# E-State Modeling of Corticosteroids Binding AffinityValidation of Model for Small Data Set

Hlaing Hlaing Maw and Lowell H. Hall*

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

Received April 15, 2001

Data for 31 steroids binding to the corticosteroid binding globulin (CBG) were modeled using E-state molecular structure descriptors and a kappa shape index. Both E-state and hydrogen E-state descriptors appear in the model in atom-level and atom-type descriptors. A four-variable model is obtained that is statistically satisfactory: $r^2 = 0.81$, $s = 0.51$; $r^2_{\text{press}} = 0.72$; $s_{\text{press}} = 0.62$. Structure interpretation is given for each variable in the model. A leave-group-out (LGO) 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 for which $r^2_{\text{LOO}} = 0.70$. These collective results support the claim that the E-state model may be useful for prediction of p*K* binding values for new compounds.

## INTRODUCTION

Corticosteroids are important for the regulation of human physiology. Glucocorticoids are involved in the regulation of carbohydrate, lipid, and protein metabolism. Mineralocorticoids influence salt and water metabolism by maintaining proper electrolyte balance. Cortisone is effective in the treatment of rheumatoid arthritis. Prednisone and prednisolone are potent antirheumatic and antiallergenic agents.[1] Researchers have found that chronic anterior uveitis (an inflammation of the uvea in the eye) responds to corticosteroid treatments.[2] Keloids, abnormal concentrations of scar tissue, can be treated by corticosteroids injections.[3] The corticosteroid drug budesonide is the preferred treatment for Crohn's disease (an inflammation of the gastrointestinal tract).[4] Baud et al. have found that the corticosteroid drug betamethasone has fewer side effects when used to prevent many pregnant women from delivery complications of premature birth.[5] Corticosteroid-receptor activation is crucial in determining steroid-mediated effects.

Many researchers have studied corticosteroid binding affinity for the purposes of quantitative structure−activity relationships (QSAR).[6−8] In this significant research effort, there is need for sound models of the relationship between the structure of steroids and the corticosteroids binding globulin (CBG). Such models can assist researchers in understanding the structure basis of binding as well as provide a basis for development of new compounds.

In this paper we develop a model for steroids binding to CBG and make use of an approach to validation of the model for a data set consisting of 31 observations. This approach to model validation was recently applied to a set of 25 tropane derivatives which bind to the dopamine transporter.[9] This steroid data set offers another opportunity to make use of this approach. Steroid data for CBG was obtained from Polanski[10] and serves as the basis both for modeling and for exploring validation.

**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[11−19] which were developed to represent potential for noncovalent intermolecular interaction. The E-state indices have been used to develop models for many activities and properties in both their atom-level[11−14] and atom-type forms.[11,15] 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 indices[11]) has extended the capability of the E-state as a powerful set of structure descriptors. Several studies have investigated QSAR models of binding.[11−13]

E-state indices have been defined and used in many QSAR and related studies.[11−19] Only a brief development is necessary here. In this topological approach to structure representation, information is developed for each atom (such as $=O$, $-F$) and each hydride group (such as $-CH_3$, $-OH$) in the molecule. For simplicity both atoms and hydride groups are often called "atoms". The E-state index, $S_i$, 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 perturbation term, $\Delta I_{ij}$, is defined as follows

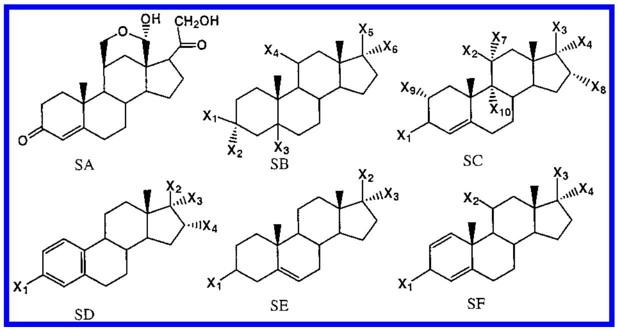
$$\Delta I_{ij} = (I_i - I_j)/r_{ij}^2 \qquad (2)$$

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 ($\Delta I_{ij}$) 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.[11] The

* Corresponding author phone: (617)745-3550; fax: (617)745-3905; e-mail: halll@enc.edu.

10.1021/ci010037i CCC: $20.00
Published on Web 08/03/2001

**Table 1.** Structures of the Corticosteroids Used in the E-State Analysis of Corticosteroid Binding Globulin Data[a]



| id | structure | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SA | | | | | | | | | | |
| 2 | SB | OH | H | H[b] | H | OH | H | | | | |
| 3 | SE | OH | OH | H | | | | | | | |
| 4 | SC | =O | H | =O | | | | H | H | H | H |
| 5 | SB | H | OH | H[b] | H | =O | | | | | |
| 6 | SC | =O | OH | $COCH_2OH$ | H | | | H | H | H | H |
| 7 | SC | =O | OH | $COCH_2OH$ | OH | | | H | H | H | H |
| 8 | SC | =O | =O | $COCH_2OH$ | OH | | | | H | H | H |
| 9 | SE | OH | =O | | | | | | | | |
| 10 | SC | =O | H | $COCH_2OH$ | H | | | H | H | H | H |
| 11 | SC | =O | H | $COCH_2OH$ | OH | | | H | H | H | H |
| 12 | SB | =O | | H[b] | H | OH | H | | | | |
| 13 | SD | OH | OH | H | H | | | | | | |
| 14 | SD | OH | OH | H | OH | | | | | | |
| 15 | SD | OH | =O | | | | | | | | |
| 16 | SB | H | OH | H[c] | H | =O | | | | | |
| 17 | SE | OH | COMe | H | | | | | | | |
| 18 | SE | OH | COMe | OH | | | | | | | |
| 19 | SC | =O | H | COMe | H | | | H | H | H | H |
| 20 | SC | =O | H | COMe | OH | | | H | H | H | H |
| 21 | SC | =O[d] | H | OH | H | | | H | H | H | H |
| 22 | SF | =O | OH | $COCH_2OH$ | OH | | | | | | |
| 23 | SC | =O | OH | $COCH_2OCOMe$ | OH | | | H | H | H | H |
| 24 | SC | =O | =O | COMe | H | | | | H | H | H |
| 25 | SC | =O | H | $COCH_2OH$ | H | | | OH | H | H | H |
| 26 | SC[e] | =O | H | OH | H | | | H | H | H | H |
| 27 | SC | =O | H | COMe | OH | | | H | OH | H | H |
| 28 | SC | =O | H | COMe | H | | | H | Me | H | H |
| 29 | SC[e] | =O | H | COMe | H | | | H | H | H | H |
| 30 | SC | =O | OH | $COCH_2OH$ | OH | | | H | H | Me | H |
| 31 | SC | =O | OH | $COCH_2OH$ | OH | | | H | H | Me | F |

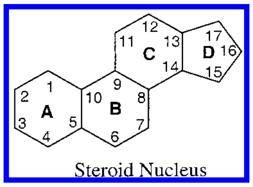
[a] Structures according to refs 26−28. [b] Of the 5-α steroid series. [c] Of the 5-*β* steroid series. [d] Assumed to be =O (testosterone) as indicated by ref 28 and not as −OH in table and further publications [compare ref 28 for mistakes by authors in previous publications]. [e] H (hydrogen) instead of Me at $C_{10}$ steroid skeleton.

atoms closest to a given atom have the greatest influence on its E-state *S* 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 for hydrogen atoms provides the basis for the hydrogen atom-level and hydrogen atom-type E-state indices. The E-state indices represent the potential for noncovalent intermolecular interaction for each atom because of the encoding of electron accessibility.

For the current data set of corticosteroids binders, there are six skeletal types among the whole data set, as shown in the steroid nucleus below (SA, SB, SC, SD, SE, SF; also see Table 1). The (atom-level) E-state and hydrogen E-state values for the common skeletal atoms can be used directly as variables in seeking a QSAR model as long as these atoms are numbered in the same way in all molecules in the data set.

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 as 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.[15] The atom-type descriptors combine three important aspects of structure information: (1) collective electron accessibility for all the atoms of a given type in a molecule, (2) presence/absence of the atom, and (3) count of the atom in the molecule. Hydrogen atom-type E-state descriptors encode very similar information except that accessibility refers to hydrogen accessibility.

In the present data set, the steroid nucleus possesses 17 common atom sites for all molecules. The atom-level E-state and hydrogen E-state indices for these 17 atoms can be used in model development along with atom-type E-state descriptors in addition to molecular connectivity chi indices and kappa shape indices.[20]



**Model Validation.** An important aspect of QSAR modeling is the development of means for validation of the model. Good direct statistical criteria for fit to the data set are not a guarantee that the model can make accurate predictions for compounds outside the data set. 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 validation, that may indicate whether 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 randomized 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 these problems. Setting a specific numerical limit for the terms "large" or "small" may not be possible nor necessary. For the present study we consider 31 compounds to be a small set. The validation approach used in this study may be called a leave-group-out method.

## METHODS

**Data Entry.** The binding data and molecular structures were taken from Polanski[10] and are given in Table 2 along with Polanski's molecule designations under the heading "symbol". The 31 compounds were entered as structure drawings with ChemDraw[21] and structure data saved as MDL mol files. The atoms-in-common in all molecules were numbered in the same way as shown in the steroid nucleus above. All structure indices used in this investigation were computed from Molconn-Z, ver 3.50.[22] Structure input was validated by visual inspection of the ChemDraw drawings[1] as well as the structure analysis provided by Molconn-Z output.

For QSAR analysis we selected all 17 common atom-level E-state and 17 hydrogen E-state indices, 21 nonzero atom-type E-state, and hydrogen atom-type E-state indices, along with molecular connectivity chi indices and kappa 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 easily interpretable in terms of molecular structure. For example, ${}^2\chi^v$ was selected from among the chi valence path indices. All the chi valence path indices from order 0 to 6 are intercorrelated; only one may be used. The second-order valence index was selected because it is the simplest of the set which yet includes the most information on skeletal branching.

**Table 2.** Observed, Calculated, and Residual Binding Data for Corticosteroids Which Bind to the Corticosteroid Binding Globulin

| obsd | symbol[a] | p*K* | calcd[b] | res[c] | press[d] |
|---|---|---|---|---|---|
| 1 | SA | 6.279 | 6.892 | −0.613 | −0.652 |
| 2 | SB | 5.000 | 5.590 | −0.590 | −0.715 |
| 3 | SE | 5.000 | 4.763 | 0.237 | 0.312 |
| 4 | SC | 5.763 | 6.303 | −0.540 | −0.656 |
| 5 | SB | 5.613 | 5.502 | 0.111 | 0.132 |
| 6 | SC | 7.881 | 7.059 | 0.822 | 0.929 |
| 7 | SC | 7.881 | 7.177 | 0.704 | 0.783 |
| 8 | SC | 6.892 | 7.001 | −0.109 | −0.123 |
| 9 | SE | 5.000 | 4.640 | 0.360 | 0.510 |
| 10 | SC | 7.653 | 7.181 | 0.472 | 0.527 |
| 11 | SC | 7.881 | 7.278 | 0.603 | 0.676 |
| 12 | SB | 5.919 | 5.795 | 0.124 | 0.139 |
| 13 | SD | 5.000 | 4.754 | 0.246 | 0.301 |
| 14 | SD | 5.000 | 5.152 | −0.152 | −0.178 |
| 15 | SD | 5.000 | 4.634 | 0.366 | 0.440 |
| 16 | SB | 5.225 | 5.502 | −0.277 | −0.330 |
| 17 | SE | 5.225 | 5.426 | −0.201 | −0.282 |
| 18 | SE | 5.000 | 5.584 | −0.584 | −0.788 |
| 19 | SC | 7.380 | 7.089 | 0.291 | 0.330 |
| 20 | SC | 7.740 | 7.222 | 0.518 | 0.617 |
| 21 | SC | 6.724 | 6.442 | 0.282 | 0.322 |
| 22 | SF | 7.512 | 6.735 | 0.777 | 0.870 |
| 23 | SC | 7.553 | 8.046 | −0.493 | −0.939 |
| 24 | SC | 6.779 | 7.011 | −0.232 | −0.254 |
| 25 | SC | 7.200 | 7.059 | 0.141 | 0.159 |
| 26 | SC | 6.144 | 6.551 | −0.407 | −0.463 |
| 27 | SC | 6.247 | 7.206 | −0.959 | −1.192 |
| 28 | SC | 7.120 | 7.325 | −0.205 | −0.238 |
| 29 | SC | 6.817 | 7.224 | −0.407 | −0.469 |
| 30 | SC | 7.688 | 7.377 | 0.311 | 0.349 |
| 31 | SC | 5.797 | 6.391 | −0.594 | −0.705 |

[a] Symbol given for compound in ref 10. [b] calcd = p*K* value calculated from eq 3. [c] res = p*K* − calc. [d] Predicted residual based on leave-one-out method.

Since the steroid nucleus is unvaried in this data set, there were many high intercorrelations in the data matrix, especially among the chi indices, the atom-level E-state indices, and hydrogen E-state indices. A commentary on intercorrelations as a function of substitution position is given in the E-state book.[11] Twenty-six variables remained for statistical analysis in model development.

**Statistical Analysis.** The data matrix was submitted for statistical analysis using the SAS system.[23] 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.

The best model consists of the kappa alpha shape index ${}^3\kappa_\alpha$, the atom-type E-state descriptor $S^T(=C<)$, the atom level E-state index $S(C_{17})$, and the hydrogen atom-level descriptor $Hs(C_6)$. 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 p*K* values are given in Table 2.

**Table 3.** Summary of p$K$ Values Predicted by LGO Method, Leaving out 20% of Data Repeatedly

| | | predicted values[a] | | | | | |
|---|---|---|---|---|---|---|---|
| obsd | p$K$ | group 1 | group 2 | group 3 | av[b] | SD[c] | res[d] |
| 1 | 6.279 | 7.06 | 6.93 | 7.10 | 7.03 | 0.089 | −0.75 |
| 2 | 5.000 | 5.76 | 5.71 | 5.70 | 5.72 | 0.032 | −0.72 |
| 3 | 5.000 | 4.52 | 4.66 | 4.73 | 4.64 | 0.107 | 0.36 |
| 4 | 5.763 | 6.46 | 6.50 | 6.42 | 6.46 | 0.040 | −0.70 |
| 5 | 5.613 | 5.48 | 5.52 | 5.43 | 5.48 | 0.045 | 0.14 |
| 6 | 7.881 | 7.00 | 6.84 | 6.95 | 6.93 | 0.082 | 0.95 |
| 7 | 7.881 | 7.05 | 7.26 | 7.23 | 7.18 | 0.114 | 0.70 |
| 8 | 6.892 | 7.46 | 7.22 | 7.06 | 7.25 | 0.201 | −0.35 |
| 9 | 5.000 | 4.48 | 4.75 | 4.45 | 4.56 | 0.165 | 0.44 |
| 10 | 7.653 | 7.12 | 6.97 | 7.03 | 7.04 | 0.075 | 0.61 |
| 11 | 7.881 | 7.28 | 7.21 | 7.21 | 7.23 | 0.040 | 0.65 |
| 12 | 5.919 | 5.60 | 5.74 | 5.90 | 5.75 | 0.150 | 0.17 |
| 13 | 5.000 | 4.44 | 4.76 | 4.82 | 4.67 | 0.204 | 0.33 |
| 14 | 5.000 | 5.27 | 5.02 | 5.22 | 5.17 | 0.132 | −0.17 |
| 15 | 5.000 | 4.30 | 4.60 | 4.48 | 4.46 | 0.151 | 0.54 |
| 16 | 5.225 | 5.53 | 5.50 | 5.61 | 5.55 | 0.057 | −0.32 |
| 17 | 5.225 | 5.43 | 5.32 | 5.38 | 5.38 | 0.055 | −0.15 |
| 18 | 5.000 | 6.12 | 5.77 | 5.80 | 5.90 | 0.194 | −0.90 |
| 19 | 7.380 | 7.00 | 7.22 | 7.00 | 7.07 | 0.127 | 0.31 |
| 20 | 7.740 | 7.02 | 7.07 | 7.17 | 7.09 | 0.076 | 0.65 |
| 21 | 6.724 | 6.37 | 6.61 | 6.19 | 6.39 | 0.211 | 0.33 |
| 22 | 7.512 | 6.63 | 6.56 | 6.77 | 6.65 | 0.107 | 0.86 |
| 23 | 7.553 | 8.80 | 8.42 | 8.83 | 8.68 | 0.229 | −1.13 |
| 24 | 6.779 | 6.91 | 7.11 | 6.93 | 6.98 | 0.110 | −0.20 |
| 25 | 7.200 | 7.01 | 6.84 | 6.88 | 6.91 | 0.089 | 0.29 |
| 26 | 6.144 | 6.63 | 6.72 | 6.62 | 6.66 | 0.055 | −0.51 |
| 27 | 6.247 | 7.80 | 7.41 | 7.38 | 7.53 | 0.234 | −1.28 |
| 28 | 7.120 | 7.29 | 7.41 | 7.48 | 7.39 | 0.096 | −0.27 |
| 29 | 6.817 | 7.28 | 7.28 | 7.32 | 7.29 | 0.023 | −0.48 |
| 30 | 7.688 | 7.35 | 7.63 | 7.33 | 7.44 | 0.168 | 0.25 |
| 31 | 5.797 | 6.81 | 6.51 | 6.47 | 6.60 | 0.186 | −0.80 |

[a] Values predicted when 20% of data set left out repeatedly so that each compound is predicted exactly once. See text. [b] Average of the three values obtained for each compound when 20% of data set is left out repeatedly. See text. [c] Standard deviation of the three group values. [d] p$K$ − av.

**Validation Study.** To obtain information on the reliability of prediction, a validation (test) data set was randomly selected as 20% (six compounds) of the whole data set. Using the four variables in eq 3, new coefficients were obtained by regression on the remaining 25 observations. Then the p$K$ 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 last validation set contained seven observations. The 31 predicted p$K$ values constitute group 1 as shown in Table 3. This validation process was repeated two more times, giving rise to the remaining two groups in Table 3. Based on these three validation groups, a consensus set of predictions was obtained as the average of the three groups. The average (average in Table 3) and standard deviation of the three validation groups are also recorded in Table 3. Finally, a residual value is obtained for each compound as p$K$ − average. This residual is recorded as the last column in Table 3. All the random selections were accomplished in EXCEL along with computation of average and standard deviation. This method of validation may be called the leave-group-out (LGO) method in analogy to the leave-one-out (LOO) method.

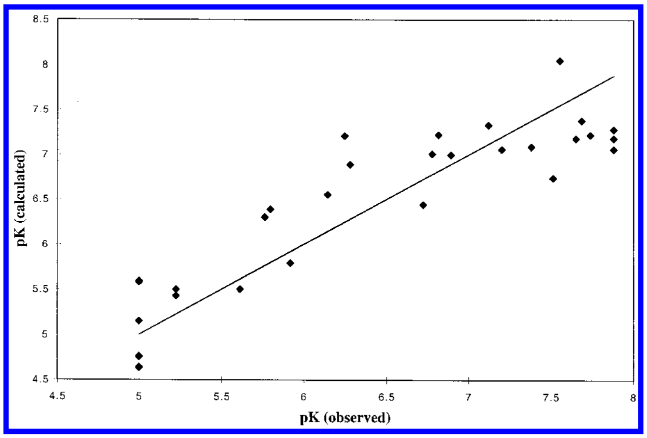


**Figure 1.** Plot of p$K$ calculated from E-state model (eq 3) versus the observed p$K$ values (Table 2).

For the randomization analysis we used EXCEL to randomize the binding data (independent variable, p$K$) and repeat the regression analysis. This randomization was performed 10 times.

## RESULTS AND DISCUSSION

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

$$pK = 2.028(\pm 0.442)\ {}^{3}\kappa_{\alpha} - 3.594(\pm 0.624)\ \mathrm{HS}(\mathrm{C}_6)\ 0.846(\pm 0.148)\ \mathrm{S}^{\mathrm{T}}({=}\mathrm{C}{<}) - 0.613(\pm 0.155)\ \mathrm{S}(\mathrm{C}_{17}) - 3.688(\pm 0.971) \quad (3)$$

$$r^2 = 0.81,\ s = 0.51,\ F = 27,\ n = 31$$

$$r^2_{\mathrm{press}} = 0.72,\ s_{\mathrm{press}} = 0.62$$

The quantities in parentheses are the standard deviations of the coefficients. A plot of the calculated p$K$ versus observed p$K$ is given in Figure 1. An examination of the plot of residuals versus observed p$K$ (not shown) revealed no trends and appears randomly distributed. The press statistics refer to the leave-one-out approach (LOO).

Each variable in the QSAR equation is a structure descriptor representing specific structure information for the steroids in this data set. The variables in the model may be examined for the structure information encoded by each as follows.

**The Kappa Shape Index, ${}^{3}\kappa_{\alpha}$.** The whole molecule index ${}^{3}\kappa_{\alpha}$ encodes information about the centrality/noncentrality of branching. The kappa index ${}^{3}\kappa_{\alpha}$ values are larger when branching is located at the extremities of the molecule.[25] ${}^{3}\kappa_{\alpha}$ is the statistically most significant variable in the model, contributing 50.6% on the average to the calculated p$K$ and ranging from 37.0% to 65.6%. As a single variable, the correlation coefficient between p$K$ and ${}^{3}\kappa_{\alpha}$ is $r^2 = 0.44$. Because of its positive coefficient, larger ${}^{3}\kappa_{\alpha}$ values are related to larger p$K$ values, suggesting that new candidates for greater binding should have more branching at the extremities of the structure. In this data set branching is most common in the vicinity of the D ring, such as a 17-$\beta$ side chain. Compound number 23 has the greatest branching at 17-$\beta$ side chain extremity and also has the largest ${}^{3}\kappa_{\alpha}$ value and one of the largest p$K$ values. The major significance of

$^3\kappa_\alpha$ is further indicated by the fact that the six compounds with the p*K* values greater than 7 have the largest $^3\kappa_\alpha$ values. Compounds with small $^3\kappa\alpha$ values tend to have small binding values.

**The Atom-Level Hydrogen E-State Index, Hs($C_6$).** The second variable, Hs($C_6$), is an atom level descriptor, the hydrogen E-state value for the hydrogen atom at position 6 on the B ring. The Hs($C_6$) value is influenced by the electronic and topological character of nearby structure features, including the double bond between $C_4$ and $C_5$ (skeletons SA, SC, and SF), the double bond between $C_5$ and $C_6$ (SE skeletons), and the aromatic character of ring A (SD skeletons). All the compounds with the SE skeleton have small p*K* values, related to the larger value for HS($C_6$), arising from the double bond between $C_5$ and $C_6$. On average Hs($C_6$) makes the second largest contribution to the calculated binding, 32.8%, and ranges from 27.6% to 44.1%. Because of the negative coefficient on Hs($C_6$), a smaller Hs($C_6$) value tends to relate to a greater binding. A double bond or aromaticity in the vicinity of $C_6$ decreases the value of Hs($C_6$). In this way, Hs($C_6$) encodes the principal effect from the A ring for compounds in this data set.

When the p*K* values are sorted by Hs($C_6$), some appearance of a nonlinear relationship occurs between the Hs($C_6$) variable and p*K* value. The smaller p*K* values are related to the smaller and larger Hs($C_6$) variable values at both data extremities. By using the square of Hs($C_6$) instead of Hs($C_6$), the proc REG yielded a four variable model with statistical information as follows:

$$r^2 = 0.82,\ s = 0.50,\ F = 28,\ n = 31,\qquad r^2_{\text{press}} = 0.73,\ s_{\text{press}} = 0.61$$

These statistics are not sufficiently improved to warrant using the variable Hs($C_6$)square instead of Hs($C_6$) but does not rule out a possible nonlinear relationship.

**The Atom Type E-State Index, $S^T$(=C<).** The third variable, $S^T$(=C<), represents the atom type E-state index for carbon connected to a double bond and two single bonds [11]. The index encodes the electron accessibility, the presence or absence of atom type, and the count of atoms that are of this type, =C<, in the molecule. A carbonyl group (O=C<) at $C_3$ and a double bond between $C_4$ and $C_5$ are necessary for superior steroid activity.[1] This atom type is present in both the A and D rings. In this present data set, the compounds with the skeleton structures of SB, SD, and SE do not possess a carbonyl group at $C_3$ and a double bond between $C_4$ and $C_5$. Twenty molecules possess a carbonyl group (O=C<) at $C_3$ and a double bond between $C_4$ and $C_5$. Five molecules (SB5, SE9, SB12, SD15, SB16) possess one carbonyl group at $C_{17}$ or $C_3$ position. Three molecules (SE3, SE18, SE17) have a double bond between $C_5$ and $C_6$. The three compounds (SC6, SC7, SC11) with the largest p*K* value possess a carbonyl group at $C_3$ and a double bond between $C_4$ and $C_5$ (in the A ring). The compound numbers SB2, SD13, and SD14 have a zero $S^T$(=C<) value because they do not contain a carbonyl group at $C_3$ and a double bond between $C_4$ and $C_5$; all have low p*K* values. The $S^T$(=C<) variable contributes on average 11.9% to the calculated binding but ranges from 0 to 23.6%. Further, because of its positive coefficient, the larger $S^T$(=C<) value tends to have the greater the value of calculated p*K*.

**The Atom Level E-State Index, S($C_{17}$).** The fourth variable, S($C_{17}$), is the atom level E-state index of atom 17 (D ring) and quantifies the electron accessibility of that atom. The S($C_{17}$) variable contributes on average 4.7% to the calculated binding but ranges from 0.4% to 12.3%. Because of its negative coefficient, the smaller S($C_{17}$) value tends to have the greater contribution to calculated p*K*. The 17-$\beta$ keto−enol side chain contributes to the increased potency of the steroids because of a larger negative or a smaller positive S($C_{17}$) value.

Other investigators have pointed out that compound 31 is dissimilar from the rest of the data set because it is the only compound with a substituent at position 9 and the only compound with a halogen substituent.[8] Patterson, Bunce, and Cramer indicated that their very poor prediction for compound 31 arises from those reasons.[8] In this present study, however, compound 31 is not poorly predicted: direct residual = −0.59, LOO residual = −0.71, and LGO residual = −0.80. Although these current results for compound 31 may be considered useful, we remain cautious because of the dissimilarity of compound 31.

In summary, the E-state model indicates that new candidate structures for increased binding should incorporate branching at extremities (primarily in the vicinity of $C_{17}$) and not at the center of steroid nucleus. They should also possess a carbonyl group (O=C<) at $C_3$ and a double bond between $C_4$ and $C_5$. These results are similar to those we obtained in our use of the atom-level E-state indices to produce a COMFA-like field, that is, increased electronegativity in the A ring area and substituents on the D ring. Equation 3 could be used to estimate the binding value p*K* for any suggested candidate structure, as illustrated below.

**Validation Study.** To assess the potential for 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 testing (validation) 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 test 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. We make use of a leave-group-out (LGO) scheme in which each observation is deleted and predicted several times; three times in this present case. 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, four sets of six compounds each and one set of seven 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. Therefore each prediction set is not dependent upon a single selection of one part of the data. This current LGO approach is similar to an approach developed earlier for toxicity of phenols to fish[24] and in our work on the antimicrobial activity of phenylpropyl ethers in which 15% of the data were left out 10 times randomly.[29] It is also most recently implemented in a modeling of tropane binding to the dopamine transporter.[9]

To assess the predictive quality of the model using this process, the mean absolute error of the average predictions (av) was computed: MAE = 0.53. The corresponding root-mean-squared error is found to be rms = 0.60. An examination of these values along with the standard error of the calculated average indicates reasonable predictive quality for the E-state model. No excessively large residuals were obtained, the largest being 1.28, which is 2.5 times the standard deviation of the regression (eq 3) (Table 3). The correlation between the observed p$K$ and the average of the predictions is $r^2 = 0.70$, which compares favorably to the press statistic, $r^2_{\mathrm{press}} = 0.72$, and to the direct statistic, $r^2 = 0.81$. These statistics also compare very favorably with the most recent analysis of this steroid data set. Liu et al. found $r^2 = 0.81$ for a model based on five principal components developed from a set of 25 descriptors.[32] Because principal components are used, direct structure interpretation is not feasible from their model.

**Randomization Study.** The independent variable, binding affinity, was randomized using the random number generator in EXCEL. The randomized p$K$ data replaced the original 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 $r^2 = 0.12$ and $F = 0.89$. The average values found are $r^2 = 0.09$ and $F = 0.64$. These data correspond to a random statistics, giving credence to the model based on the topological variables in eq 3, indicating that the model is significantly different from a random model. This current work is similar to our earlier work on randomization.[30,31]

**Prediction Study.** 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 its press statistic $s_{\mathrm{press}} = 0.62$ is similar to the rms for the consensus model, 0.60, suggesting that this equation may be useful for prediction.

Therefore based on the interpretation given to the variables in eq 3, we have designed a steroid which has SC skeleton with $X_1 \rightarrow {=}O$, $X_2 \rightarrow -OH$, $X_3 \rightarrow -COCHMeCOMe$, $X_4 \rightarrow -OH$, $X_7 \rightarrow -H$, $X_8 \rightarrow -H$, $X_9 \rightarrow -H$, and $X_{10} \rightarrow -H$. This "designed" compound is similar to compound SC23 with an additional branch point on X3 arising from a methyl group. This compound has somewhat increased branching near the D ring (increasing the ${}^3\kappa_\alpha$ value), additional $=C<$ atoms (increasing the $S^T(=C<)$ value), and decreased value of S(C17) due to the presence of additional $>C=O$ groups. Using eq 3, the binding value is predicted as p$K$ = 7.989, a somewhat larger value than for SC23.

**Second**, one can make several predictions of p$K$ 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 structure is predicted from that diminished set. Then, a second randomly selected diminished set of 80% is selected and used to predict the candidate binding value. This process can be repeated several times. In this manner, several predictions are obtained for each candidate. For example, for the compound described above, necessary structure variables were calculated by Molconn-Z. A first subset set was obtained by deleting compounds 5, 6, 17, 22, 26, and 29 (randomly selected). The p$K$ value predicted from the 80% remaining set is found to be 7.796. Following the same procedure, another set was obtained by deleting compounds 2, 7, 10, 14, 20, and 24 and predicting, finding the p$K$ value = 7.887. Finally, for a third time by deleting the set 9, 16, 19, 21, 25, and 30, the predicted value for p$K$ = 7.975. To represent the p$K$ value for the new compound, the mean and standard deviation of these three values was computed: mean p$K$ = 7.886; standard deviation = 0.090. We note that a consistent set of three predictions is obtained in this manner. The mean p$K$ value was obtained by sampling three independently selected portions of the data set to minimize the potential bias arising from only one set. Three predicted values were used here but the method can be extended to more values if desired.

The value predicted by the LGO approach (7.886) and by the full eq 7.989 are not significantly different. However, the LGO approach also yields a standard deviation which may indicate some degree of reliability not available from the single prediction based on eq 3. It remains to be seen whether the standard deviation from the LGO approach is actually a good measure of prediction reliability.

## CONCLUSIONS

For steroids binding to the corticosteroids binding globulin, a statistically satisfactory QSAR model is developed with one kappa shape index and three E-state structure descriptors. Structure information encoded in the descriptors indicates that the large binding (p$K$) values result from the structure effects represented by all four structure indices working together. To obtain the full effect of all structure information encoded in the four descriptors, the p$K$ values for new compounds must be calculated from the whole equation.

This steroid data set has been described by Coats as a benchmark data set for 3D methods.[33] This present study clearly indicates that 3D-based information is not necessary for quality a QSAR. For this present study it was not necessary to use information from computation of 3D-based descriptors such as quantum mechanical charges or HOMO/LUMO energies nor information based on 3D alignment of a series of structures. The E-state structure descriptors adequately encode the information necessary to represent the relationship between structure and binding.

The structure interpretation obtained from this present model is similar to that obtained earlier.[6,7] This present study is much simpler to carry out because it does not require three-dimensional information nor the skeletal alignment of a data set as is required in CoMFA. The detailed structure interpretation is also very straightforward in this present study. In ref 6, a structure entry problem on the last compound produced a model based only on atom-level E-state indices. Because the problem occurred in only one structure, the structure interpretation of this present model is very similar to that of ref 6.

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 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 (p$K$) 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

L.H.H. wishes to give acknowledgment to the Dupont Chemical Corporation for a grant which provided partial support for this research investigation.

**Note.** In an earlier study of this steroid data set,[6] a data entry problem in one structure (#30, ref 6) was made using an old data entry method which is more difficult to validate and is no longer used. The details of structure interpretation given in that paper should no longer be considered reliable although the general features are similar to those given above. We express our regrets for any inconveniences caused by our mistake.

## REFERENCES AND NOTES

(1) Foye, W. O. *Principles of Medicinal Chemistry*, 3rd ed.; Lea & Febiger: Philadelphia, 1989; pp 433−477.

(2) McCluskey, P. J.; Towler, H. M.; Lightman, S. *Br. Med. J.* **2000**, *320*, 555.

(3) Slavkin, H. C. *J. Am. Dental Assoc.* **2000**, *131*, 362.

(4) Rampton, D. S. *Br. Med. J.* **1999**, *319*, 1480.

(5) Baud, O.**;** Foix-L'Helias, L.; Kaminski, M.; Audibert, F.; Jarreau, P.; Papiernik, E.; Huon, C.; Lepercq, J.; Dehan, M.; Lacaze-Masmonteil, T. *New Engl. J. Med.* **1999**, *341*, 1190.

(6) Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR Modeling with the Electrotopological State Indices: Corticosteroids. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 557.

(7) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-State Fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513.

(8) Cramer, R. D., III.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effects of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc*. **1988**, *110*, 5959−5967.

(9) Maw, H. H.; Hall, L. H. E-State Modeling of Dopamine Transporter Binding. Validation of the Model for a Small Data Set. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1270−1275.

(10) Polanski, J. The Receptor-Like Neural Network for Modeling Corticosteroid and Testosterone Binding Globulins. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 553.

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

(12) Kier, L. B.; Hall, L. H. 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.

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

(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.; 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.

(17) Gough, J.; Hall, L. H. Modeling the Toxicity of Amide Herbicides using the Electrotopological State. *Environ. Tox. Chem*. **1999**, *18*, 1069−1075.

(18) Kier, L. B.; Hall, L. H. The E-State in Database Analysis: The PCBs as an Example. *Il Farmico* **1999**, *54*, 346−353.

(19) Kier, L. B.; Hall, L. H. Database Organization and Similarity Searching with E-State Indices. In *Symposium on Computer Methods for Structure Representation*; Kluwer Publishing Co.: Amsterdam, The Netherlands (in press).

(20) Hall, L. H.; Kier, L. B. 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 307−360.

(21) *ChemDraw*, ver 4.5; CambridgeSoft: Cambridge, MA.

(22) *Molconn-Z*, ver 3.50; Hall Associates Consulting: Quincy, MA.

(23) *SAS*, ver 6.12; SAS Institute: Cary, NC.

(24) Hall, L. H.; Kier, L. B. Molecular Connectivity of Phenols and their Toxicity to Fish. *Bull. Environ. Contam. Toxicol*. **1984**, *32*, 354−362.

(25) Hall, L. H.; Kier, L. B. 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 455−489.

(26) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid-Protein Interactions. Human Corticosteroid Binding Globulin: Some Physicochemical Properties and Binding Specificity. *Biochemistry* **1981**, *20*, 6211−6218.

(27) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR from Similarity Matrixes. Technique Validation and Application in the Comparison of Different Similarity Evaluations Methods. *J. Med. Chem*. **1993**, *36*, 433−438.

(28) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulins and Cytostolic AH Receptor Activity by Neural Networks. *J. Am. Chem. Soc*. **1995**, *117*, 7769−7775.

(29) Hall, L. H.; Kier, L. B. Molecular Connectivity and Substructure Analysis. *J. Pharm. Sci*. **1978**, *67*, 1743−1747.

(30) Hall, L. H.; Kier, L. B. A Molecular Connectivity Study of the Muscarinic Receptor Affinity of Acetylcholine Antagonists. *J. Pharm. Sci*. **1978**, *67*, 1408−1412.

(31) Kier, L. B.; Hall, L. H. Structure−Activity Studies on Hallucinogenic Amphetamines Using Molecular Connectivity. *J. Med. Chem*. **1977**, *20*, 1631−1636.

(32) Liu, S.-S.; Yin, C.-S.; Li, Z.-L.; Cai, S.-X. QSAR Study of Steroid Benchmark and Dipepteides Based on MEDV-13. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 321−329.

(33) Coats, E. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug. Discov. Design*. **1998**, *12/13/14*, 199−213.

CI010037I