Development of Quantitative Structure−Activity Relationship and Classification Models for a Set of Carbonic Anhydrase Inhibitors

Brian E. Mattioni and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, 152 Davey Laboratory, University Park, Pennsylvania 16802

Received July 21, 2001

Mathematical models are developed to find quantitative structure−activity relationships that correlate chemical structure and inhibition toward three carbonic anhydrase (CA) isozymes: CA I, II, and IV. Numerical descriptors are generated to encode important topological, geometric, and electronic features of molecular structure. After descriptor generation, multiple linear regression, and computational neural network (CNN) analyses are performed on various descriptor subsets to find superior models for prediction. Committees of five CNNs were utilized to average final predicted values for the 142-compound data set. For inhibitors of CA I, an 8−5−1 CNN committee produced a training set rms error of 0.105 log K_i ($r^2 = 0.994$) and prediction set rms error of 0.208 log K_i ($r^2 = 0.980$). Training and prediction set rms errors of 0.140 log K_i ($r^2 = 0.992$) and 0.231 log K_i ($r^2 = 0.971$), respectively, were produced by a 9−5−1 CNN committee for inhibitors of CA II. For prediction of CA IV inhibitors, an 8−5−1 CNN committee produced training and prediction set rms errors of 0.147 log K_i ($r^2 = 0.992$) and 0.211 log K_i ($r^2 = 0.991$), respectively. In addition, classification models were built using *k*-nearest neighbor (*k*NN) analysis to solve two- and three-class problems for inhibitors of CA IV. A three-descriptor classification model proved superior in labeling compounds as active or inactive inhibitors for the two-class problem. Training and prediction set percent classification rates of 100% and 87.1%, respectively, were obtained. For the three-class (active/moderate/inactive) problem, a five-descriptor model was deemed optimal producing a training set percent classification rate of 98.8% and prediction set rate of 79.0%.

INTRODUCTION

Many different forms of the carbonic anhydrase (CA) enzyme appear in the mammalian body, each having specific functionality.[1] Diseases caused by problematic acid−base secretion chemistry in the body, particularly in the eye, have been linked to the dysfunctional activities of several types of carbonic anhydrases.[2] Excess secretion of aqueous humor in the eye can cause pressure gradients to occur permanently damaging eye tissue. Diseases such as macular edema and open-angle glaucoma can be treated by employing drugs which reduce the rate of formation of aqueous humor. It is believed that certain CA enzymes contribute to the creation of eye humor through production of bicarbonate ions.[1] Drugs inhibiting the activity of the CA isozymes that exist in the eye have been successful in relieving symptoms of these diseases. The synthesis and testing of a wide variety of new drugs which could inhibit CA secretory activity is a continuing goal in the medicinal community.

Quantitative structure−activity relationship (QSAR) methodology can be helpful in screening a large library of possible drug candidates for selectivity and potency.[3−6] Mathematical models are formed that correlate molecular structure to an activity or property of interest. Molecular structure is encoded through the generation of descriptors, which are numerical values corresponding to topological, geometric, or electronic structural features. Lead compounds that could perform a required task (e.g., CA inhibition) can be found computationally, eliminating time and money spent on the synthesis and testing of drugs with poor activities. For example, in the present study, compounds could be screened to identify those which have strong inhibition toward the CA isozymes. Those compounds could then be synthesized and be tested.

Models can also be generated for the purpose of classifying compounds into respective groups.[7] Different statistical classifier methods can be utilized to find models which can label the biological activity (e.g., active/moderate/inactive) of compounds depending on their molecular structure. Even though quantitative information is not obtained from this procedure, large libraries of drug candidates could be screened to identify promising lead compounds based on their predicted activity.

The goal of this work is to develop several QSAR models to predict inhibition values of compounds toward three CA isozymes. Additionally, a classification model which could label inhibition activities toward CA IV was built for all compounds in the data set. These models could improve lead optimization techniques for finding potential pressure-lowering drugs to treat eye diseases such as glaucoma or macular edema. Potential new inhibitors could be screened beforehand, to reduce time and money spent on synthesis and testing. Furthermore, possible insight could be obtained as to what molecular features are deemed important when developing inhibitors of the CA enzyme, with hope that the pathway or mechanism of inhibition can be more clearly understood.

* Corresponding author phone: (814)865-3739; fax: (814)865-3314; e-mail: pcj@psu.edu.

10.1021/ci0100696 CCC: $22.00
Published on Web 11/27/2001

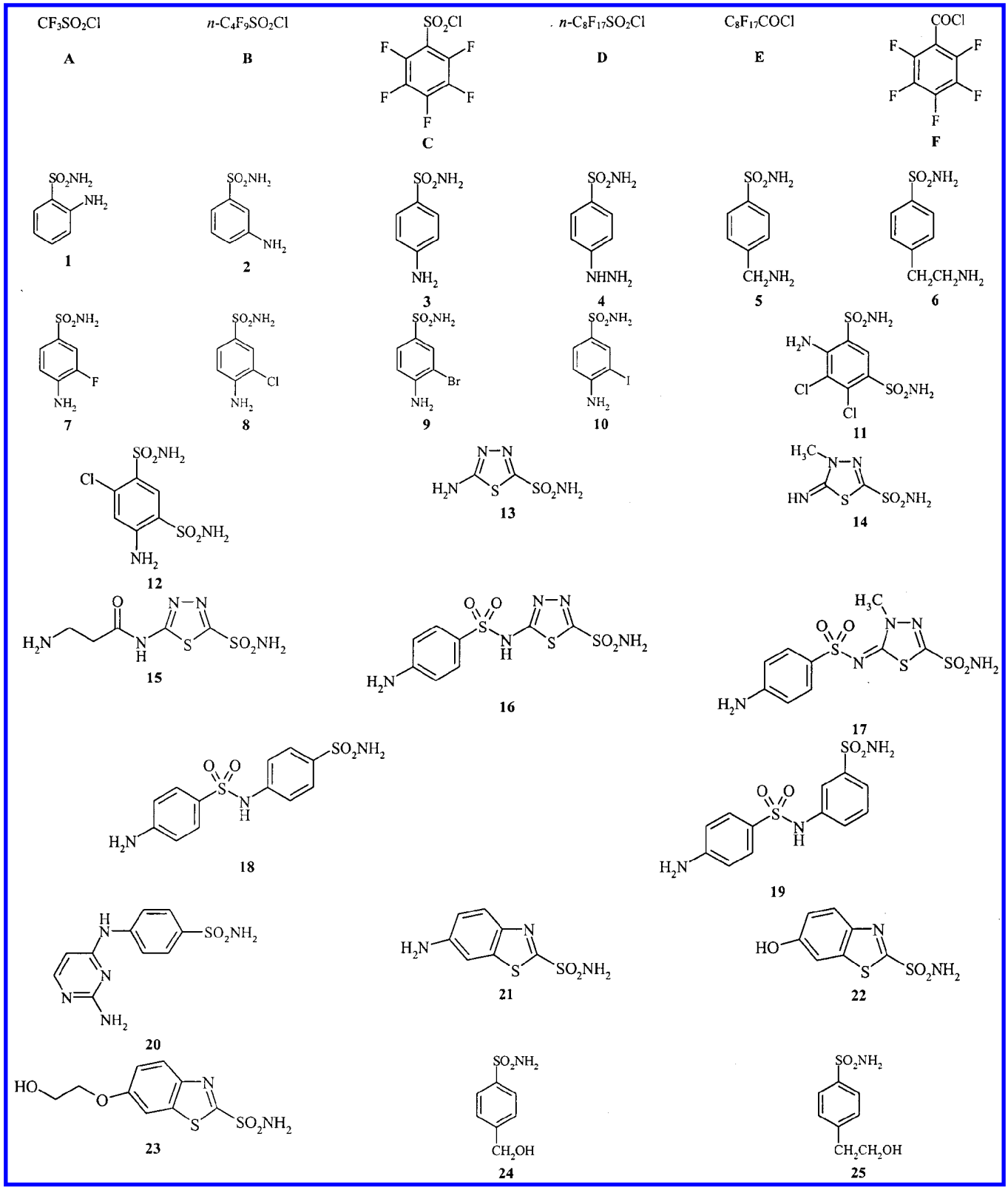

Figure 1. Sulfonamides and perfluorosulfonic/carboxylic acid moieties used in this study.

EXPERIMENTAL SECTION

Data Sets. All data used in this work came from the literature.[8] A set of 25 sulfonamide compounds (**1**−**25** in Figure 1) were derivatized with perfluoroalkyl/arylsulfonyl chlorides or perfluoroalkyl/arylcarbonyl chlorides (**A**−**E** in Figure 1) to produce a data set of 150 compounds. Due to software limitations, eight compounds (**D16**−**D19** and **E16**−**E19**) were removed from the data set leaving 142 structures. The new compounds were assayed for inhibition of three CA isozymes: CA I, II, and IV. The reported inhibition values (K_i) for each compound are reported in Table 1 of reference **8**. The dependent variable used for this study was the logarithm of the K_i value. Log K_i values ranged from 1.23 to 5.32 (mean = 3.17), − 0.70 to 4.31 (mean = 1.43),

and 0.70 to 4.99 (mean = 2.30), respectively, for the CA I, II, and IV isozymes. The molecular weight of the compounds ranged from 304 to 802 (mean = 519) atomic mass units. Three separate QSARs were created, one for each isozyme. A 14-member prediction set was removed for each QSAR study leaving a 128-member training set. Care was taken to ensure that the prediction set dependent variables spanned the range of the entire data set, thereby making the training and prediction set members different for each QSAR experiment. In addition, classification models were developed using *k*-nearest neighbor (*k*NN) analysis to label the inhibitory activity of compounds toward CA IV. A two-class (active/inactive) split and a three-class (active/moderate/inactive) split were investigated. For classification, a 62-member prediction set was randomly removed from the data set leaving an 80-member training set. Since all the models reported were developed with this data set which has limited structural diversity, the models are applicable only to predictions for compounds that are structurally similar to those in the data set.

Model Development. Models were developed using numerical descriptors that were calculated from molecular structure. Topological, geometric, and electronic features of molecules can be captured by the descriptor routines. Information-rich subsets of descriptors were used to form models which were tested for statistical strength and predictive ability. For QSAR, finding models which possessed low training set rms errors was the final goal. On the other hand, classification models were deemed optimal when the training set correct classification percentage reached a maximum. These models were then tested for predictive ability with the external prediction set compounds.

Linear and nonlinear QSAR models were built using multiple linear regression analysis[9] (MLRA) and computational neural networks[10] (CNNs), respectively. For each isozyme study, a 14-member cross-validation set was randomly removed from the training set leaving a new 114-member training set along with the original 14-member prediction set. The cross-validation sets in each study were used to prevent over-training of the neural networks.

All computations were performed on a DEC 3000 AXP Model 500 workstation using the Unix operating system. The **A**utomated **D**ata **A**nalysis and **P**attern recognition **T**oolkit (ADAPT) software,[11,12] simulated annealing,[13] genetic algorithm,[14] *k*NN, and CNN[10] routines developed in our laboratory at Penn State were used to develop QSARs and classification models. Model development involves the following steps: (1) structure entry and optimization *(QSAR/Classification)*, (2) descriptor generation with objective feature selection *(QSAR/Classification)*, (3) model formation and validation by linear means *(QSAR)*, (4) model formation and validation by nonlinear means *(QSAR)*, and (5) model formation and validation using a classifier fitness evaluator *(Classification)*.

Structure Entry and Optimization. Compounds were sketched into HyperChem (HyperCube, Inc. Waterloo, ON) on a PC. HyperChem provided reasonable starting geometries for each compound. Structures were then transferred onto a Unix platform where they were further optimized by the semiempirical molecular orbital package MOPAC.[15] The PM3 Hamiltonian[16] was used to obtain low-energy geometry-optimized structures. Charge information needed by various descriptor routines was calculated using the AM1 Hamiltonian.[17] Two Hamiltonians were used because of the appropriateness of each in providing information from molecular structure.[18]

Descriptor Generation and Objective Feature Selection. Descriptors are numerical values that encode topological, geometric, and electronic features of each molecule. Topological descriptors provide information about molecular size[19] and connectivity.[20,21] Examples include molecular distance edge information[22] and the number of self-avoiding connected paths in a molecule.[23] Geometric descriptors encode molecular features such as the length-to-breadth ratios, shadow projections,[24] and solvent accessible surface areas.[25] Descriptors can also aid in defining the electronic environment of molecules by calculating values such as the partial charge on each atom[26] or dipole moment. In addition, molecular information can be combined to produce hybrid descriptors such as the charged partial surface area (CPSA) descriptors.[27] CPSA descriptors provide information about the extent of interactions between molecules relative to their solvent accessible surface area. Initially, 231 descriptors were calculated for the compounds in the data set.

Redundant or highly correlated descriptors were removed from the descriptor pool during objective feature selection. Redundancy lessens the discriminating power of descriptors, thereby reducing their worth in model development. A descriptor was removed if it had the same value for over 90% of the training set compounds. Furthermore, highly correlated descriptors provide nearly identical information, and only one is needed for model development. Pairwise correlations were performed on members of the descriptor pool, removing one of two descriptors randomly if their correlation coefficient exceeded 0.90. Subjective feature selection was also done to ensure that the ratio of the number of descriptors to training set compounds was below 0.6, which reduces the risk of chance correlations during model formation.[28] The descriptor pool was reduced to 63 members through objective feature selection.

Quantitative Structure−Activity Relationships. Type I: Linear Feature Selection and Linear Modeling. MLRA accompanied by a simulated annealing optimization algorithm was employed to survey the reduced descriptor pool to find models with low training set rms errors. Descriptor subsets were examined to find those which minimized the training set rms error. Subset size was increased sequentially until no significant improvement was seen by addition of another descriptor. The smallest subset of descriptors that provided low training set rms errors was considered optimal.

Several statistical tests were performed on the optimal descriptor subsets to identify the presence of outlier compounds or multicollinearities between descriptors in the subset. If outliers were detected, they were removed to evaluate their contribution to the overall error of the model. A variance inflation factor ($VIF = 1/(1-r^2)$) was calculated for each descriptor in the model by regressing it against the others. All models with VIFs over 10 were not considered for further analysis. Finally, models were validated with the external prediction set of compounds.

Type II: Nonlinear Models Using Best Type I Descriptors. Descriptors from the best Type I model were then passed to a three-layer, fully connected, feed-forward CNN for analysis. CNNs have the ability to generate nonlinear

models with the descriptors to produce predicted values comparable to experimental values. If a nonlinear relationship exists between chemical structure and K_i values, CNNs should be able to identify this correlation.

CNNs consist of three layers: input, hidden, and output. Each layer is comprised of a certain number of neurons, which are connected by paths having adjustable weights. The number of input layer neurons is equal to the number of descriptors in the model. The sole purpose of input layer neurons is to transform descriptor values onto a range of 0.05−0.95. This is done to ensure that some descriptors are not weighted more heavily due to their numerical magnitude. Input values are then passed to the hidden layer where a weighted summation is performed and passed to a nonlinear sigmoidal function to produce an output for each neuron. This same process is repeated by passing hidden layer data to the single output layer neuron. The output is transformed back onto the original dependent variable range and compared to experimental values.

During model development, the number of hidden layer neurons was increased sequentially until no further improvement was seen for that model. The number of hidden layers was restricted, though, to ensure that the ratio of training set compounds to adjustable parameters does not fall below two, thereby reducing the risk of chance correlations.[29] Also, the cross-validation set is used to guide network training to preserve the model's ability to generalize. Periodically predicting values for the cross-validation set compounds can identify the point where network training should be halted to provide optimal results. A point occurs during training where the rms error of the cross-validation set begins to rise, while the rms error for the training set continues to decrease. After this point, the network is starting to memorize specific structural features of only the training set. The weights that produce the minimum cross-validation set error are used as starting points for full network training by the BFGS (Broyden-Fletcher-Goldfarb-Shanno) quasi-Newton optimization algorithm.[30] Finally, model validation is performed employing the prediction set.

Type III: Nonlinear Feature Selection and Nonlinear Modeling. Nonlinear feature selection is performed using simulated annealing and genetic algorithm routines accompanied by a nonlinear CNN. Simulated annealing proved superior in finding optimal models for CA II and IV, while the genetic algorithm found better models for CA I. Performance is based on minimizing the cost function:

$$\mathrm{COST} = \mathrm{TSET}_{\mathrm{rms}} + 0.4|\mathrm{TSET}_{\mathrm{rms}} - \mathrm{CVSET}_{\mathrm{rms}}|$$

Low training set errors are desired, but the cost function ensures that the difference between the training set and cross-validation set does not get too large. The best descriptor subsets are fed to a CNN for complete analysis by the same methodology used for Type II models. The prediction set was used for final validation.

Classification. Model Formation and Validation. Subsets of descriptors were examined using a classifier fitness evaluator accompanied by a genetic algorithm optimization routine. In our implementation, descriptor values were transformed onto the range of 0−1. Models were developed to handle two-class (active/inactive) and three-class (active/moderate/inactive) problems for inhibitors of the CA IV isozyme. Classification models were not developed for CA I and II because a clear separation point did not exist within the dependent variable range to discriminate between classes. Models were tested on their ability to minimize the number of incorrect classifications of the training set compounds.

*k*NN analysis classifies compounds based on the shortest Euclidean distances from their *k* nearest neighbors in descriptor space. Three was chosen for the value of *k* so ties could not exist between classes. *k*NN is a supervised learning method, which requires a set of compounds with known class membership (training set), so descriptor subsets can be found which minimize the number of compounds classified incorrectly. A leave-one-compound-out method is used in our implementation to set up initial classes for the training set compounds. To analyze an unknown compound, the Euclidean distance is calculated from the unknown to its *k* nearest neighbors. The class that possesses the majority of nearest neighbors is assigned to the unknown compound. Final validation was done using the external prediction set.

RESULTS AND DISCUSSION

Quantitative Structure−Activity Relationships (QSAR). CA I. Subset size was varied ranging from three to 10 descriptors during Type I modeling. Only subsets of descriptors that contained T-values with a magnitude greater than four were considered. A T-value with a magnitude greater than four ensures that the standard error of a MLRA coefficient does not exceed 25% of the coefficient value itself. The best model contained eight descriptors which produced a training set rms error of 0.383 log K_i ($r^2 = 0.921$) and a prediction set rms error of 0.452 log K_i ($r^2 = 0.943$). Pairwise correlations for the eight descriptors in this model ranged from 0.389 to 0.894 (mean = 0.720).

These descriptors were then fed as input values to a CNN to form Type II models. A committee of five neural network trainings was used to average output values from the CNN.[5] CNN architectures were tested ranging from 8−2−1 to 8−5−1. A 8−4−1 architecture was deemed best in accordance with training set rms error. This model produced a training set rms error of 0.169 log K_i ($r^2 = 0.985$), cross-validation set rms error of 0.165 log K_i ($r^2 = 0.989$), and prediction set rms error of 0.327 log K_i ($r^2 = 0.955$). This is a 56% improvement for the training set and 28% improvement for the prediction set compared to the Type I results. Due to the improvements seen by the CNN analysis, it was believed that nonlinear Type III modeling would provide even better results.

Fully nonlinear CNN models with various 8-*x*-1 ($x = 2...5$) architectures were formed using the genetic algorithm, the best model being a 8−5−1 CNN. The eight descriptors chosen are shown in Table 1. A calculated versus experimental plot of the best Type III model is shown in Figure 2. A committee of five 8−5−1 CNNs produced rms errors of 0.105 log K_i ($r^2 = 0.994$), 0.133 log K_i ($r^2 = 0.990$), and 0.208 log K_i ($r^2 = 0.980$) for the training, cross-validation, and prediction sets, respectively. This is a 38% improvement for the training set, 19% cross-validation set improvement, and a 36% improvement for the prediction set over the Type II results. Note that all descriptors except one are topological, which are generally preferred for some applications due to the low computation cost that is required during generation.

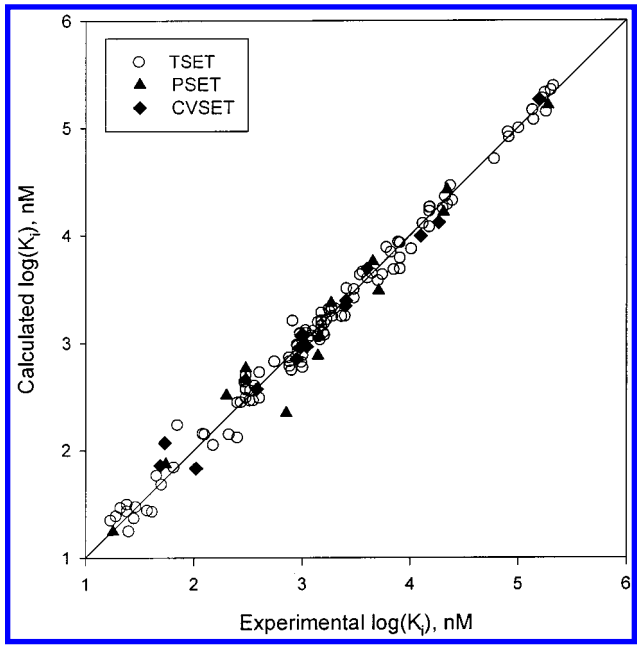

Figure 2. Plot of calculated versus experimental log K_i values for inhibition of the CA I isozyme using nonlinear CNN models. Training set members include 114 compounds, prediction and cross-validation set members − 14 compounds each.

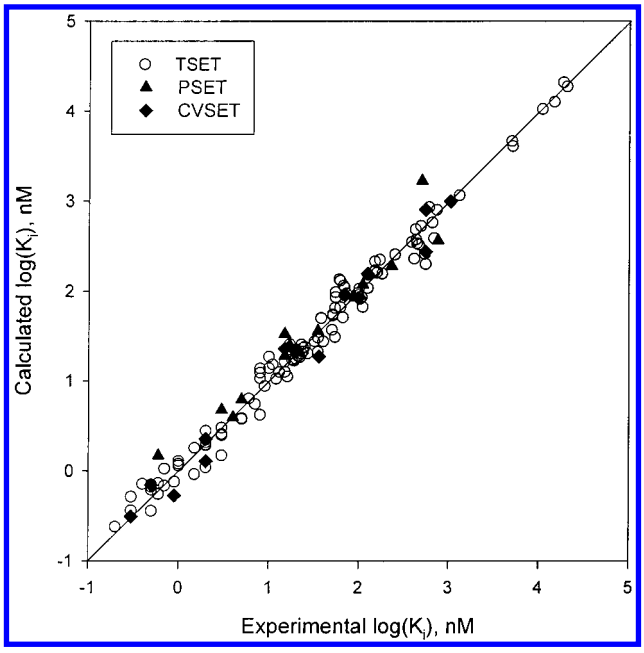

Figure 3. Plot of calculated versus experimental log K_i values for inhibition of the CA II isozyme using nonlinear CNN models. Training set members include 114 compounds, prediction and cross-validation set members − 14 compounds each.

The descriptor ALLP-2[23] encodes information about molecularly connectivity within the compound, which provides information about molecular size and branching. Furthermore, the descriptors V6P-7 and V6PC-14 (valence-corrected sixth-order path/path cluster χ values[20]) contain additional information about size and degree of branching for the data set, as does the descriptor N7CH-20 (number of seventh-order chains (rings)[20]). A count of oxygen atoms per compound is represented by NO-3. The descriptor, WTPT-2,[31] combines information about molecular connectivity and path counts to calculate a molecular ID for each structure. Molecular ID values are structurally dependent, enabling discrimination between molecules. The number of tertiary sp^2-hybridized carbons is accounted for by the descriptor 3SP2-1. This molecular configuration appears in many of the data set structures, particularly in five- and six-member ring moieties. Finally, the hybrid descriptor FNSA-1[27] encodes the extent of interaction between molecules based on their solvent accessible surface areas.

CA II. Type I modeling for inhibition of the CA II isozyme was performed using the procedure outlined above. The best Type I model contained nine descriptors with pairwise correlations ranging from 0.561 to 0.926 (mean = 0.821). The model produced rms errors of 0.325 log K_i (r^2 = 0.954) and 0.427 log K_i (r^2 = 0.881) for the training and prediction sets, respectively. Due to the poor generalization of the linear Type I model, the same descriptors were fed as input values to a CNN, to see if a nonlinear relationship existed between chemical structure and inhibition values.

Model architectures ranging from 9−2−1 to 9−5−1 were investigated. The best Type II model was found by using a committee of five 9−4−1 CNNs. This model produced a training set rms error of 0.170 log K_i (r^2 = 0.988), cross-validation set rms error of 0.201 log K_i (r^2 = 0.988), and prediction set rms error of 0.343 log K_i (r^2 = 0.930). This is a 48% improvement for the training set and 20% improvement for the prediction set over the Type I results. Realizing this improvement, Type III models were explored to further improve results.

Fully nonlinear CNN models with various 9-x-1 (x = 2...5) architectures were formed using simulated annealing with the best model being a 9−5−1 CNN. The nine descriptors chosen are shown in Table 2. A calculated versus experimental plot of the best Type III model for predicting inhibition of the CA II isozyme appears in Figure 3. A committee of five 9−5−1 CNNs produced a training set rms error of 0.140 log K_i (r^2 = 0.992), cross-validation set rms error of 0.163 log K_i (r^2 = 0.990), and prediction set rms error of 0.231 log K_i (r^2 = 0.971). This is a 18% improvement for the training set, 19% cross-validation set improvement, and 33% prediction set improvement over the Type II results.

The descriptors ALLP-2 and FNSA-1 have been described previously. Molecular distance edge[22] information is provided by MDE-33 and MDE-34. Furthermore, the topological descriptor EAVE-2[32] encodes the extent of atomic and molecular interactions that can occur between compounds. The geometric descriptor, L/B-2,[33] is supplying insight about structural size and shape through calculation of molecular length-to-breadth ratios. The electronic environment is being estimated by the descriptor, ENEG-0, by calculating electronegativities for each compound. Two hydrogen bonding descriptors[34,35] (CHAA-3 and SCAA-2) appear in the model, which combine charge and surface area information to form a hybrid class of descriptors. These two descriptors encode the potential for molecular interactions due to hydrogen bonding.

CA IV. Procedures described above were used for model development. An eight-descriptor model proved to be optimal in accordance with training set rms error while forming Type I models. Pairwise correlations among the descriptors ranged from 0.374 to 0.946, producing an average of 0.734. Errors (rms) for the training and prediction sets were 0.411 log K_i (r^2 = 0.935) and 0.404 log K_i (r^2 = 0.944), respectively, demonstrating the ability of the model to generalize.

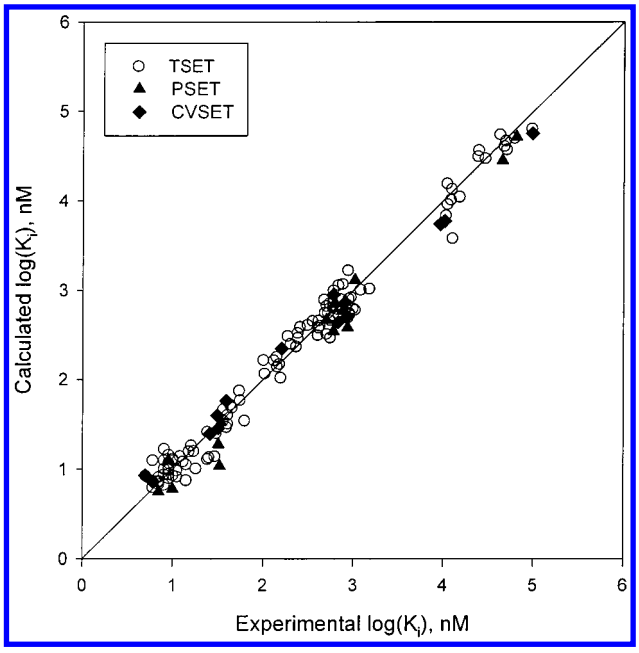

Figure 4. Plot of calculated versus experimental log K_i values for inhibition of the CA IV isozyme using nonlinear CNN models. Training set members include 114 compounds, prediction and cross-validation set members − 14 compounds each.

Descriptors from the best Type I model were then fed to a CNN for nonlinear analysis. CNN architectures ranging from 8−2−1 to 8−5−1 were examined. The best model found employed a committee of five 8−4−1 CNNs, which produced a training set rms error of 0.205 log K_i ($r^2 = 0.984$), cross-validation set rms error of 0.186 log K_i ($r^2 = 0.993$), and prediction set rms error of 0.576 log K_i ($r^2 = 0.908$). This is a 50% improvement for the training set, but 30% deterioration for the prediction set over the Type I results. Three compounds (**A23, B4,** and **D4**) in the prediction set appeared as major outliers contributing to the poor rms error of the model. It is interesting to note that two out of the three, **B4** and **D4**, originate from the same sulfonamide precursor.

Type III models were produced in an attempt to improve the overall predictive ability and to better encode the aforementioned outliers. A committee of five 8−5−1 CNNs produced the best nonlinear Type III model. Rms errors for the training, cross-validation, and prediction sets of 0.147 log K_i ($r^2 = 0.992$), 0.170 log K_i ($r^2 = 0.995$), and 0.211 log K_i ($r^2 = 0.991$), respectively, were obtained from the model. Improvements of 28%, 9%, and 63% were seen for the training, cross-validation, and prediction sets, respectively, over the Type II results. A calculated versus experimental plot for the best Type III model is shown in Figure 4. The descriptors used in this model appear in Table 3. Again, the model is comprised of all topological descriptors except one.

The descriptors N7CH-20 and SCAA-2 have been previously described. The descriptor ALLP-4 encodes molecular connectivity,[31] which describes molecular size and branching. The χ index, S6P-7,[20] provides additional details about molecular connectivity. The descriptors NN-4 and NCl-7 are simple atom counts. In addition, molar refractivity, MREF-1,[34,35] was deemed important for predicting inhibitor values of the CA IV isozyme. Last, electrotopological state indices,[32] such as EMIN-1, are a measure of the reactivity of each atom and meant to encode information regarding intermolecular interactions.

Chance Correlations. Many precautions have been taken to prevent chance correlations throughout model development. Additional experiments were performed to prove the validity of the models developed in this study. The dependent variable (log K_i) was randomly scrambled to produce a new set of dependent variables, which was used to construct models. Under these conditions, no relationship exists between the structure of a compound and its inhibition value. If true quantitative structure−activity relationship models were produced with the real dependent variable, the models formed with the randomly scrambled variables should be very poor.

The same experimental conditions (e.g., subset size, CNN architecture) that produced the best models above were duplicated with the random dependent variable to find models. The best Type III model during the random experiment for inhibition of the CA I isozyme, produced training, cross-validation, and prediction set rms errors of 0.530 log K_i ($r^2 = 0.722$), 0.964 log K_i ($r^2 = 0.107$), and 1.236 log K_i ($r^2 = 0.0004$), respectively. For prediction of inhibition for the CA II isozyme, the best Type III random model produced a training set rms error of 0.650 log K_i ($r^2 = 0.652$), cross-validation set rms error of 0.875 log K_i ($r^2 = 0.529$), and prediction set rms error of 1.848 log K_i ($r^2 = 0.242$). Finally, the best random model produced with Type III methodology for inhibition of the CA IV isozyme produced rms errors of 0.651 log K_i ($r^2 = 0.715$), 0.734 log K_i ($r^2 = 0.481$), and 1.325 log K_i ($r^2 = 0.013$) for the training, cross-validation, and prediction set, respectively. The very poor rms errors and r^2-values of the external prediction set compounds compared to the values obtained with the real dependent variables supplies evidence that the results obtained with the real data were not due to chance effects.

Classification of Compounds Regarding the Inhibition of CA IV. Two-Class Problem. For classification modeling, compounds were labeled as active for log K_i values < 2 and inactive for values ≥ 2. The cutoff was chosen by examining a histogram of dependent variable values for the data set. The distribution of compounds into training and prediction sets is shown in Table 4. A genetic algorithm routine surveyed the descriptor space to find optimal models ranging from two to 10 descriptors. The best model found contained the three descriptors shown in Table 5. Note that this model consisted of only topological descriptors. The model produced a training set correct classification rate of 100% and prediction set classification rate of 87.1%. A confusion matrix for the training and prediction set is shown in Table 6. Eight compounds (**A13−18−19, B12−18−19, C11, F11**) in the prediction set were misclassified, of which three were false positives and five were false negatives.

The two descriptors NN-4 and N7CH-20[20] are counts of nitrogen atoms and seventh-order chains. The appearance of NN-4 is not surprising due to the wide variety of nitrogen containing compounds in the data set. Also, the data set is comprised of compounds which have the molecular configuration that produce varying counts of seventh-order chains. Figure 5 shows the molecular configurations specific to this data set, which will generate seventh-order chains. This descriptor is revealing that five- and six-member rings could be important toward the inhibition of CA IV. A simple fifth-

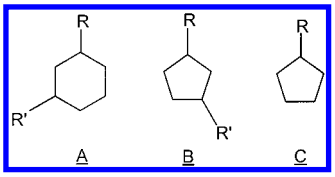

Figure 5. Molecular configurations that produce seventh-order chain counts specific to this data set. R-group = heavy atoms (hydrogens excluded). (A) Substituted six-member ring where each R contributes one count, (B) multisubstituted five-member ring where two different R-groups contribute one count, and (C) substituted five-member ring where R is a chain with two or more heavy atoms.

Table 1. Descriptors Used in Nonlinear Type III Modeling for Prediction of CA I Inhibition

descriptor[a]	type	range
ALLP−2	topological	14.3−60
V6P−7	topological	0.42−2.44
V6PC−14	topological	1.27−11.16
N7CH−20	topological	2−17
NO−3	topological	3−6
WTPT−2	topological	1.86−1.98
3SP2−1	topological	0−2
FNSA−1	hybrid	0.57−0.82

[a] ALLP-2, (total number of paths)/(number of atoms);[23] V6P-7, valence-corrected sixth-order path χ index;[20] V6PC-14, valence-corrected sixth-order path cluster χ index;[20] N7CH-20, number of seventh-order chains;[20] NO-3, number of oxygen atoms; WTPT-2 (molecular ID)/(number of atoms);[31] 3SP2−1, number of 3° sp^2 hybridized carbons; FNSA-1, $[(\Sigma(-SA_i))/SA_{tot}]$, where $-SA_i$ is the surface area contribution of the *i*th negative atom in the molecule and SA_{tot} is the total molecular surface area.[27]

Table 2. Descriptors Used in Nonlinear Type III Modeling for Prediction of CA II inhibition

descriptor[a]	type	range
ALLP−2	topological	14.3−60
MDE−33	topological	0−14
MDE−34	topological	0−37.9
EAVE−2	topological	6.1−11.4
L/B−2	geometric	1.1−3.1
ENEG−0	electronic	5.1−7.2
FNSA−1	hybrid	0.57−0.82
CHAA−3	hybrid	$-4.3 \times 10^{-3} - 8.1 \times 10^{-4}$
SCAA−2	hybrid	−14 − −4.5

[a] ALLP-2, FNSA-1, see Table 1 caption; MDE−33, MDE−34, molecular distance edge between 3°/3° and 3°/4°carbons, respectively;[22] EAVE−2, average E-state value over all heteroatoms;[32] L/B-2, molecular length-to-breadth ratio oriented to give minimum area;[33] ENEG-0, electronegativity $[0.5(E_{HOMO} + E_{LUMO})]$; CHAA-3, [(charge summation over all acceptor atoms)/(total molecular surface area)];[34,35] SCAA-2, [((surface area * charge) summation over all acceptor atoms)/(total number of acceptor atoms)].[34,35]

order chain χ index is providing information about molecular size and degree of branching, which are due to five-member rings in the structures. This adds further support to the theory that five-member rings could be significant when determining CA IV inhibition.

Descriptor ranges and averages from the best model also appear in Table 5. Interestingly, a major trend was that averages for the active compounds tended to be higher than for the inactive compounds. Also, it is interesting to note that only one compound (**A13**) labeled as inactive had a nonzero value for S5CH-17 and was misclassified during the *k*NN analysis. All eight compounds that were misclassified possessed log K_i values close to the cutoff value of 2. This cutoff might not have been the optimal breaking point to distinguish between active and inactive inhibition toward CA IV. Furthermore, in the case of every misclassified compound, discrepancies appeared with descriptor values that did not follow the major trend of descriptor values for the active compounds being higher than the inactive compounds.

Table 3. Descriptors Used in Nonlinear Type III Modeling for Prediction of CA IV Inhibition

descriptor[a]	type	range
ALLP−4	topological	1.9−2.3
S6P−7	topological	1.5−4.8
N7CH−20	topological	2−17
NN−4	topological	1−6
NCl−7	topological	0−2
MREF−1	topological	54−110
EMIN−1	topological	−9.2 − −2.3
SCAA−2	hybrid	−14.4 − −4.5

[a] SCAA-2, see Table 2 caption, N7CH-20, see Table 1 caption; ALLP-4, (total number of weighted paths)/(number of atoms);[23] S6P-7, simple sixth-order path χ index;[20] NN-4, number of nitrogen atoms; NCl-7, number of chlorine atoms; MREF-1, molar refractivity;[36] EMIN-1, minimum atomic E-state value.[32]

Table 4. Distribution of Training and Prediction Set Compounds for Two-Class and Three-Class Problems

	training set	prediction set
Two-Class Problem		
active (log K_i < 2)	40	20
inactive (log $K_i \geq 2$)	40	42
Three-Class Problem		
active (log K_i < 2)	32	28
moderate ($2 \leq \log K_i < 3.9$)	32	29
inactive (log $K_i \geq 3.9$)	16	5

Table 5. Topological Descriptors Defining the Optimal Two-Class Model

	range		average	
descriptor[a]	active	inactive	active	inactive
S5CH-17	0−0.1179	0−0.1179[b]	0.0744	0.0014
N7CH-20	4−17	2−11	9.5	4.5
NN-4	2−6	1−4	3.7	2.1

[a] S5CH-17, simple fifth-order chain (ring) χ index;[20] N7CH-20, see Table 2 caption; NN-4, see Table 3 caption. [b] One inactive compound had nonzero value.

Table 6. Confusion Matrix for Training and Prediction Set Compounds Using the Optimal Model for the Two-Class Problem

	predicted class		
actual class	active	inactive	% correct
Training Set			
active	40	0	100
inactive	0	40	100
Prediction Set			
active	15	5	75.0
inactive	3	39	92.9

Three-Class Problem. For the three-class problem, compounds were labeled as active if log K_i values < 2, moderate if $2 \leq \log K_i < 3.9$, and inactive if log $K_i \geq 3.9$. Table 4 shows the distribution of training and prediction set compounds into their respective classes. The same classification procedure outlined above was used to find the best models.

Table 7. Topological Descriptors Defining the Optimal Three-Class Model

	range			average		
descriptor[a]	active	moderate	inactive	active	moderate	inactive
S6P−7	1.7−4.9	1.5−4.8	1.5−3.0	3.6	3.1	2.2
S6CH−18	0.06−0.33	0.05−0.14	0.07−0.08	0.21	0.09	0.08
NO−3	3−6	3−6	4−5	4.7	4.0	4.3
NN−4	2−6	1−4	1−3	3.7	2.2	1.7
EMIN−1	−9.1 − −2.3	−9.3 − −2.4	−9.1 − −5.6	−5.6	−5.9	−7.0

[a] S6P-7, see Table 3 caption; S6CH-18, simple sixth-order chain χ index;[20] NO-3, see Table 2 caption; NN-4, EMIN-1, see Table 3 caption.

Table 8. Confusion Matrix for Training and Prediction Set Compounds Using the Optimal Model for the Three-Class Problem.

	predicted class			
actual class	active	moderate	inactive	% correct
		Training Set		
active	32	0	0	100
moderate	0	32	0	100
inactive	0	1	15	93.8
		Prediction Set		
active	23	5	0	82.1
moderate	4	23	2	79.3
inactive	0	2	3	60.0

A five-descriptor model was deemed optimal and produced a training set percent correct classification rate of 98.8% and prediction set correct classification of 79.0%. Note that only one compound (**E25**) in the training set was misclassified. Significantly, only topological descriptors comprise this model, which is shown in Table 7. A confusion matrix for the training and prediction set structures is shown in Table 8. The same trend was seen as in the two-class problem in that averages tend to decrease as they go from active → moderate → inactive with the exception of NO-3. Descriptor values for the misclassified compounds deviated from the trend seen above. Five out of eight compounds that were misclassified for the two-class problem were also misclassified for the three-class problem. Furthermore, the confusion matrix reveals that the model never misclassified an active compound as inactive or vice versa. This shows that the model has a good degree of accuracy in labeling compounds for potential inhibitory activity toward CA IV.

The descriptors S6P-7 and S6CH-18[20] are simple χ indices that can encode molecular shape and degree of branching. Both descriptors involve sixth-order paths, which adds to the hypothesis mentioned for the two-class problem above that six-member rings could be deemed important when designing drugs to inhibit the CA IV isozyme. Counts of oxygen and nitrogen atoms are being provided by NO-3 and NN-4, respectively. Last, the descriptor EMIN-1[32] attempts to encode the extent of molecular interactions that can occur between atoms in a molecule.

Chance Correlations. Scrambling experiments were performed to validate the best classification models above and to provide confirmation that good models were not developed by chance. Class labels for the compounds were randomly scrambled 10 times. After each scrambling, the label was used with the same reduced pool of descriptors to try and find good models. Averaging was performed on the percent correct values for the 10 runs on the training and prediction sets. Good models being found using the random class labels provides strong evidence that models produced with the true class label could be chance correlations.

For the two-class problem training set, the 10 models built with random classes produced a classification accuracy of 74.3 ± 3.6%. Classification of the prediction set compounds produced an accuracy rate of 52.1 ± 5.9%. Based on compound distribution, random class labeling would produce an accuracy of 50%, which is close to the value obtained for the prediction set.

The 10 random models found for the three-class problem produced a training set classification rate of 61.1 ± 6.3%. The prediction set classification accuracy was 37.3 ± 6.0%. Random assignment based on compound distribution for the three-class problem would produce a 36% classification rate, which is very close to the one obtained for the prediction set. This offers further proof that chance effects were not playing a role during model development.

CONCLUSIONS

Quantitative structure−activity relationship and classification models were developed with descriptors generated from molecular structure. QSAR models were mostly comprised of topological descriptors which is significant due to the relatively low computational cost required for generation. In addition, models built for classification were comprised of only topological descriptors. If drug-screening models were built to identify possible candidates for inhibition of CA IV, models comprised of topological descriptors could save computational time and money.

For inhibition of the CA I isozyme, the best QSAR model developed produced training, cross-validation, and prediction set rms errors of 0.105 log K_i ($r^2 = 0.994$), 0.133 log K_i ($r^2 = 0.990$), and 0.208 log K_i ($r^2 = 0.980$), respectively. The best model found for CA II inhibition produced a training set rms error of 0.140 log K_i ($r^2 = 0.992$), cross-validation set rms error of 0.163 log K_i ($r^2 = 0.990$), and prediction set rms error of 0.231 log K_i ($r^2 = 0.971$). Last, rms errors of 0.147 log K_i ($r^2 = 0.992$), 0.170 log K_i ($r^2 = 0.995$), and 0.211 log K_i ($r^2 = 0.991$) for the training, cross-validation, and prediction sets, respectively, were obtained for inhibition of the CA IV isozyme.

Classification models were also developed with structural descriptors to label the inhibitory activity toward CA IV. A two- and three-class problem was investigated to classify compounds as active/inactive or active/moderate/inactive, respectively. A three-descriptor model proved optimal for the two-class problem producing a training set correct classification rate of 100% and a prediction set classification rate of 87.1%. Five descriptors were deemed superior for the three-class problem. Training and prediction set correct classification rates produced by the model were 98.8% and 79.0%, respectively.

REFERENCES AND NOTES

(1) *The Carbonic Anhydrases: New Horizons*; Chegwidden, W. R., Carter, N. D., Edwards, Y. H., Eds.; Birkhauser Verlag: Basel − Boston − Berlin, 2000.

(2) Maren, T. H. The Development of Topical Carbonic Anhydrase Inhibitors. *J. Glaucoma* **1995**, *4*, 49−62.

(3) Kireev, D. B.; Chretien, J. R.; Grierson, D. S.; Monneret, C. A. A 3D QSAR Study of a Series of HEPT Analogues: The Influence of Conformational Mobility on HIV-1 Reverse Transcriptase Inhibition. *J. Med. Chem.* **1997**, *40*, 0.

(4) Patankar, 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.

(5) Kauffman, G. W.; Jurs, P. C. Prediction of Inhibition of the Sodium Ion-Proton Antiporter by Benzoylguanidine Derivatives from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 753−761.

(6) Wessel, M. D.; Jurs, P. C. Prediction of Human Intestinal Absorption of Drug Compounds From Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.

(7) Bakken, G. A.; Jurs, P. C. Classification of Multidrug-Resistance Reversal Agents Using Structure-Based Descriptors and Linear Discriminant Analysis. *J. Med. Chem.* **2000**, *43*, 4534−4541.

(8) Scozzafava, A.; Menabuoni, L.; Mincione, F.; Briganti, F.; Mincione, G.; Supuran, C. T. Caronic Anhydrase Inhibitors: Perfluoroalkyl/Aryl-Substituted Derivatives of Aromatic/Heterocyclic Sulfonamides as Topical Intraocular Pressure-Lowering Agents with Prolonged Duration of Action. *J. Med. Chem.* **2000**, *43*, 4542−4551.

(9) Jurs, P. C. *Computer Software Applications in Chemistry*, 2nd ed.; John Wiley and Sons: New York, 1996.

(10) Lu, X.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.

(11) *Computer-Assisted Drug Design*; Jurs, P. C., Chou, J. T., Yuan, M., Eds.; American Chemical Society: Washington, DC, 1979.

(12) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.

(13) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulating Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

(14) Wessel, M. D. Ph.D. Computer-Assisted Development of Quantitative Structure−Property Relationships and Design of Feature Selection Routines, The Pennsylvania State University, 1997.

(15) Stewart, J. P. P. MOPAC 6.0, Quantum Chemistry Program Exchange; Program 455, Indiana University.

(16) Stewart, J. P. P. Mopac: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.

(17) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. P. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(18) Aleman, C.; Luque, F. J.; Orozco, M. Suitability of the PM3-Derived Molecular Electrostatic Potentials. *J. Comput. Chem.* **1993**, *14*, 799−808.

(19) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.

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

(21) Kier, L. B.; Hall, L. H. Intermolecular Accessibility: The Meaning of Molecular Connectivity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792−795.

(22) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.

(23) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17.

(24) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.

(25) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980.

(26) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492−504.

(27) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(28) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative-Structure Property Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(29) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.

(30) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.

(31) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.

(32) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.

(33) Rohrbaugh, R. H.; Jurs, P. C. Molecular Shape and the Prediction of High-Performance Liquid Chromatographic Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1987**, *59*, 1048−1054.

(34) Pimentel, G. C.; McClellan, A. L. *The Hydrogen Bond*; Reinhold Pub. Corp.: New York, 1960.

(35) Vinogradov, S. N.; Linnell, R. H. *Hydrogen Bonding*; Van Nostrand Reinhold: New York, 1971.

(36) Vogel, A. I. *Textbook of Organic Chemistry*; Chaucer: 1977.

CI0100696