucil	−1.700	−0.407	**−1.293**	−1.309
70[g]	Y−G15	−1.300	0.431	**−1.731**	−1.767
79[g]	Y−G16	−1.400	0.085	**−1.485**	−1.513

[a] Observation number. Note: Numbers out of sequence (listed at the bottom of the table) are observations that were dropped from the model due to large residuals (see text). [b] Compound name where available. (See Figure 1 for structure drawings). [c] Observed logBB value.[11] [d] logBB value calculated from QSAR model, eq 4. [e] logBB-calc. [f] Residual for compound during leave-one-out (LOO) cross-validation. [g] Compounds 46, 52, 70, and 79 were dropped from the training set during the modeling process (see text).

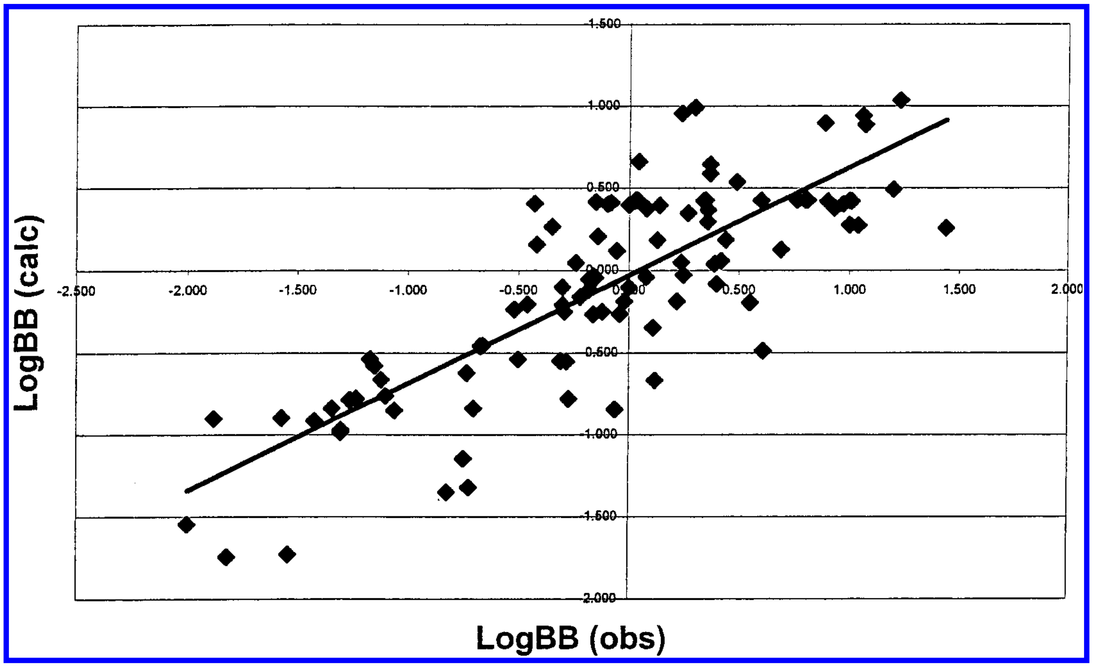


Figure 2. Plot of calculated logBB values (QSAR model, eq 4) versus observed logBB for the blood-brain barrier partitioning data set.

The statistical criteria for selection of the QSAR model for the logBB data are based on the following: r^2 and q^2 for the final model ($n = 102$); the mean absolute error (MAE) for the test set ($n = 20$); and the MAE for the cross-validation (5-fold cross-validation). The MAE value was considered the most important statistic because an equation should not be considered a QSAR model unless it has demonstrated predictive value. (See below.)

These statistical results compare favorably with others reported in the literature.[11b,e,12] Smaller data sets were used

Table 2. CNS ± Prediction Results, External Validation Set

name[a]	log BB[b]	observed ±[a]	predicted ±[c]
clobazam	0.523	+	+
diazepam	0.677	+	+
diphenhydramine	1.015	+	+
estradiol	0.466	+	+
nordazepam	0.272	+	+
morphine	0.503	+	+
progesterone	0.344	+	+
testosterone	0.377	+	+
mefloquine	−2.588	-	-
carbidopa	−2.279	-	-
cetirizine	−0.289	-	-
ciprofloxacin	−1.250	-	-
difloxacin	−1.485	-	-
dopamine	−1.031	-	-
enoxacin	−1.730	-	-
terfenadine	1.055	-	+
fleroxacin	−2.056	-	-
furosemide	−1.117	-	-
isoxicam	−0.747	-	-
lomefloxacin	−1.936	-	-
meloxicam	−0.365	-	-
corticosterone	−0.595	-	-
norfloxacin	−1.552	-	-
ofloxacin	−1.227	-	-
perfloxacin	−1.133	-	-
piroxicam	−0.290	-	-
tenoxicam	−0.332	-	-
rufloxacin	−0.649	-	-

[a] Name and observed ± data as given in ref 1. [b] logBB value calculated from QSAR model, eq 4. [c] ± designation as derived from calculated logBB, eq 4. If logBB <0, compound is CNS -, otherwise compounds is CNS +.

in the other investigations. Keseru and Molnar, using a solvation model based on 3D structure information, developed a relation for a set of 60 compounds.[12] Luco used a data set of 61 compounds.[11e] In each of those cases, a model was developed, several outliers were removed, and predictions were made on test sets. In this present investigation, all the data reported by the other investigators have been included in the present data set. The results of predictions on the validation test set as well as on the full cross-validation tests (see below) gave mean absolute errors either comparable to or lower than those obtained by the other authors.[11,12]

Methods based on 3D structure information and detailed atom−atom interaction calculations require significantly more computation time. Keseru and Molnar report that their calculations require 6 s per molecule (10 molecules computed per minute).[12] The model presented in this work, however, is much faster, producing about 5000−6000 molecules computed per minute.

Interpretation of the Model. The three-variable model (eq 4) adequately represents the logBB data, based on direct statistics as well as validation methods. Each of the variables is a descriptor of an aspect of molecular structure and will be discussed to indicate the specific structure information encoded.

The HS^T(HBd) Descriptor. The variable with the single best correlation to logBB is HS^T(HBd), a representation of hydrogen bond donation ability.[3,19,24] The descriptor, HS^T(HBd), is the sum of the hydrogen E-State values for groups that act as hydrogen bond donors. In this present data set, these groups include $-OH$, $-NH_2$, and $-NH-$ donors. The numerical contribution to HS^T(HBd) is greater for $-OH$ than for either amine. The negative coefficient on HS^T(HBd) indicates that hydrogen bond donor groups lead to negative or low values of logBB. This effect arises from increased water solubility and decreased membrane affinity of compounds with hydrogen bond donor groups. The HS^T(HBd) variable contributes 32.7% on average to the calculated logBB value and ranges from 0.0% to 100.0% across the data set.

An important consideration in the structure interpretation from this model is the ability to estimate the contribution of hydrogen bonding groups to the calculated logBB value. HS^T(HBd) is the sum of Hs values; the contribution to logBB is the product of the hydrogen E-State value (Hs) of a donor group times the coefficient of HS^T(HBd) in the model, −0.202. For the data set, the average contribution of an $-OH$ group is −0.53 and ranges from −0.46 to −0.55. The contribution varies from molecule to molecule because of the influence of the structure of the rest of the molecule on an $-OH$ group. The encoding of the influence of molecular environment on hydrogen bond donating ability results from the nature of the hydrogen E-State formalism. This encoding is not a mere count of donors, but it is a variable measure of hydrogen bond donating ability arising from hydrogen atom accessibility for intermolecular interactions.[3,15,19]

For a secondary amine group, $-NH-$, the average contribution to logBB is −0.45 and varies from −0.29 to −0.80. This range of values is an indication of the wide range of structure contexts in which $-NH-$ groups are found in this data set. For a primary amine, $-NH_2$, the contribution to logBB is −0.33 and ranges from −0.28 to −0.36 in this data set. This structure information on these hydrogen bond donor groups may be of assistance in the design of new compounds.

The HS^T(arom) Descriptor. The second variable in the model is the square of the atom-type hydrogen E-State descriptor for aromatic CH groups, HS^T(arom). This descriptor of nonpolar aromatic CHs is the sum of the hydrogen atom-level E-State indices for all aromatic hydrogen atoms in the molecule.[3,21,26] The HS^T(arom) descriptor contributes 21.6% on average to the calculated logBB values and ranges from 0.0% to 99.8%. Because of the positive coefficient on HS^T(arom), larger values are related to larger logBB values.

The contribution of a single aromatic CH group can be computed from the product of the square of its Hs value times the coefficient of $[HS^T(arom)]^2$, 0.00627. A typical aromatic CH contribution to logBB is +0.011. For a phenyl group, the HS^T(arom) contribution to logBB is +0.27; for two phenyl groups in a molecule the contribution is +1.08. Again it should be noted that the HS^T(arom) descriptor is not a simple count; the numerical value of the descriptor is influenced by the structure context in which the aromatic CH group is located. Nearby electronegative groups tend to increase Hs values.

This discussion indicates the combined significance of specific hydrogen bond donating groups and of nonpolar, aromatic regions toward logBB. The third structure descriptor provides detailed molecular architecture information.

The $d^2\chi^v$ Descriptor. The third variable in the model is the square of the second-order valence molecular connectivity difference chi index, $d^2\chi^v$. This variable increases with increased branching in the structure; it is a measure of overall skeletal structure variation.[20,22] The $d^2\chi^v$ index is derived from

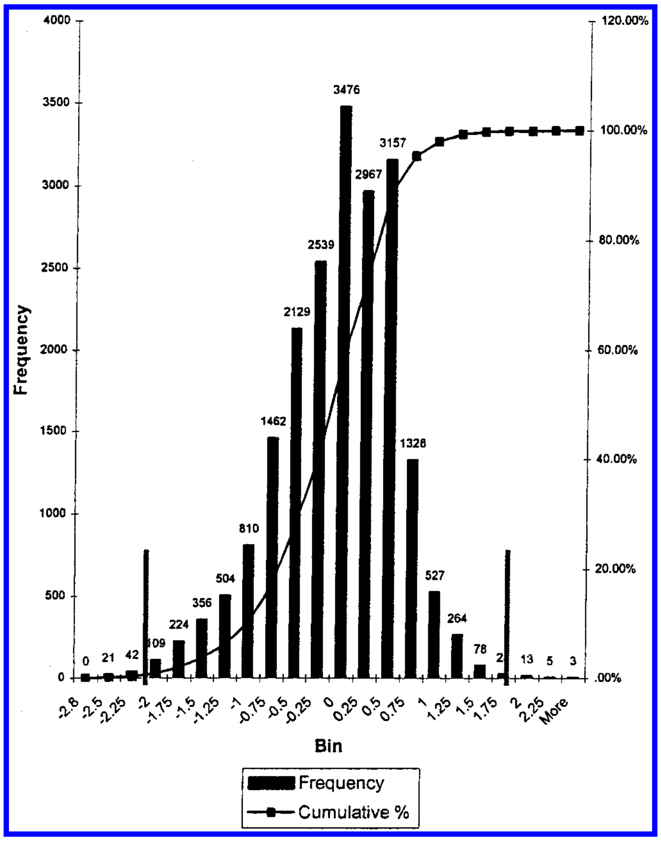


Figure 3. Histogram of predictions (QSAR model, eq 4) for large external database of 20 039 drug and drug-like compounds. The vertical lines mark the boundaries of the logBB values in the training set.

the ${}^2\chi^v$ descriptor, an index that increases with molecular size. The difference chi index is a normalized index, obtained by subtracting the index value for a reference structure from the ${}^2\chi^v$ index. The reference structure is the unbranched structure of the same number of atoms.

$$d^2\chi^v = {}^2\chi^v - {}^2\chi_{ref}v$$

(for unbranched structure with same atoms)

The result of this normalization is to eliminate the dependence on molecular size. The $d^2\chi^v$ index contributes 45.6% on the average to the calculated logBB and ranges from 0.0% to 100.0% across the data set. Because of the negative coefficient on $(d^2\chi^v)^2$, larger values are related to more negative logBB values.

The $d^2\chi^v$ index encodes variation in the molecular architecture, especially skeletal branching. For example, to see how degree of skeletal branching is encoded, consider the set of hexane isomers, as shown in Figure 4. The ranking of $d^2\chi^v$ values for the isomers clearly indicates that the greater the degree of skeletal branching, the greater the numerical value of $d^2\chi^v$. Furthermore, the more electronegative the heteroatoms in the structure, the larger and more negative is the $d^2\chi^v$ index value, as indicated by the set of phenethyl structures shown in Figure 4. Because the coefficient of $[d^2\chi^v]^2$ is −0.105, the more branched the skeleton, the greater is the $d^2\chi^v$ value and the more negative the contribution to logBB value. Also the greater the number of electronegative atoms, the greater is the magnitude of $[d^2\chi^v]^2$ and the more negative the contribution to the logBB value.

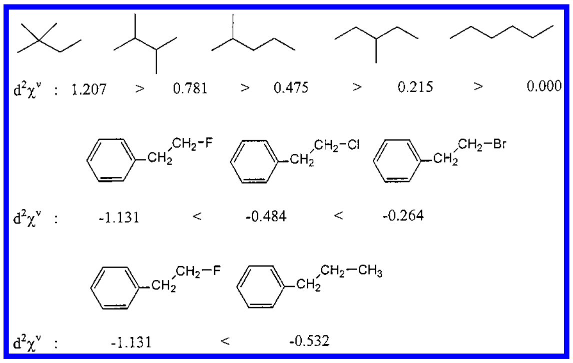

d²χᵛ : 1.207 > 0.781 > 0.475 > 0.215 > 0.000
CH₂–F
CH₂
CH₂–Cl
CH₂
CH₂–Br
CH₂
d²χᵛ : -1.131 < -0.484 < -0.264
CH₂–F
CH₂
CH₂–CH₃
CH₂
d²χᵛ : -1.131 < -0.532

Figure 4. A set of molecular structures that illustrates the relations of the $d^2\chi^v$ index to trends in molecular skeleton variation.

In summary, the model given as eq 4 indicates that blood-barrier penetration is increased for compounds with aromatic CH groups, less skeletal branching, and fewer or weaker hydrogen bond donor groups. The model permits the direct numerical estimation of logBB for organic structures and blends together the structure information of the three structure descriptors in eq 4.

Validation Studies. Four approaches to validation of the model were adopted for this data set.

1. Validation Test Set. A model based on $HS^T(HBd)$, $[HS^T(arom)]^2$, and $[d^2\chi^v]^2$ was obtained for the 86 observations remaining after removal of the validation test set of 20 observations. The model was then used to predict logBB values for the validation test set. The mean absolute error obtained was MAE = 0.32. The computed root-mean-square for these predicted logBB values is rms = 0.38. Although no accurate estimate is available for the experimental error in this eclectic data set, an MAE value of 0.33 (rms = 0.38) appears to be quite reasonable. This result tends to confirm the significance of the three selected structure descriptors and the model based on them.

2. Cross-Validation. For each group of observations left out (20%, 21 compounds) of the whole data set, a model was developed from the remaining 80% of the data. The 21 compounds left out were predicted by that model. This process was carried out five times on five unique subsets. In this way, every observation was predicted once (in its group of left-out observations). The mean absolute errors for the five groups are as follows: MAE = 0.32, 0.38, 0.51, 0.35, and 0.34. The overall mean absolute error was found to be MAE = 0.38 (rms = 0.47). For a 20% full leave-out, cross-validation procedure, this level of mean absolute error MAE (and rms) is good confirmation of the predictive quality of the model.

3. External Prediction of CNS ± Data. For the data set of Testa,[1] the logBB value of each compound was predicted with the model in eq 4. A "+" value was assigned if logBB(predicted) > 0; otherwise, a "−" value was assigned if logBB(predicted) < 0. The values are recorded in Table 2. The model of eq 4 correctly predicted 27/28 of the observations. This excellent result lends further strong confirmation to the validity of the model.

4. Prediction of Large Database of Drug-like Molecules. Using eq 4, the logBB values were computed for 20 039 drug and drug-like compounds obtained from the MedChem database. The results are shown as a histogram in Figure 3 in which frequency of computed logBB values is plotted against logBB. Experimental logBB values are not available for these compounds (although some of the entries are also found in our data set of 106 compounds). The objective of this predictive experiment is to determine whether the predicted logBB values are found in a range of values that is reasonable for a set of drugs and drug-like molecules. The histogram in Figure 4 indicates that the predicted values fall in a range consistent with drug-like molecule properties. No predicted values of logBB are unreasonably large positive nor negative values. A very small number of values (135) actually lie outside the logBB range of our data set, −2.15 to +1.44; these values could be expected to correspond to reasonable experimental values for those compounds.

The overall indication of these validation studies is that the topological logBB model appears reasonable and likely to give useful estimations of logBB values.

CONCLUSIONS

For the blood/brain barrier partition data set, an excellent QSAR model is developed with two E-State structure descriptors and one difference molecular connectivity chi index. Furthermore, validation and cross-validation of the QSAR model support the claim that the model may be used to make predictions for compounds not in the original data set. The structure information encoded in the three descriptors is presented and discussed, indicating significant and specific structure information that may be useful for new compound design. This encoded structure information may be interpreted directly in terms of structure features that can be communicated to synthetic chemists: presence and donating ability of hydrogen bond donors and presence and nonpolar nature of aromatic CH groups as well as degree of skeletal branching and electronegative atom content. These qualitative statements may be converted into numerical values of logBB by the QSAR model.

The reason for success of this topologically based method is the nature of the structure information used in algorithms that lead to the topological representation of molecular structure. For drug-receptor interactions and physicochemical properties, noncovalent interactions are dominant. These interactions among molecules are known to arise from the electron distribution across each molecule. In considering the nature of these noncovalent interactions, many investigators have found it practical to separate what are termed electronic attributes from what are called steric attributes. These two attributes are typically represented separately and entered into QSAR models independently.

At a fundamental level, however, both these aspects of structure (electronic and steric) arise from the same basis, the electron distribution across a molecule. In the E-State representation, the central feature is the intrinsic state term, I_i, which encodes in an integrated fashion both electronic and topological attributes. The intrinsic state is derived from the ratio of valence state electronegativity of an atom to a measure of its local topological character. These attributes of structure, essential in describing intermolecular interactions and providing the basis for relationships, are integrated in a single expression and employed in a unified fashion in QSAR models based on the E-State.[4] In this manner, the E-State indices describe the electron accessibility at each atom; the hydrogen E-State indices represent the hydrogen accessibility at each hydrogen atom. These accessibility attributes are very significant in noncovalent interactions essential to the development of QSAR models. In this sense, essential features of noncovalent intermolecular interaction are encoded in the E-State descriptors.

Topological descriptors such as the E-State and chi indices are representations of molecular structure that arise from the chemical identity of each atom, including valence state, and the nature of the set of connections in the molecular skeleton, the chemical bonding pattern.[4,15,19] Through appropriate mathematical processes described in the literature, this encoded information is transformed into significant structure information such as valence state electronegativity, whose atom-by-atom differences are strongly related to electron distribution.[4,15,19] Furthermore, significant relations among structure features are obtained from analysis of the network of chemical bonds,[20−23] leading to molecular connectivity chi indices of skeletal branching and kappa indices of molecular shape. The combination of valence state electronegativity and skeletal characterization leads to electron accessibility for the E-State formalism.[4] Furthermore, it has been recently shown that the potential for noncovalent intermolecular interaction, intermolecular accessibility, is the essence of the molecular connectivity formalism.[31]

ACKNOWLEDGMENT

K. Rose acknowledges partial support for this work by a Norris-Richards Summer Undergraduate Scholarship sponsored by the Northeast Section of the American Chemical Society.

REFERENCES AND NOTES

(1) Crivori, P.; Cruciani, G.; Carrupt P.; Testa, B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204−2216.

(2) Greer, J.; Erickson, J. W.; Baldwin, J. J.; Varney, M. D. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Design. *J. Med. Chem.* **1994**, *37*, 1035−1054.

(3) Bugg, C. E.; Carson, W. M.; Montgomery, J. A. Drugs by Design. *Sci. Am.* **1993**, December, 92−98.

(4) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, 1999.

(5) Kier, L. B.; Hall, L. H. Inhibition of Salicylamide Binding: An Electrotopological State Analysis. *Med. Chem. Res.* **1992**, *2*, 497−502.

(6) 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.

(7) 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.

(8) Huuskonen, J. QSAR Modeling with the Electrotopological State: TIBO Derivatives, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425−429.

(9) Pantakar, S. J.; Jurs, P. C. Prediction of IC_{50} Values for ACAT Inhibitors from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 706−723.

(10) Gozolbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. Molecular Search of New Active Drugs Against *Toxoplasma Gondii*. *SAR and QSAR Environ. Res.* **1999**, *10*, 47−60.

(11) (a) Young, R. C. Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *J. Med. Chem.* **1988**, *31*, 656−671. (b) Abraham, M. H.; Chadha H. S.; Mitchell, R. C. Hydrogen Bonding. Part 33. Factors that influence the distribution of solutes between blood and brain. *J. Pharm. Sci.* **1994**, *83*, 1257−1268. (c) Salminem, T.; Pulli, A.; Taskinen, J. Relationship between immobilized artificial membrane chromatographic retention and the brain penetration of structurally diverse drugs. *J. Pharm. Biomed. Anal.* **1997**, *15*, 469−677. (d) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sci.* **1999**, *83*, 815−821. (e) Luco, J. M. Prediction of brain-blood distribution of a

large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396−404. (f) Yazdanian, M.; Glynn, S. L. In vitro blood-brain barrier permeability of nevirapine compared to other HIV antiretroviral agents. *J. Pharm. Sci*. **1998**, *87*, 306−310. (g) Grieg, N. H.; Brossi, A.; Xue-Feng, P.; Ingram, D. K.; Soncrant, T. In *New Concepts of a Blood-Brain Barrier*; Greenwood J., et al., Eds.; Plenum: New York, 1995; pp 251−264. (h) Lin, J. H.; Chen, I.; Lin, T. Blood-brain barrier permeability and in vivo activity of partial agonists of benzodiazepine receptor: a study of L-663, 581 and its metabolites in rats. *J. Pharmacol. Exptl. Therapeut.* **1994**, *271*, 1197−1202. (i) Lombardo, F.; Blake, J. F.; Curatolo, W. Computation of brain-blood partitioning of organic solutes via free energy calculations. *J. Med. Chem.*. **1996**, *39*, 4750−4755. (j) Van Belle, K.; Sarre, S.; Ebinger, G; Michotte, Y. Brain, liver, and blood distribution kinetics of carbamazepine and its metabolic interaction with clomipramine in rats: a quantitative microdialysis study. *J. Pharmacol. Exptl. Therapeut*. **1995**, *272*, 1217−1222. (k) Calder, J. A. D.; Ganellin, R. Predicting the brain-penetrating capability of histaminergic compounds. *Drug Design Discovery* **1994**, *11*, 259−268.

(12) KeserÜ, G.; Molnár, L. High-throughput prediction of blood-brain partitioning: a thermodynamic approach. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 120−128.

(13) Parham, M. E.; Interactive Analysis, 6 Reuben Duren Way, Bedford, MA 01730; www. InterActiveAnalysis.com.

(14) Hall, L. H.; Mohney, 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.

(15) 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.

(16) Gough, J. D.; Hall, L. H. Modeling the Toxicity of Amide Herbicides using the Electrotopological State. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 356−361.

(17) Gough, J. D.; Hall, L. H. QSAR Models of the Antileukemic Potency of Carboquinones: Electrotopological State and Chi Indices. *Environ. Toc. Chem.* **1999**, *18*, 1069−1075.

(18) 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.

(19) 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.

(20) Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure−Activity Analysis; Research Studies Press: John Wiley and Sons: Chichester, U.K., 1986.

(21) 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.

(22) 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.

(23) 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.

(24) 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.

(25) (a) 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. (b) 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) 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. (b) Kier, L. B.; Hall, L. H. Database Organization and Searching with E-State Indices. *SAR QSAR Environ. Sci*. **2001**, *12*, 55−74.

(27) ChemDraw, ver 4.5; CambridgeSoft: Cambridge, MA 02139.

(28) Molconn-Z, ver 3.50; available from Hall Associates Consulting, 2 Davis Street, Quincy, MA, 02170; also from EduSoft, LC, P.O. Box 1811, Ashland, VA 23005; and SciVision, Inc., 200 Wheeler Road, Burlington, MA 01803.

(29) SAS, ver 8.0; SAS Institute: Cary, NC 27513.

(30) 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.

(31) (a) Kier, L. B.; Hall, L. H. Intermolecular accessibility: The meaning of molecular connectivity. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 792−795. (b) Kier, L. B.; Hall, L. H. Molecular connectivity: Intermolecular accessibility and encounter simulation. *J. Mol. Graphics Model*. **2001**, *20*, 76−83.

CI010127N