

E-State Modeling of Dopamine Transporter Binding. Validation of the Model for a Small Data Set

Hlaing Hlaing Maw and Lowell H. Hall*

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

Received March 18, 2000

Data for 25 tropane analogues binding to the dopamine transporter were modeled using E-state molecular structure descriptors. Both E-state and hydrogen E-state descriptors appear in the model in both atom-level and atom-type descriptors. A statistically satisfactory four-variable model is obtained, and structure interpretation is given for each variable, emphasizing substituent influence on nonpolar parts of the molecule as well as the role of hydrogen bonding. A leave-group-out approach to model validation is presented in which each observation is removed from the data set three times in random groups of 20% of the whole data set. The average of the resulting predicted values constitutes consensus predictions for these data and supports the claim that the E-state model may be useful for prediction of pIC_{50} values for new compounds.

INTRODUCTION

The importance of the dopamine transporter system, including a role in cocaine addiction, has been discussed by several investigators.^{1–7} Of the several possible pathways which might be involved in cocaine addiction, the one involving dopamine has been identified.^{6,7} (–)-Cocaine is known to inhibit the reuptake of dopamine, although details of the mechanism are not known. Hence, researchers recently have been investigating this important area.^{7,8} The dopamine hypothesis has suggested that the dopamine transporter is the significant monoamine transporter which is most closely related to cocaine effects, especially reinforcing effects. The implications of the dopamine hypothesis is that binding of cocaine at the dopamine transporter blocks transport of dopamine from the synapse, thus making dopamine's effect felt at the postsynaptic receptor. Some cocaine analogues have been investigated for use in cocaine abuse treatment; however, these compounds also tend to have effects similar to those of cocaine. More recently certain tropane derivatives have found promise because they do not show a cocaine-like profile.⁸ In this significant research effort there is need for sound models of the relationship between molecular structure and the transporter binding affinity. Such models can assist researchers to understand the structure basis of binding as well as provide a basis for development and evaluation of new compounds.

In this paper we develop a model for dopamine transporter binding and explore an approach to validation of the model for a data set consisting of 25 observations. Data for binding of tropane derivatives were obtained from Carroll et al.⁹ and serve as the basis both for modeling and for exploring validation. Carroll et al. used CoMFA⁹ as well as the classical Hansch approach⁹ to develop models. In two cases Carroll used 12 of the 25 compounds to create the CoMFA model and then to predict the remaining 13. The occurrence of some large residuals seems to indicate that the training set may

be small for these purposes although the overall statistics were reasonable.

E-State Descriptors. An important objective of modeling is to obtain useful information about the structure features which influence the property being modeled. For this present case we use the molecular structure descriptors known as electrotopological state indices.^{10–18} The E-state indices have been used to develop models for many activities and properties in both their atom-level^{10–13} and atom-type forms.^{10,14} E-state QSAR models yield structure information which reveals structure features significantly related to activity. Further, the more recent development of hydrogen E-state values (and hydrogen atom-type E-state indices¹⁰) has extended the capability of the E-state as a powerful set of structure descriptors. Several studies have investigated QSAR models of binding.^{10–12}

E-state indices have been defined and used in many QSAR and related studies.^{10–18} A complete development is not necessary here. In this topological approach to structure representation, information is developed for each atom (such as $>N-$, or $=O$, or $-Cl$) and each hydride group (such as $-CH_3$, $-NH_2$, or $-OH$) in the molecule. For simplicity both atoms and hydride groups are often called "atoms". The E-state index for atom i in a molecule is computed as follows:

$$S_i = I_i + \sum_j \Delta I_{ij} \quad (\text{sum over all other atoms } j) \quad (1)$$

The perturbation term is as follows:

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

in which r_{ij} is the number of atoms in the shortest path between atoms i and j . The E-state index is constituted from the atom intrinsic state (I_i) plus perturbations (ΔI_{ij}) by all other atoms in the molecule. In this manner each atom's E-state value contains electronic and topological structure information from all other atoms within the structure.¹⁰ The atoms closest to a given atom have the greatest influence on its E-state S value. Influence diminishes for atoms separated

* To whom correspondence should be addressed. Phone: (617) 745-3550; Fax: (617) 745-3905. E-mail: halll@enc.edu.

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 hydrogen atom-level E-state indices. For a data set such as the current dopamine transporter binders, there is a common skeleton among the whole data set. The E-state values for these common skeletal atoms can be used directly as variables in seeking a QSAR model, likewise for the corresponding hydrogen atom-level E-state indices.

For all data sets, including those with a common skeletal core and those with a heterogeneous group of molecules, the atom-type E-state indices provide much useful information. Each atom (or hydride group) in the molecule is classified into an atom type. The atom-type E-state index is the sum of the individual atom level E-state values for a particular atom type.¹⁴ The atom-type descriptors combine three important aspects of structure information: (1) electron accessibility at the atom, (2) presence/absence of the atom, and (3) count of the atom. Hydrogen atom-type E-state descriptors encode very similar information except that accessibility refers to proton accessibility.

In the present dopamine transporter binding data set, the tropane derivatives possess 19 atom sites in common for all molecules. The 19 atom-level E-state and hydrogen E-state indices can be used in model development along with atom-type descriptors in addition to molecular connectivity χ indices and κ shape indices.¹⁹

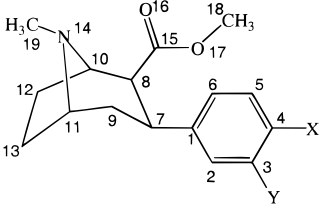
Model Validation. An important aspect of QSAR modeling is the development of means for validation of the model. Good statistical criteria for a fit to the data set are not a guarantee that the model can make accurate predictions for compounds outside the data set. For some time the leave-one-out (LOO) press statistic has been used as a means of demonstrating predictive capability. Alternatively, one may set aside a randomly selected part of the data (validation or test set) which is not used in any way to develop the model. Typically the validation set may consist of 10% of the whole data set. After the model is created, it is used to predict the validation set. The reported statistics are expected to indicate predictive ability. However, it may well be that different parts of the data behave differently with respect to prediction. A single validation set may not clearly indicate predictive ability. Another approach to indicate that the model is significantly better than a random model is to randomize the activity data and repeat the regression. We have performed that analysis with 10 random activity sets.

Creating a validation set is particularly difficult for small data sets. It may be that 90% of the data is too small for adequate modeling and that the 10% validation set may not be large enough for satisfactory validation. Studies of large data sets do not suffer from this problem. Setting a specific numerical limit for the term "large" may not be possible nor necessary. For the present study we consider 25 compounds to be a small set. This validation approach may be called a leave-group-out method (LGO).

METHODS

Data Entry. The binding data and molecular structures were taken from Carroll et al.⁹ and are given in Table 1 along with Carroll's molecule designations under the heading "symbol". The 25 compounds were entered as structure

Table 1. Observed, Calculated, and Residual Binding Data for Phenyltropane Derivatives Which Bind to the Dopamine Transporter



obsd	symbol ^a	X	Y	pIC ₅₀	calcd ^b	res ^c	press ^d
1	2a	H	H	7.64	7.56	0.08	-0.11
2	2b	-F	H	7.86	8.41	-0.55	0.62
3	2c	-Cl	H	8.95	8.64	0.31	-0.34
4	2d	-Br	H	8.77	8.72	0.05	-0.05
5	2e	-NO ₂	H	7.88	8.16	-0.28	0.33
6	2f	-NH ₂	H	8.01	8.18	-0.17	0.19
7	2g	CH ₃ CONH ⁻	H	7.19	7.19	0.00	0.00
8	2h	C ₂ H ₅ CONH ⁻	H	6.92	6.78	0.14	-0.18
9	2i	C ₂ H ₅ OCONH ⁻	H	6.50	6.52	-0.02	0.03
10	2j	-CH ₃	H	8.77	8.25	0.52	-0.57
11	2k	-I	H	8.90	8.76	0.14	-0.15
12	2l	-CF ₃	H	7.89	7.91	-0.02	0.02
13	2m	CH ₃ O ⁻	H	8.04	7.89	0.15	-0.17
14	2n	-N ₃	H	8.67	8.29	0.38	-0.44
15	2o	C ₂ H ₅ ⁻	H	7.26	7.90	-0.64	0.71
16	2p	-OH	H	7.92	7.71	0.21	-0.29
17	2q	-Sn(CH ₃) ₃	H	6.84	6.96	-0.12	0.43
18	2r	H	-F	7.79	7.65	0.14	-0.21
19	2s	H	-Cl	8.20	7.94	0.26	-0.35
20	2t	H	-I	7.58	8.09	-0.51	0.68
21	2u	-NH ₂	-I	8.87	8.81	0.06	-0.08
22	2v	-NH ₂	-Br	8.41	8.77	-0.36	0.49
23	2w	-Cl	-Cl	9.10	9.20	-0.10	0.11
24	2x	-F	-CH ₃	8.53	8.56	-0.03	0.03
25	2y	-Cl	-CH ₃	9.09	8.74	0.36	-0.42

^a Symbol given for the compound in ref 9. ^b calcd = pIC₅₀ value calculated from eq 3. ^c res = pIC₅₀ - calcd. ^d Predicted residual based on the leave-one-out method.

drawings with ChemDraw²⁰ and structure data saved as MDL mol files. The atoms in all molecules were numbered as shown in Table 1. All structure indices used in this investigation were computed from Molconn-Z, ver. 3.50.²¹ 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 all 19 atom-level E-state and 14 (nonzero) hydrogen E-state indices, 20 nonzero atom-type E-state and hydrogen E-state indices, and molecular connectivity χ indices and κ shape indices, which have nonzero variance. 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 kept is primarily experience-based. Preference is given to variables thought to be more interpretable in terms of molecular structure. For example, the atom-type E-state index ST(-CH₃) is correlated with the first-order χ valence index $^1\chi^v$; we selected the E-state index because it is more easily interpretable as a group index. (See the interpretation in a later section.) Since the tropane ring is unvaried in this data set, there were many high intercorrelations in the data matrix, especially among the χ indices and the atom-level E-state indices for atoms remote from the positions of substitution. A commentary on intercorrelations as a function of substitution position is given in the E-state book.¹⁰ Twenty-six

variables remained for statistical analysis in model development.

Statistical Analysis. The data matrix was submitted for statistical analysis using the SAS system.²² The RSQUARE selection method in proc REG was used to examine every QSAR model from one to four variables, listing the top 10 most statistically significant. RSQUARE is not a stepwise procedure; all possible sets of variables are considered, and those with the largest F values are listed. There are two four-variable models with equivalent statistics ($r^2 = 0.84$, $s = 0.32$, $F = 27$). Both models contain the hydrogen atom-type E-state variables $HS^T(\text{other})$ and $HS^T(\text{HBd})$ and the hydrogen atom-level descriptor $Hs(C4)$. One of the two models contains the atom-type E-state descriptor $S^T(-CH_3)$, and the other contains the molecular connectivity χ chain-six-valence descriptor, ${}^6\chi_{CH}^v$. We selected the model with $S^T(-CH_3)$ because that variable is more simply interpreted. A full statistical treatment was done with SAS proc REG based on the four-variable model. The QSAR equation and accompanying statistics are given as eq 3. The observed, calculated, and residual pIC_{50} values are given in Table 1.

Validation Study. To obtain information on the reliability of prediction, a validation (test) data set was randomly selected as 20% (five compounds) of the whole data set. Using the four variables in eq 3, new coefficients were obtained by regression on the remaining 20 observations. Then the pIC_{50} values for the compounds in the validation set were predicted. A second validation set was randomly selected with the requirement that none of the first validation set was included. This selection process was continued for a total of five times so that each compound was selected exactly once. The 25 predicted pIC_{50} values constitute group 1 as shown in Table 2. This validation process was repeated two more times, giving rise to the remaining two groups in Table 2. On the basis of these three validation groups, a consensus set of predictions was obtained as the average of the three groups. The average (av in Table 2) and standard deviation of the three validation groups are also recorded in Table 2. Finally, a residual value is obtained for each compound as $pIC_{50} - \text{av}$. This residual is recorded as the last column in Table 2. All the random selections were accomplished in EXCEL along with computation of average and standard deviation. For the randomization analysis we used EXCEL to randomize the binding data (independent variable) and repeat the regression analysis. This randomization was performed 10 times.

RESULTS AND DISCUSSION

E-State QSAR Model. The model based on the selected four variables yielded statistical information as follows:

$$pIC_{50} = -0.589(\pm 0.086)HS^T(\text{other}) - 0.327(\pm 0.080)HS^T(\text{HBd}) - 0.568(\pm 0.15)Hs(C4) - 0.117(\pm 0.046)S^T(-CH_3) - 16.342(\pm 1.02) \quad (3)$$

$$r^2 = 0.84, s = 0.32, F = 27, n = 25$$

$$r^2_{\text{press}} = 0.77, s_{\text{press}} = 0.39$$

The quantities in parentheses are the standard deviations of the coefficients. A plot of the calculated pIC_{50} versus

Table 2. Summary of pIC_{50} Values Predicted by Leaving out 20% of the Data Repeatedly

obsd	pIC_{50}	predicted values ^a			av ^b	std dev ^c	res ^d
		group 1	group 2	group 3			
1	7.64	7.57	7.75	7.72	7.68	0.096	-0.04
2	7.86	8.46	8.46	8.48	8.47	0.012	-0.61
3	8.95	8.55	8.62	8.58	8.58	0.035	0.37
4	8.77	8.72	8.76	8.77	8.75	0.026	0.02
5	7.88	8.19	8.14	8.12	8.15	0.036	-0.27
6	8.01	8.14	8.14	8.20	8.16	0.035	-0.15
7	7.19	7.31	7.16	7.12	7.20	0.100	-0.01
8	6.92	6.71	6.77	6.73	6.74	0.031	0.18
9	6.50	6.62	6.52	6.47	6.54	0.076	-0.04
10	8.77	8.22	8.18	8.14	8.18	0.040	0.59
11	8.90	8.69	8.78	8.71	8.73	0.047	0.17
12	7.89	7.69	7.88	7.89	7.82	0.113	0.07
13	8.04	7.90	7.94	7.95	7.93	0.026	0.11
14	8.67	8.14	8.21	8.24	8.20	0.051	0.47
15	7.26	8.16	7.96	7.96	8.03	0.115	-0.77
16	7.92	7.64	7.66	7.62	7.64	0.020	0.28
17	6.84	7.83	7.13	7.30	7.42	0.365	-0.58
18	7.79	7.60	7.39	7.61	7.53	0.124	0.26
19	8.20	8.07	7.70	7.87	7.88	0.185	0.32
20	7.58	8.22	8.27	8.29	8.26	0.036	-0.68
21	8.87	8.70	8.90	8.78	8.79	0.101	0.08
22	8.41	8.93	8.86	8.83	8.87	0.051	-0.46
23	9.10	9.20	9.13	9.25	9.19	0.060	-0.09
24	8.53	8.55	8.48	8.59	8.54	0.056	-0.01
25	9.09	8.62	8.71	8.59	8.64	0.062	0.45

^a Values predicted when 20% of the data set is left out repeatedly so that each compound is predicted exactly once. See the text. ^b Average of the three values obtained for each compound when 20% of the data is set left out repeatedly. See the text. ^c Standard deviation of the three group values. ^d $pIC_{50} - \text{av}$.

observed pIC_{50} is given in Figure 1. An examination of the plot of residuals versus observed pIC_{50} (not shown) revealed no trends and appears randomly distributed. The variables in the model may be examined for the structure information encoded by each as follows.

$HS^T(\text{other})$ is the sum of the hydrogen atom-level E-state indices for all nonpolar hydrogen atoms in the molecule.¹⁴ $HS^T(\text{other})$ is the statistically most significant variable in the model, contributing 90.4% on the average to the calculated pIC_{50} . As a single variable, the correlation coefficient between pIC_{50} and $HS^T(\text{other})$ is $r^2 = 0.65$. Because of the negative coefficient on $HS^T(\text{other})$, smaller values are related to larger pIC_{50} values, suggesting that new candidates for greater binding should have substituents at X and Y which decrease the nonpolar character of the molecule. Variation in $HS^T(\text{other})$ and all variables in the model arises solely from the variation in the substituents X and Y.

It is possible to partition $HS^T(\text{other})$ among three atom-type hydrogen E-state indices: $HS^T(\text{other}) = HS^T(\text{Csats}) + HS^T(\text{Csatu}) + HS^T(\text{arom})$. These descriptors represent hydrogen atoms in three different environments: $HS^T(\text{Csats})$, hydrogens on saturated carbon atoms bonded to saturated carbon atoms; $HS^T(\text{Csatu})$, hydrogens on saturated carbon atoms bonded to unsaturated carbon atoms; $HS^T(\text{arom})$, hydrogens on aromatic carbon atoms. When each of these three variables replaces $HS^T(\text{other})$ in the model, the correlation is degraded significantly: $r^2 < 0.67$. This analysis indicates that none of these three regions of the carbon skeleton dominate the $HS^T(\text{other})$ descriptor.

The major significance of $HS^T(\text{other})$ is further indicated by the fact that compounds with the three largest pIC_{50} values

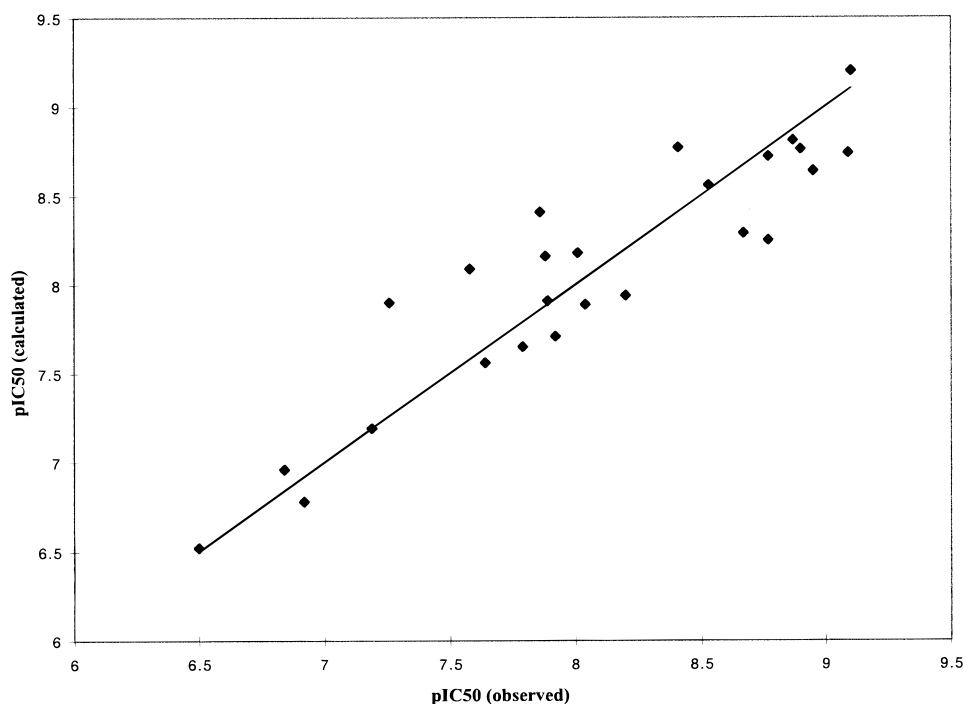


Figure 1. Plot of pIC_{50} calculated from E-State model (equation 3) versus the observed pIC_{50} values (Table 1).

also have the three smallest $HS^T(\text{other})$ values. Also compounds with small $HS^T(\text{other})$ values tend to have large binding values.

The second variable, $HS^T(\text{HBd})$, represents hydrogen bond donation ability.¹⁰ It is the sum of the hydrogen E-state values for groups which act as hydrogen bond donors. In this present data set these groups include $-\text{CONH}$, $-\text{OCONH}$, $-\text{NH}_2$, and $-\text{OH}$. They are present in 7 of the 25 molecules; each possesses only one hydrogen bond donor group present in any one molecule. The $HS^T(\text{HBd})$ variable contributes on average only 2.0% to the calculated binding but ranges from 0% to 9.7%. Thus, for the seven compounds with a hydrogen bond donor, the descriptor $HS^T(\text{HBd})$ contributes about 10% to the calculated binding (others being zero), indicating a significant but secondary feature. Further, because of its negative coefficient, the smaller the value of $HS^T(\text{HBd})$, the greater the value of calculated pIC_{50} . Three of the four compounds with the largest $HS^T(\text{HBd})$ values have the smallest pIC_{50} values. Further, the descriptor for hydrogen bond acceptors, $HS^T(\text{HBa})$, was not selected in the modeling process, although it was available in the list of descriptors from Molconn-Z. A model based only on $HS^T(\text{other})$ and $HS^T(\text{HBd})$, the best two-variable model, yields $r^2 = 0.72$.

The third variable, $Hs(\text{C4})$, is the hydrogen E-state value for a hydrogen atom at position 4 on the phenyl ring (position of X substituents). $Hs(\text{C4})$ is zero when there is a substituent at X but possesses a variable value otherwise. Its value is influenced by both the presence/absence of X substituents and the electronic and topological character of the Y substituents. On average $Hs(\text{C4})$ contributes 1.3% to the calculated binding but ranges from 0% to 8.8%. It should be noted that $Hs(\text{C4})$ is not the descriptor for either substituent X or substituent Y but does encode electronic and topological influence of Y on position 4 in the phenyl ring. For compounds not substituted at X, the $Hs(\text{C4})$ descriptor contributes about 9% to the calculated value of

pIC_{50} ; otherwise $Hs(\text{C4})$ is zero. Because of the negative coefficient on $Hs(\text{C4})$, smaller $Hs(\text{C4})$ values relate to greater binding. A model based on $HS^T(\text{other})$, $HS^T(\text{HBd})$, and $Hs(\text{C4})$, the best three-variable model, yields $r^2 = 0.79$.

The fourth variable is $S^T(-\text{CH}_3)$, the atom type E-state index for $-\text{CH}_3$ groups in the molecule. All molecules possess at least two methyl groups, located on the ester group, $-\text{CO}_2\text{CH}_3$, and the tertiary amine, $>\text{NCH}_3$. Nine other molecules possess one or more additional methyl groups; three molecules have a methyl group directly on the phenyl ring. In this data set the $S^T(-\text{CH}_3)$ variable contributes on average 6.3% to the calculated binding and ranges from 4.7% to 13.7%. Because of the negative coefficient on the $S^T(-\text{CH}_3)$, smaller values of $S^T(-\text{CH}_3)$ are associated with greater binding. Thus, methyl groups bonded to polar groups tend to increase binding.

The $S^T(-\text{CH}_3)$ descriptor may be partitioned between contributions from methyl groups directly on the phenyl ring and those which are not, that is, which are part of a substituent (and on the ester group or tertiary amine). When values are used in the model for methyl groups just in substituents (and on the ester group and tertiary amine), the correlation coefficient ($r^2 = 0.86$) is significantly greater than for those which are only directly on the phenyl ring ($r^2 = 0.80$). Further, the coefficient of the methyl term for those in substituents is about 3 times greater than for those directly on the phenyl ring. This analysis indicates the greater significance of methyl groups on the substituents. Although there is no simple, clear picture of methyl groups in substituents (and ester group and tertiary amine), their presence does contribute to the hydrophobicity of the substituted phenyltropanes.

In summary, the E-State model indicates that new candidate structures for increased binding should incorporate polar substituents, possess weak hydrogen-bonding groups, and have substituents at Y which are polar or electronegative.

Validation Study. To assess the potential for the predictive ability of the E-state model developed here (eq 3), we considered the potential difficulties created by selecting too small a training set or too small a validation (test) set. When the training set is too small (for a small data set), it may be that the relationship between activity and structure is not discernible by statistical methods. A satisfactory model may not be obtained, or a model may not contain all the relevant structure information of the whole data set. On the other hand, if the test set is too small, the predictions may not give a reliable picture of predictability. One particular set may fortuitously give a false impression of high reliability or, on the other hand, too negative a picture of low reliability. Since a large training set implies a small test set and a large test set implies a small training set, these two alternatives for a single train/test set approach seem less than optimal.

To deal with these problems for small data sets, we adopted an alternative approach to selecting one training and only one test set or the common leave-one-out LOO approach. Instead we propose a leave-group-out LGO scheme in which each observation is deleted and predicted at least three times. Each deleted observation is in a set of 20% of the data; all deletion (test) sets are unique. To obtain a group of predicted values consisting of each compound deleted once, five sets of five compounds each were selected randomly. This whole process was repeated three times. In this manner three predictions were obtained for each compound in the whole data set. These predictions are not dependent upon a single selection of one part of the data. This current leave-group-out LGO approach is similar to an approach developed earlier for toxicity of phenols to fish²³ and in our work on the antimicrobial activity of phenyl propyl ethers in which 15% of the data were left out 10 times randomly.²⁴

To assess the predictive quality of the model using this process, the mean absolute error of the average predictions (av) was computed: MAE = 0.28. The corresponding root-mean-squared error is found to be 0.36. An examination of these values along with the standard error of the calculated average indicates reasonable predictive quality for the E-State model. There were no large residuals, the largest being 0.77, which is 2.4 times the standard deviation of the regression (eq 3) (Table 2). The correlation between the observed pIC₅₀ and the average of the predictions is $r^2 = 0.75$. For small data sets we suggest consideration of this LGO approach to determination of predictive ability or validation.

To predict binding for new compounds, that is, compounds not in the original data set, we suggest two approaches.

First, one can use the model based on the full data set, eq 3. We note that the press statistic for standard error 0.39 is only slightly larger than for the consensus model, 0.36, suggesting that this equation may be useful for prediction.

Second, one can make several predictions of pIC₅₀ binding values from subsets of the whole data set in the same manner in which validation test sets were predicted. A diminished set consisting of 80% of the data is used as the basis for a regression model with the four variables of eq 3. The candidate structures are predicted from that diminished set. Then, a second randomly selected diminished set of 80% is selected and used to predict the candidate binding values. This process can be repeated several times. In this manner, several predictions are obtained for each candidate. For example, consider a new structure in which X = -F and Y =

-Cl. Necessary structure variables were calculated by Molconn-Z. A first diminished set was obtained by deleting compounds 4, 6, 13, 15, and 23 (randomly selected). The pIC₅₀ value predicted from the 80% set is found to be 9.07. Following the same procedure, another diminished set was obtained by deleting compounds 3, 12, 14, 22, and 24 and predicting, finding the pIC₅₀ value 8.98. Finally, for a third time by deleting the set 1, 2, 4, 20, and 25, the predicted value for pIC₅₀ is 9.04. To represent the pIC₅₀ value for the new compound, the mean and standard deviation of these three values was computed: mean pIC₅₀ = 9.03; standard deviation 0.046. We note that a consistent set of three predictions is obtained in this manner. This mean value is obtained by sampling three independently selected portions of the data set to minimize potential bias of using only one set.

Randomization Study. The independent variable, binding affinity, was randomized using the random number generator in EXCEL. The randomized data replaced the binding data in the data set, and the regression was repeated. This process was carried out 10 times. The largest r^2 and F values found are $r^2 = 0.34$ and $F = 2.5$. The average r^2 was found to be 0.16, and the average F was found to be 1.0. These data correspond to a random statistics, giving credence to the model based on the topological variables used in eq 3. This current work is an extension of our earlier work on randomization.^{25,26}

CONCLUSIONS

For phenyltropane binding to the dopamine transporter, a statistically satisfactory QSAR model is developed with four E-state structure descriptors. Structure information encoded in the descriptors indicates structurally significant features: (1) although a large portion of the phenyltropanes is nonpolar and important to binding, that general nonpolarity must be moderated by the polarity of the substituents; (2) hydrogen bond donating groups in the X position tend to add to binding (no hydrogen-bonding groups were present in the Y position); (3) hydrogen bond donors should be weaker rather than stronger; and (4) methyl groups, especially part of the substituents rather than directly on the phenyl ring, add to binding strength.

An approach to model validation for small data sets is described. Observations are deleted randomly in unique test sets (LGO) of 20% (of the total set) and predicted from the remaining 80% so that each observation is predicted three times. The mean of the three predictions is compared to the observed values. Predictive ability is based on the statistical comparison of the average predictions with the observed values and found to be of good quality for this model. Binding (pIC₅₀) values for new candidate molecules can be predicted from the four-variable model based on the whole data set or from several sets of randomly selected partial sets of the data.

ACKNOWLEDGMENT

LHH wishes to acknowledge the DuPont Chemical Corporation for a grant which provides partial support for this research investigation.

REFERENCES AND NOTES

- (1) Kuhar, M.; Ritz, M. C.; Boja, J. W. The Dopamine Hypothesis of the Reinforcing Properties of Cocaine. *Trends Neurosci.* **1991**, *14*, 299–302.
- (2) Self, D. W.; Nestler, E. Molecular Mechanisms of Drug Reinforcement and Addiction. *Annu. Rev. Neurosci.* **1995**, 463–495.

- (3) Cline, E. J.; Terry, P.; Carroll, R. I.; Kuhar, M. J.; Katz, J. L. Stimulus Generalization from Cocaine Analogues with High In Vitro Affinity for Dopamine Uptake Sites. *Behav. Pharmacol.* **1992**, *3*, 113–116.
- (4) Newman, A. H.; Allen, A. C.; Izenwasser, S.; Katz, J. L. Novel 3- α -(diphenylmethoxy)tropane Profiles. *J. Med. Chem.* **1994**, *37*, 2258–2261.
- (5) Newman, A. H.; Kline, A. H.; Allen, A. C.; Izenwasser, S.; George, C.; Katz, J. L. Novel 4'-Substituted and 4',4''-Disubstituted 3- α -(diphenylmethoxy)tropane Analogs as Potent and Selective Dopamine Uptake Inhibitors. *J. Med. Chem.* **1995**, *38*, 3933–3940.
- (6) Carroll, F. I.; Lewin, A. H.; Boja, J. W.; Kuhar, M. J. Cocaine Receptor: Biochemical Characterization and Structure–Activity Relationships for the Dopamine Transporter. *J. Med. Chem.* **1992**, *35*, 969–981.
- (7) Ritz, M. C.; Lamb, R. J.; Goldberg, S. R.; Kuhar, M. J. Cocaine Receptors on Dopamine Transporters are Related to Self-Administration of Cocaine. *Science* **1987**, *237*, 1219–1223.
- (8) Newman, A. H. Novel Dopamine Transporter Ligands: The State of the Art. *Med. Chem. Rev.* **1998**, *8*, 1–11.
- (9) Carroll, F. I.; Mascarella, S. W.; Kuzemko, M. A.; Gao, Y.; Abraham, P.; Lewin, A. H.; Boja, J. W.; Kuhar, M. J. Synthesis, Ligand Binding, and QSAR (CoMFA and Classical) Study of 3 β -(3-Substituted phenyl)-, 3 β -(4'-Substituted phenyl)-, and 3 β -(3',4'-Disubstituted phenyl)tropane-2 β -carboxylic Acid Methyl Esters. *J. Med. Chem.* **1994**, *37*, 2865–2873.
- (10) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrototopological State*; Lemont, B., Ed.; Academic Press: New York, 1999.
- (11) Kier, L. B.; Hall, L. H. The Electrototopological 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.
- (12) Kier, L. B.; Hall, L. H. Inhibition of Salicylamide Binding: A Electrototopological State Analysis. *Med. Chem. Res.* **1992**, *2*, 497–502.
- (13) Hall, L. H.; Mohny, B. K.; Kier, L. B. Comparison of Electrototopological State Indexes with Molecular Orbital Parameters: Inhibition of MAO by Hydrazides. *Quant. Struct.-Act. Relat.* **1993**, *12*, 44–48.
- (14) Hall, L. H.; Kier, L. B. Electrototopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
- (15) Gough, J.; Hall, L. H. QSAR Models of the Antileukemic Potency of Carboquinones: Electrototopological State and Chi Indices. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 356–361.
- (16) Gough, J.; Hall, L. H. Modeling the Toxicity of Amide Herbicides using the Electrototopological State. *Environ. Toxicol. Chem.* **1999**, *18*, 1069–1075.
- (17) Kier, L. B.; Hall, L. H. The E-State in Database Analysis: The PCBs as an Example. *Il Farmico* **1999**, *54*, 346–353.
- (18) Database Organization and Similarity Searching with E-State Indices. *Symposium on Computer Methods for Structure Representation*, Kier, L. B., Hall, L. H. Kluwer Publishing Co.: Amsterdam, The Netherlands, in press.
- (19) 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.
- (20) ChemDraw, ver. 4.5, CambridgeSoft, Cambridge, MA 02139.
- (21) Molconn-Z, ver. 3.50, is available from Hall Associates Consulting, 2 Davis St., Quincy, MA 02170, from EduSoft, LC, P.O. Box 1811, Ashland, VA 23005, and from SciVision, Inc. 200 Wheeler Rd., Burlington, MA 01803.
- (22) SAS, ver. 6.12, SAS Institute, Cary, NC 27513.
- (23) Hall, L. H.; Kier, L. B. Molecular Connectivity of Phenols and Their Toxicity to Fish. *Bull. Environ. Contam. Toxicol.* **1984**, *32*, 354–362.
- (24) Hall, L. H.; Kier, L. B. Molecular Connectivity and Substructure Analysis. *J. Pharm. Sci.* **1978**, *67*, 1743–1747.
- (25) Kier, L. B.; Hall, L. H. Molecular Connectivity Study of Muscarinic Receptor Activity of Acetylcholine Antagonists. *J. Pharm. Sci.* **1978**, *67*, 1408–1412.
- (26) Kier, L. B.; Hall, L. H. Structure–Activity Studies on Hallucinogenic Amphetamines Using Molecular Connectivity. *J. Med. Chem.* **1977**, *20*, 1631–1636.

CI000023X