QSAR and *k*-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-Based Numerical Descriptors

Gregory W. Kauffman and Peter C. Jurs*

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

Received July 21, 2001

Experimental IC_{50} data for 314 selective cyclooxygenase-2 (COX-2) inhibitors are used to develop quantitation and classification models as a potential screening mechanism for larger libraries of target compounds. Experimental $\log(IC_{50})$ values ranged from 0.23 to ≥ 5.00. Numerical descriptors encoding solely topological information are calculated for all structures and are used as inputs for linear regression, computational neural network, and classification analysis routines. Evolutionary optimization algorithms are then used to search the descriptor space for information-rich subsets which minimize the rms error of a diverse training set of compounds. An eight-descriptor model was identified as a robust predictor of experimental $\log(IC_{50})$ values, producing a root-mean-square error of 0.625 log units for an external prediction set of inhibitors which took no part in model development. A *k*-nearest neighbor classification study of the data set discriminating between active and inactive members produced a nine-descriptor model able to accurately classify 83.3% of the prediction set compounds correctly.

INTRODUCTION

Selective inhibitors of the cyclooxygenase-2 (COX-2) enzyme have received a great deal of attention in the recent literature as superior nonsteroidal antiinflammatory drugs (NSAIDs).[1−4] Among the therapeutic advantages of COX-2 inhibitors over traditional NSAIDs are reduced complications in renal and gastrointestinal function. Research indicates that these complications arise from inhibition of the COX-1 isoform which is constitutively expressed in many cells.[5,6] In contrast, the COX-2 enzyme is thought to be primarily an inducible form which is responsible for the production of prostaglandins triggering pain and inflammation; however, recent reports suggest that COX-2 function may be constitutive in other physiological processes.[7,8]

While the two COX isoforms are approximately 60% structurally homologous, the ability to inhibit one isoform selectively is attributed to the substitution difference at position 523.[2] The COX-1 enzyme has an isoleucine residue at this position, while the corresponding position on the COX-2 enzyme has a valine residue. This seemingly minute difference in structure results in a larger central channel in the COX-2 isoform. As a result, drug substrates which are too large to enter the COX-1 channel may be able to enter the COX-2 channel based solely on size discrimination. Exploitation of this phenomenon has been the key to advances in developing selective COX-2 inhibitors.

A variety of computational approaches for development of COX-2 inhibitors have appeared in the literature over the past several years. The methods employed include comparative molecular field analysis (CoMFA),[9,10] receptor surface analysis (RSA),[9] simple quantum mechanical correlations,[11] the Fujita-Ban approach,[12−14] and the classical Hansch QSAR approach.[13−15] All models presented are of sufficiently high quality to be used as an estimator of biological activity of unknown compounds; however, each has minor limitations. Most of the data sets examined were small and limited in structural diversity. In addition, many of the reports utilized 3D-QSAR approaches. In the later stages of development, 3D-QSAR approaches are invaluable for identifying specific drug/active site interactions; however, for larger libraries of compounds the computational burden of obtaining accurate structural geometries for all compounds can be impractical for initial screening purposes.

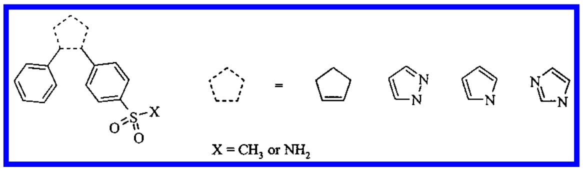


Figure 1. Generic structure for the 314 compounds used in this study. The central five-member ring core is either a cyclopentene, pyrazole, pyrrole, or imidazole.

The goal of this work is to develop QSAR models which could serve as practical screening tools for accurate and rapid quantitation and classification of a set of specific COX-2 inhibitors. We have used only topologically based numerical descriptors, which require no specific orientation or through-space distances, thus alleviating the need for geometry optimization of the structures. This approach to QSAR analysis has been increasingly employed in quantitation[16−18] and classification[19−21] problems as this type of descriptors has evolved into a more information-rich set of indices.

EXPERIMENTAL SECTION

Data Set. The 314-member data set has four series of compounds, each with a five-membered ring as its core structural element (Figure 1). The breakdown of compounds

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

10.1021/ci010073h CCC: $20.00
Published on Web 09/26/2001

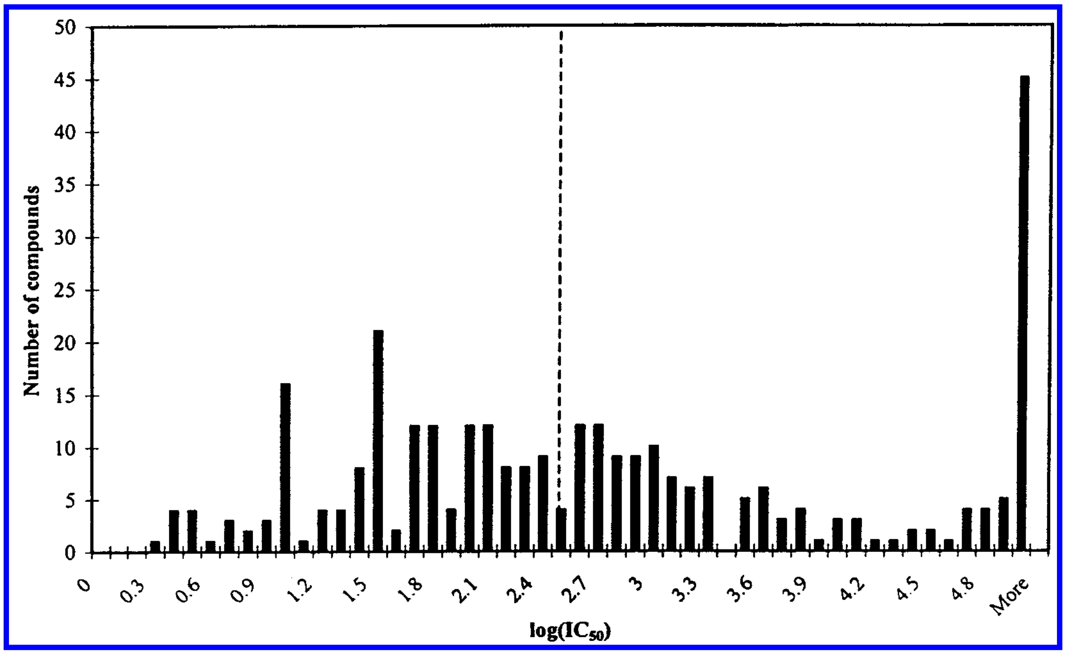


Figure 2. Histogram of dependent variable values and the split point for active and inactive classification (dashed line at 2.5 log(IC_{50})). Dependent variable values less than 2.5 are considered active and those above 2.5 are considered inactive.

shows that 30 compounds contain a cyclopentene core, 110 contain a pyrazole core, 34 contain a pyrrole core, and 140 contain an imidazole core. Furthermore, each compound possesses two aryl substituents in a 1,2 relationship on the five-member ring. The final structural homology between all members of this data set is the presence of a sulfonate or sulfonamide moiety on at least one of the aryl substituents and always in a 1,4 relationship to the five-membered ring core. This scaffold is the common theme among two selective COX-2 inhibitors recently marketed, Celecoxib[22] and Rofecoxib.[23] With the success of these two inhibitors, development of predictive models for this structural class of compounds may offer advantageous tools for high throughput screening in search of more highly selective inhibitors.

One precaution that must be taken when developing predictive models with biological data is to ensure that experimental conditions and assays be nearly identical or strictly compatible. To guarantee the quality of data used in this study, IC_{50} values for the 314 antiinflammatory target compounds were taken from five literature sources published by the same laboratory.[22,24−27] All data are in vitro and were evaluated in human recombinant COX enzyme assays described by Gierse and co-workers.[28] This data set was chosen for the thoroughness of SAR analysis performed by the authors. While structural diversity exists among the four series of compounds, systematic substitutions at key positions on the scaffold provide ample homology to test the capabilities of theoretical descriptors to capture biophoric information at an atomic level.

The IC_{50} values range from 1.7 to $\geq$100 000 nM and were reported as the average of duplicate or triplicate measurements. Inequality values were reported for 41 of the compounds; therefore, the 273 structures with definitive values were used for the QSAR analysis. For analysis purposes, log(IC_{50}) values were used as the dependent variables, evening the distribution of the data on the range 0.23−5.00 log units.

For the QSAR portion of this work, 10% of the compounds were set aside to serve as an external prediction set for validation of the models developed. Furthermore, for models employing CNNs, an additional 10% of the data was removed to serve as a cross-validation set, which was used during CNN training. The choice of sets was made by binning the range of experimental values and randomly selecting an even distribution of compounds from each bin. This guaranteed that members of the cross-validation and prediction sets spanned the entire range of the experimental measurements and would be numerically representative of the data set. In total, the training set contained 220 compounds, the cross-validation set 26 compounds, and the prediction set 27 compounds. The breakdown of compound series distribution for the cross validation/prediction sets, respectively, are as follows: 1/1 from the cyclopentene series, 13/6 from the pyrazole series, 1/4 from the pyrrole series, and 11/16 from the imidazole series.

All 314 compounds were used for the classification analysis. An active/inactive split point of 2.5 log(IC_{50}) units was chosen (Figure 2) generating evenly sized classes while compromising few compounds near the split point. This resulted in a 153-member class of active compounds and a 161-member class of inactive compounds. Of these, 30 active and 30 inactive compounds were chosen for an external prediction set leaving a 254-member training set. The prediction set compounds were randomly chosen from each of the four structural series, thus ensuring a representative sampling of compounds were withheld for validation purposes. The specific breakdown of compounds in the prediction set are as follows: 10 from the cyclopentene series, 21 from the pyrazole series, 9 from the pyrrole series, and 20 from the imidazole series.

Computational Methodology. This work was performed using the Automated Data Analysis and Pattern recognition Toolkit (ADAPT) software package.[29,30] ADAPT has been used to find highly predictive models for various pharmacological[19,31−33] and toxicological[34] properties. All descriptor calculation, simulated annealing,[35] genetic algorithm,[36] computational neural network,[37] and classification algorithms are written in FORTRAN and were run on a DEC 3000 AXP Model 500 workstation under a Unix operating system. All structures were sketched using Hyperchem (Hypercube, Inc., Waterloo, ON) on a Pentium desktop PC. This generated connection tables for each of the compounds with bond and atom adjacency information to be used with the ADAPT software. Analysis of the COX-2 inhibitors proceeded through four stages: (1) descriptor generation and feature selection, (2) quantitative multiple linear regression model development and validation, (3) quantitative computational neural network model development and validation, and (4) active versus inactive classification model development and validation.

Descriptor Generation and Feature Selection. Only topological descriptors were employed for QSAR and classification model development. These descriptors include κ indices,[38,39] χ indices,[40] weighted path indices,[41] molecular distance edge measures (λ),[42] superpendentic indices,[43] electrotopological state indices,[44] simple atom and bond counts, and carbon hybridization descriptors. These descriptors are designed to encode wholistic properties of the compounds such as size, shape, branching, and constitution from an atomic level. A total of 135 descriptors were calculated for each compound and were subsequently reduced by objective feature selection. The first reduction method removed descriptors for which greater than 90% of the values were identical across all compounds in the training set. The second reduction method randomly removed one of two descriptors if their pairwise correlation coefficient was greater than 0.95. The experimental data had no influence on this phase of the descriptor reduction process. This type of feature selection is solely driven by descriptor values for the training set compounds. Therefore, while the reduced pool for the QSAR analysis contained 75 descriptors and the reduced pool for the classification analysis contained 74 descriptors, the composition of each pool was marginally different.

Multiple Linear Regression Model Development. The 75-descriptor reduced pool was submitted to subjective feature selection routines which searched for subsets of information-rich descriptors. Searches were conducted using simulated annealing and genetic algorithm optimization techniques coupled with a multiple linear regression (MLR) fitness evaluator to find model sizes employing 5−12 descriptors. During the model screening process, for each descriptor in a subset the ratio of its model coefficient to its respective standard error was calculated, termed a *T*-value. A *T*-value threshold of four was used in this study ensuring that a standard error did not exceed 25% of the corresponding coefficient value. In addition, variance inflation factors (VIF) were calculated to identify whether excessively high multicollinearity coefficients existed among the descriptors in a subset. A VIF greater than 10 corresponds to a multiple correlation coefficient of greater than 0.95, which was used as the threshold for this work. The final selection criteria were based upon the number of descriptors in the subset, with small subsets receiving preference over larger ones, and minimization of the root-mean-square (rms) error. The model which passed all statistical diagnostics and generated the lowest rms error with as few descriptors as possible was chosen as the optimal linear model. Validation of the model was then performed on the external prediction set of compounds withheld from training.

Computational Neural Network Model Development. The descriptors used in the MLR model were submitted to computational neural networks to improve the model performance. CNNs are well suited for QSAR applications when a high degree of nonlinear character may be present in the data. The CNNs used for these analyses are three-layer, fully connected, feed-forward networks, and they have been described in detail previously.[37,45] The number of neurons in the input layer corresponds to the number of descriptors in the model. In this layer, the descriptor values are transformed on the range {0.05, 0.95}. The number of hidden layer neurons controls the flexibility of the network and was adjusted until the optimal network architecture was identified. Care was taken to restrict the number of hidden neurons within the adjustable parameter guideline proposed by Livingstone and Manallack, which suggests that a ratio of training set compounds to adjustable parameters be greater than two to reduce the risk of chance correlations.[46]

Network training was directed by optimization of the weights and biases using the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm.[47−50] Each network training was terminated at the minimum cross-validation set rms error and was evaluated using the cost function defined by eq 1, where T_{rms} and $\mathrm{CV}_{\mathrm{rms}}$ are the rms errors for the training and cross-validation sets, respectively. This cost function

$$\mathrm{cost} = T_{\mathrm{rms}} + 0.4^{*}|T_{\mathrm{rms}} - \mathrm{CV}_{\mathrm{rms}}| \qquad (1)$$

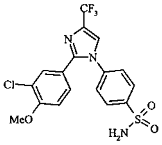
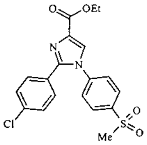
assesses both the minimization of the training set rms error as well as the difference between the training and cross-validation set rms errors. At the minimum cross-validation set rms error, the network ceases to use global information about the structures and begins memorizing idiosyncrasies of the training set compounds, thus leading to poor validation results. Therefore, the network architecture which generated the lowest and most consistent training set and cross-validation set rms errors was deemed optimal. Finally, an average of 10 network trainings was used to calculate the final rms errors for each compound set to reduce the risk of overrated or underrated results from any one network training.

Classification Model Development. The descriptors in the 74-member reduced pool were scaled on the range {0, 1} and then submitted to a single-point crossover genetic algorithm optimization coupled with a classifier fitness evaluator. Both *k*-nearest neighbor (*k*NN) and linear discriminant analysis classifiers were examined; however, the results for the *k*NN classification were superior and will therefore be discussed in detail in this paper. *k*NN is an algorithmically simple, supervised method which classifies an unknown compound based on the class membership of its *k* nearest neighbor compounds. To determine the nearest neighbors, an Euclidean distance matrix of all samples in the training set is scanned for the *k* shortest distances to the observation of interest. In the present implementation, a leave-one-compound-out approach is used to assign class membership to an unknown. The class membership of the majority of the *k* neighbors is then assigned to the unknown. For this work, $k = 3$. The descriptor space was searched for

Table 1. Descriptors Included in the MLR and CNN QSAR Models

descriptor	coefficient	error	explanation
constant	3.914	0.746	
S5P-6	2.561	0.293	simple fifth-order path χ index
S7CH-19	−7.500	1.328	simple seventh-order chain χ index
MOLC-5	−1.239	0.219	third-order path molecular connectivity
2SP2-1	−0.421	4.904×10^{-2}	no. sp^2 carbons attached to two carbon atoms
2SP3-1	−0.460	9.308×10^{-2}	no. sp^3 carbons attached to two carbon atoms
PND-5	1.165×10^{-2}	1.607×10^{-3}	superpendentic index from pendant oxygens
EAVE-1	−31.000	3.681	av electrotopological state index over all heavy atoms
EAVE-2	0.513	0.101	av electrotopological state index over all heteroatoms

Table 2. Two Representative Compounds and Their Descriptor Values for the Eight-Descriptor QSAR Model

descriptors			
label	range		
S5P-6	4.84−8.70	6.07	5.95
S7CH-19	0.411−0.965	0.634	0.569
MOLC-5	3.07−5.98	4.28	4.25
2SP2-1	4−15	7	8
2SP3-1	0−4	0	0
PND-5	0−258.0	41.4	87.4
EAVE-1	0.159−0.344	0.288	0.235
EAVE-2	5.27−10.6	8.02	7.24
exptl log(IC_{50})		1.301	3.756
predicted log(IC_{50})		1.638	3.927

subsets which minimized the number of incorrect classifications. Model sizes from 5−12 descriptors were examined, and the optimal model size was chosen solely by the lowest percentage of total misclassifications in the training set. Validation of the best model was then performed using the external prediction set.

RESULTS AND DISCUSSION

Multiple Linear Regression Model. The best linear model is shown in Table 1. Descriptor values for two representative and structurally similar compounds (**234** and **256**) from the data set along with descriptor value ranges are included in Table 2. The three χ indices, S5P-6, S7CH-19, and MOLC-5, are the simple fifth-order path index, the simple seventh-order chain index, and the valence and ring-corrected third-order path molecular connectivity index, respectively.[40] The values of S5P-6 and S7CH-19 are sensitive to the order of the bond vertices along the path of each respective length they encode. In other words, higher degrees of branching along paths of length five and seven will result in larger descriptor values for each of these. The values are particularly sensitive to substitutions on ring systems which is demonstrated by the difference in the values of these two descriptors for the compounds in Table 2. This is potentially capturing information about the branching and the steric bulk of each molecule. MOLC-5 is a more sophisticated χ index which takes into account valence corrections for heteroatoms and aromatic ring atoms in paths of length three. This could be important for delineating electronic differences among the core five-membered ring structures and for highly homologous compounds where one heteroatom is replaced by another. For the two pyrazole compounds in Table 2, the values for this descriptor are nearly identical; however, compounds from the cyclopentene series had an average value 0.5 units higher than either the pyrazole or imidazole series and approximately 0.1 units higher than the pyrrole series. This indicates that a trend may exist between the magnitude of this descriptor and the number of nitrogen atoms in the core ring structure.

The descriptors 2SP2-1 and 2SP3-1 are simply a count of the number of sp^2 and sp^3 carbon atoms attached to two other carbon atoms, respectively. The additional chlorine substitution on the aromatic ring in **234** accounts for the difference in the value of 2SP2-1 between the two example compounds. The descriptor PND-5 is a variation of the superpendentic index which measures distance terms from only pendant oxygen atoms on a hydrogen-suppressed graph.[43] Like the χ indices, higher degrees of branching result in larger descriptor values for this index, thus providing useful information about the steric and bulk size and shape of each molecule relative to the number of pendant oxygen atoms. Compound **256** in Table 2, with an additional pendant oxygen as part of the ester carbonyl moiety, demonstrates the dramatic effect that the introduction of one more terminal oxygen can have on this descriptor value. Finally, the descriptors EAVE-1 and EAVE-2 are the average electrotopological state (e-state) indices over all heavy atoms (non-hydrogen atoms) and all heteroatoms (non-hydrogen atoms excluding carbon atoms), respectively.[44] The e-state index is highly correlated to free-valence theory which has been used to identify the most reactive atoms in chemical structures. The average e-state descriptors here may be encoding the overall propensity for a molecule to participate in intermolecular interactions. Compound **234** has larger values than **256** for both of the these descriptors. This is likely explained by the presence of three more heteroatoms in **234** than in **256**. Compound **234** contains a dense area of very electronegative fluorine atoms as well as a sulfonamide group (versus a sulfonate group in **256**) which may rationalize the differences in descriptor values for the two compounds.

All eight descriptors had absolute *T*-values greater than four and VIFs less than 10 indicating that the model coefficients were statistically sound and that the multicollinearities between the descriptors were acceptable. The range of pairwise correlations between the descriptors was −0.443 to 0.784 with an average value of 0.153. The descriptor coefficients, standard coefficient errors, and descriptor value ranges are included in Table 1. The rms error for the training set was 0.845 log(IC_{50}) units (R = 0.669) and for the prediction set was 0.655 log(IC_{50}) units (R = 0.666). The linear relationship between structure and the experimental log(IC_{50}) values is not very strong. Therefore, the descriptors listed in Table 1 were submitted to CNNs to determine whether nonlinearity in the training method would produce superior results.

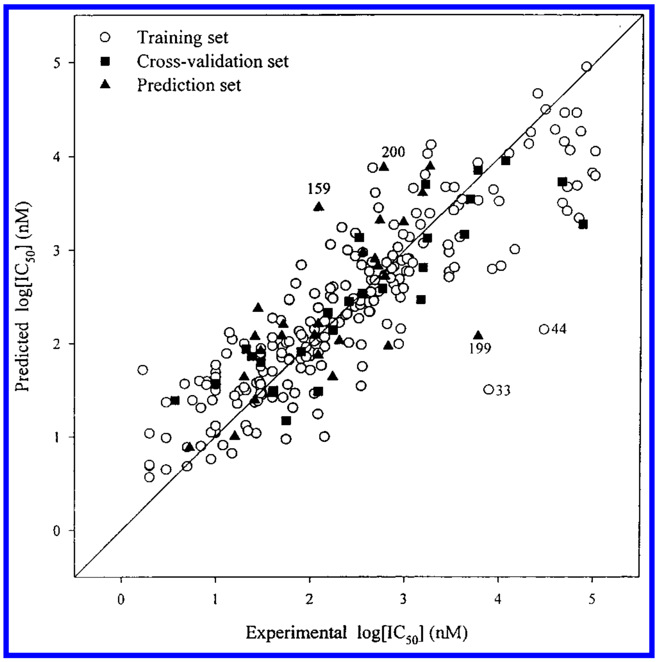


Figure 3. Correlation plot for the eight-descriptor neural network model. The diagonal line denotes a perfect one-to-one correlation. Outliers in the training set (**33** and **44**) and prediction set (**159**, **199**, and **200**) are labeled for clarity.

Computational Neural Network Model. Neural network architectures ranging from 8-2-1 (21 adjustable parameters) to 8-8-1 (81 adjustable parameters) were evaluated using the cost function defined by eq 1. Network architectures which produced the lowest individual cost functions (8-4-1, 8-5-1, and 8-6-1) were then separately evaluated based on the lowest average rms errors for a committee of 10 networks. The committee of 10 8-5-1 networks (51 adjustable parameters) was found to give the best results. The resultant rms errors for this analysis were 0.551 log(IC_{50}) units ($R = 0.876$) for the training set, 0.535 log(IC_{50}) units ($R = 0.883$) for the cross-validation set, and 0.625 log(IC_{50}) units ($R = 0.719$) for the prediction set. The predicted versus experimental activity plot for this model is shown in Figure 3. The CNN results for training were considerably improved over the MLR model; however, the results for the prediction set were only marginally improved. It should be noted that a feature selection algorithm which employs a CNN as its fitness evaluator was also used; however, the results showed no improvement over descriptors selected by MLR.

As is clear to see from the correlation plot shown in Figure 3, several outliers do exist within the training, cross-validation, and prediction sets. Examination of these outliers showed that 15 compounds in the training set had prediction residuals greater than one log(IC_{50}) unit. Two of these compounds had prediction residuals greater than two log units. These two most poorly predicted compounds **33** and **44** (Figure 4) belong to the pyrazole series of compounds. The difficulty in predicting the activity of these compounds is likely attributed to the activity of these two compounds being extremely high relative to the other compounds in this structurally homologous series. Therefore, these two compounds are essentially being predicted at the average activity of the other structurally similar compounds in the training set. If these two compounds were removed from the rms error calculation, a reduction of the error to 0.506 log(IC_{50}) units ($R = 0.897$) would result; a significant difference considering the size of the training set.

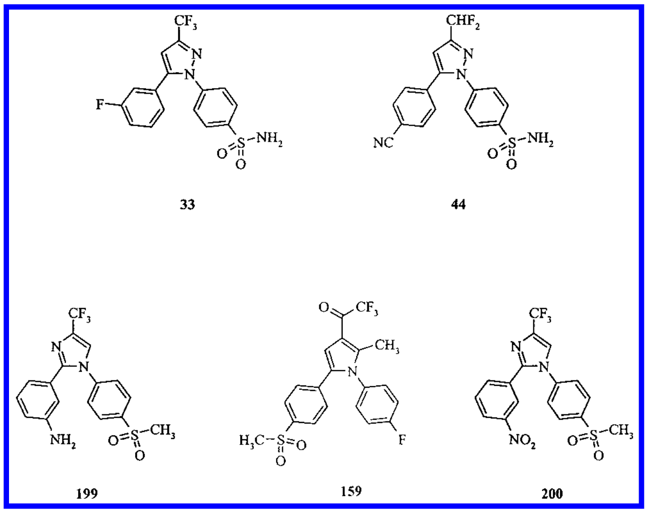


Figure 4. Notable training set (**33** and **44**) and prediction set (**159**, **199**, and **200**) compound outliers for the nonlinear CNN model.

Table 3. Comparison of Compound Series in QSAR Analysis

core ring structure	av residual	rms error
cyclopentene	0.281	0.391
pyrazole	0.494	0.672
pyrrole	0.430	0.588
imidazole	0.362	0.484

A similar argument can be made about the three largest outliers in the prediction set **199**, **159**, and **200** (Figure 4). While the activities of compounds **199** and **200** are quite different, the activities of the other compounds in that subseries of structurally related imidazoles are much lower than both, perhaps making structural discrimination of these two compounds from the rest extremely difficult. The same is true for compound **159** among other members of the pyrrole series. Recalculation of the prediction set rms error without **199** resulted in an rms error of 0.543 log(IC_{50}) units ($R = 0.817$) which is on the order of the training and cross-validation set rms errors. Further removal of compounds **159** and **200** in addition to compound **199** produced an rms error of 0.436 log(IC_{50}) units ($R = 0.863$). The effect of these poor predictions clearly illustrates that only a few compounds have skewed the rms error of the prediction set from that of the training and cross-validation sets. For each core ring series, Table 3 shows the average residual and the rms error for all compounds in that series regardless of set membership. It can be seen that predictions were much better overall for the cyclopentene and imidazole series than for the other two series.

Monte Carlo Experiments. The possibility of chance effects playing a role in model development must be investigated in QSAR. Therefore, Monte Carlo randomization experiments were performed for both model types. For the QSAR analysis, the dependent variable values were scrambled such that each member of the data set was assigned the experimental log(IC_{50}) value of another member of the data set. Simulated annealing/MLR optimizations were performed on 10 different sets of randomly scrambled dependent variables to identify the best eight-descriptor models. The training set rms errors for these models ranged from 1.055 to 1.076 log(IC_{50}) units with correlation coefficients ranging

Table 4. Descriptors Included in the *k*NN Classification Model

descriptors	range	explanation
KAPA-6	3.39−6.74	atom-corrected third-order κ index
2SP2-1	4.00−15.0	no. sp^2 carbons attached to two C atoms
3SP2-1	0.0−5.0	no. sp^3 carbons attached to two C atoms
MDE-12	0.0−4.26	distance edge between 1° and 2° C atoms
MDE-14	0.0−16.3	distance edge between 1° and 4° C atoms
MDE-23	0.0−24.5	distance edge between 2° and 3° C atoms
MDE-33	9.49−77.6	distance edge between 3° and 3° C atoms
MDE-44	5.36−22.8	distance edge between 4° and 4° C atoms
PND-5	0.0−258.0	superpendentic index from pendant O atoms

Table 5. Confusion Matrix for the 254-Member Training Set and the 60-Member Prediction Set

	training set		prediction set	
class	active	inactive	active	inactive
active	105	18	26	4
inactive	26	105	6	24

Table 6. Comparison of Misclassified Samples by Compound Series

	no. of misclassified samples	
core structure	training set (of 254)	prediction set (of 60)
cyclopentene	1	1
pyrazole	20	5
pyrrole	6	2
imidazole	17	2

from 0.195 to 0.398. The average prediction set rms errors ranged from 1.113 to 1.255 log(IC_{50}) units with correlation coefficients ranging from −0.377 to 0.001.

For each of the 10 linear models, an average of 10 neural network trainings were performed on each. The training set rms errors ranged from 0.847 to 1.014 log(IC_{50}) units with correlation coefficients ranging from 0.407 to 0.658. The cross-validation set rms errors ranged from 0.817 to 1.238 log(IC_{50}) units with correlation coefficients ranging from −0.048 to 0.590. Finally, the prediction set rms errors ranged from 0.905 to 1.505 log(IC_{50}) units with correlation coefficients ranging from −0.295 to 0.197. Since significant results were not obtained in the experiments using 10 sets of scrambled dependent variables, the likelihood that chance correlations are responsible for the quality of the real models is low.

Classification Model. Using the *k*NN classification procedure described, the nine-descriptor model listed in Table 4 was identified as optimal. The descriptor KAPA-6 is the atom type-corrected $^3\kappa$ index (three-bond fragments).[39] The κ index is designed to encode information about the shape of molecules based on molecular graph theory. The two descriptors, 2SP2-1 and 3SP2-1, are a count of the sp^2 hybridized carbons attached to two other carbon atoms and of the sp^2 hybridized carbons attached to three other carbon atoms, respectively. The five descriptors, MDE-12, MDE-14, MDE-23, MDE-33, and MDE-44, are all from the molecular distance edge class of descriptors.[42] Distance edge measures can be calculated for all combinations of primary, secondary, tertiary, and quaternary carbon atom connectivity, which the number label on each descriptor denotes (i.e., MDE-12 is the λ measure between all primary and secondary carbon atoms). These descriptors encode atom adjacency and branching information such that structural isomers of a compound will have unique distance edge values. Clearly, these descriptors have a profound effect on the discrimination of this data set. Finally, the descriptor, PND-5, is the superpendentic index from pendant oxygen atoms.[43] As previously discussed for this descriptor in the QSAR model, information about the bulk size and shape of each molecule relative to terminal oxygen atoms is being encoded.

The model was able to correctly classify 82.7% of the training set compounds (210 of 254 compounds) and 83.3% of the prediction set compounds (50 of 60 compounds). Table 5 shows the confusion matrices for the training and prediction sets. For the training set compounds, a relatively even distribution of misclassifications was observed. While 18 active compounds were incorrectly predicted as inactive (false negative), the model predicted that 26 of the experimentally inactive compounds were active (false positive). Of the 44 misclassifications in the training set, 15 have dependent variable values within ±0.5 log units of the split point (2.5 log units). Unfortunately, only five of the 18 false negatives fell within ±0.5 log units of 2.5 log(IC_{50}) indicating that active compounds near the split point are not the only ones being misclassified.

An interesting observation is that nine of the inactive compounds with experimental log(IC_{50}) values > 5.0 were incorrectly misclassified as active. Due to the extent of homology among the compounds in the data set, relative to those which were misclassified in particular, it is difficult to speculate on structural features which contributed to this result. An examination of the descriptor values for each of the misclassified compounds revealed that each had values embedded within the range of all compounds in the data set. Furthermore, average descriptor values for the active and inactive compounds followed no particular trend which might elucidate an interpretable rationale for the good discrimination by this model. This obviates the need for a more complete set of descriptors capable of discriminating the active and inactive COX-2 inhibitors.

Validation of the model was performed using the 60-compound prediction set. In this set, four of the active compounds were misclassified as inactive, and six of the inactive compounds were misclassified as active. Three of the four false negatives had experimental values within ±0.5 log(IC_{50}) units of the split point, indicating that the active compounds which are being incorrectly classified are generally among the less active within that class. Two of the false positives had experimental values within ±0.5 log(IC_{50}) units from the split point value, while the other four had experimental inequality values (log(IC_{50}) > 5.0). Again, despite the fact that clearly inactive compounds are being incorrectly labeled active, the high classification accuracies for both of the sets demonstrates that discrimination of active and inactive COX-2 inhibitors of this general structure (Figure 1) can be obtained quickly and successfully.

An examination of misclassifications relative to compound series indicated that the most poorly classified set of compounds (relative to the number of compounds in the data set) had the pyrazole-based scaffold. Of the 10 misclassifications in the prediction set, five were from this series of compounds. Interestingly, compound **33** (Figure 4), which predicted poorly in the QSAR analysis, also proved problematic in the classification experiments. Table 6 includes the breakdown of misclassifications in the training and prediction sets for all four compound series.

Monte Carlo Experiments. The Monte Carlo experiments for classification analysis were conducted in an analogous fashion to the QSAR analysis. The class labels for active and inactive compounds were scrambled 10 times. Searches for the best nine-descriptor models were conducted for each of the 10 scrambled dependent variables as described for the normal classification experiments. The percent correct classifications for the 10 models ranged from 63.14% to 70.90% (67.65% average) for the training set and from 40.0% to 60.0% (49.99% average) for the prediction set. Due to the slightly uneven distribution of active and inactive compounds in the training set, a random assignment of classes would be 50.06%. With an average prediction set classification rate of approximately 50% over 10 different scrambled models, it can be concluded that chance effects played a limited role, if any, in the formation of models for classifying the compounds based on their true activity.

CONCLUSIONS

Successful QSAR and classification models have been developed for a diverse set of specific COX-2 inhibitors using only topological descriptors. An extensive descriptor space was screened using a linear regression fitness evaluator to find an eight-descriptor model which when passed to a CNN was able to accurately estimate the experimental inhibitory concentrations of 273 potential inhibitors. This application also demonstrates the benefit of using a committee of CNNs for applications in which nonlinear relationships exist and where a high degree of uncertainty and variance may be present in the data. Furthermore, randomization experiments were conducted to validate that the predictions made by these models were likely not due to chance correlations.

In addition, a nine-descriptor model was generated for the classification of active and inactive COX-2 inhibitors using solely topological information. For this, a *k*NN classifier was used which provides the benefit of simplicity and speed, two factors which are key in the realm of high throughput screening and lead optimization of drug compounds. Despite the high number of extremely inactive compounds being incorrectly classified as active, for the purposes of drug screening, a false positive classification is preferable over a false negative even if the compound fails to show activity later in the screening process. As with the QSAR work, Monte Carlo scrambling experiments were performed which verified that sound results were obtained using this model.

It should be noted that models were also developed during the course of this work which employed geometry-dependent descriptors. One limitation, and shortcoming, of using this information is that the lowest energy conformations were calculated in vacuo. The results of this may be somewhat misleading since flexible drug molecules generally will be oriented with the active site such that they may not be in the lowest energy conformation. Furthermore, other factors such as water and larger biomolecules in the binding environment will invariably alter the conformation of the inhibitor compound. In any event, the use of geometry-based descriptors provided little improvement in the rms errors of the models over those presented in this paper using solely topological information. With the exception of a few compound outliers, these topological descriptors have proven to be sufficiently information-rich to delineate minute structural variations across a large data set for quantitation and classification. While the calculated descriptors used in our models provide limited insight into the major structural contributions to potent COX-2 inhibition, this work clearly demonstrates that they can be used as a screening tool for larger libraries of potential target compounds which share a similar core scaffold.

Supporting Information Available: A table of all 314 compound numbers, structures, and respective log(IC_{50}) values. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Berenbaum, F. Selective Cyclooxygenase-2 Inhibitors: Hope and Facts. *Joint Bone Spine* **2000**, *67*, 499−501.

(2) Buttar, N. S.; Wang, K. K. The "Aspirin" of the New Millenium: Cyclooxygenase-2 Inhibitors. *Mayo Clin. Proc.* **2000**, *75*, 1027−1038.

(3) Needleman, P.; Isakson, P. C. The Discovery and Function of COX-2. *J. Rheumatol.* **1997**, *24*, 6−8.

(4) Saag, K.; van der Heidje, D.; Fischer, C.; Samara, A.; DeTora, L.; Bolognese, J.; Sperling, R.; Daniels, B. Rofecoxib, a New Cyclooxygenase-2 Inhibitor, Shows Sustained Efficacy, Comparable With Other Non-Steroidal Antiinflammatory Drugs − A 6-Week and a 1-Year Trial in Patients With Osteoarthritis. *Arch. Fam. Med.* **2000**, *9*, 1124−1134.

(5) McGettigan, P.; Henry, D. Current Problems With Nonspecific COX Inhibitors. *Curr. Pharm. Design* **2000**, *6*, 1693−1724.

(6) Venturini, C. M.; Isakson, P.; Needleman, P. Non-Steroidal Anti-inflammatory Drug-induced Renal Failure: A Brief Review of the Role of Cyclooxygenase Isoforms. *Curr. Opin. Nephrol. Hy.* **1998**, *7*, 79−82.

(7) Gilroy, D. W.; Colville-Nash, P. R.; Willis, D.; Chivers, W. S.; Paul-Clarke, M. J.; Willoughby, D. A. Inducible Cyclooxygenase May Have Antiinflammatory Properties. *Nat. Med.* **1999**, *5*, 698−701.

(8) Giercksky, K. E.; Haglund, U.; Rask-Madsun, J. Selective Inhibitors of COX-2 − Are They Safe For the Stomach? *Scand. J. Gastroentero.* **2000**, *35*, 1121−1124.

(9) Desiraju, G. R.; Gopalakrishnan, B.; Jetti, R. K. R.; Raveendra, D.; Sarma, J. A. R. P.; Subramanya, H. S. Three-dimensional Quantitative Structural Activity Relationship (3D-QSAR) Studies of Some 1,5-diarylpyrazoles: Analogue-Based Design of Selective Cyclooxygenase-2 Inhibitors. *Molecules* **2000**, *7*, 945−955.

(10) Marot, C.; Chavette, P.; Lesieur, D. Comparative Molecular Field Analysis of Selective Cyclooxygenase-2 (COX-2) Inhibitors. *Quant. Struct.-Act. Relat.* **2000**, *19*, 127−134.

(11) Zoete, V.; Bailly, F.; Maglia, F.; Rougee, M.; Bensasson, R. V. Molecular Orbital Theory Applied to the Study of Nonsteroidal Antiinflammatory Drug Efficiency. *Free Radical Bio. Med.* **1999**, *26*, 1261−1266.

(12) Singh, P.; Kumar, R. Novel Inhibitors of Cyclooxygenase-2: The Sulfones and Sulfonamides of 1,2-diaryl-4,5-difluorobenzene. Analysis of Quantitative Structure−Activity Relationship. *J. Enzymol. Inhib.* **1998**, *13*, 409−417.

(13) Singh, P.; Kumar, R. 1,2-Diarylimidazoles as Inhibitors of Cyclooxygenase-2: A Quantitative Structure−Activity Relationship Study. *J. Enzymol. Inhib.* **1999**, *14*, 277−288.

(14) Kumar, R.; Singh, P. Diarylspiro[2.4]heptenes as Selective Cyclooxygenase-2 Inhibitors: A Quantitative Structure−Activity Relationship Analysis. *Indian J. Chem. B* **1997**, *36*, 1164−1168.

(15) Hadjipavlou-Litna, D. Quantitative Structure−Activity Relationship (QSAR) Studies On Non-Steroidal Antiinflammatory Drugs (NSAIDs). *Curr. Med. Chem.* **2000**, *7*, 375−388.

(16) Galvez, J.; R., G.-D.; Gomez-Lechon, M. J.; Castell, J. V. Use of Molecular Topology in the Selection of New Cytostatic Drugs. *J. Mol. Struct.-Theochem* **2000**, *504*, 241−248.

(17) Jaen-Oltra, J.; Salabert-Salvador, T.; Garcia-March, F. J.; Perez-Giminez, F.; Thomas-Vert, F. Artificial Neural Network Applied to Prediction of Fluoroquinolone Antibacterial Activity by Topological Methods. *J. Med. Chem.* **2000**, *43*, 1143−1148.

(18) Hoffman, B.; Cho, S. J.; Zheng, W. F.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative Structure−Activity Relationship Modeling of Dopamine D-1 Antagonists Using Comparative Molecular Field Analysis, Genetic Algorithms-Partial Least-Squares, and K Nearest Neighbor Methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.

(19) 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.

(20) Gozalbes, R.; Galvez, J.; R., G.-D.; Derouin, F. Molecular Search of New Active Drugs Against *Taxoplasma Gondii*. *SAR QSAR Environ. Res.* **1999**, *10*, 47−60.

(21) Dixon, S. L.; Villar, H. O. Investigation of Classification Methods For the Prediction of Activity in Diverse Chemical Libraries. *J. Comput. Aid. Mol. Des.* **1999**, *13*, 533−545.

(22) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Doctor, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)-1*H*-pyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* **1997**, *40*, 1347−1365.

(23) Prasit, P.; Wang, Z.; Brideau, C.; Chan, C.-C.; Charleson, S.; Cromlish, W.; Ethier, D.; Evans, J. F.; Ford-Hutchinson, A. W.; Gauthier, J. Y.; Gordon, R.; Guay, J.; Gresser, M.; Kargman, S.; Kennedy, B.; Leblanc, Y.; Leger, S.; Mancini, J.; O'Neill, G. P.; Ouellet, M.; Percival, M. D.; Perrier, H.; Reindeau, D.; Rodger, I.; Tagari, P.; Therian, M.; Vickers, P.; Wong, E.; Xu, L.-J.; Young, R. N.; Zamboni, R.; Boyce, S.; Rupniak, N.; Forrest, M.; Visco, D.; Patrick, D. The Discovery of Rofecoxib. [MK-996, Vioxx, 4-(4′-Methylsulfonylphenyl)-3-Phenyl-2(5H)-Furanone]. An Orally Active Cyclooxygenase-2 Inhibitor. *Biorg. Med. Chem. Lett.* **1999**, *9*, 1773−1778.

(24) Khanna, I. K.; Weier, R. M.; Yu, Y.; Collins, P. W.; Miyashiro, J. M.; Koboldt, C. M.; Veenhuizen, A. W.; Currie, J. L.; Seibert, K.; Isakson, P. C. 1,2-Diarylpyrroles as Potent and Selective Inhibitors of Cyclooxygenase-2. *J. Med. Chem.* **1997**, *40*, 1619−1633.

(25) Khanna, I. K.; Weier, R. M.; Yu, Y.; Xu, X. D.; Koszyk, F. J.; Collins, P. W.; Koboldt, C. M.; Veenhuizen, A. W.; Perkins, W. E.; Casler, J. J.; Masferrer, J. L.; Zhang, Y. Y.; Gregory, S. A.; Seibert, K.; Isakson, P. C. 1,2-Diarylimidazoles as Potent, Cyclooxygenase-2 Selective, and Orally Active Antiiflammatory Agents. *J. Med. Chem.* **1997**, *40*, 1634−1647.

(26) Khanna, I. K.; Yu, Y.; Huff, R. M.; Weier, R. M.; Xu, X.; Koszyk, F. J.; Collins, P. W.; Cogburn, J. N.; Isakson, P. C.; Koboldt, C. M.; Masferrer, J. L.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Yuan, J.; Yang, D.; Zhang, Y. Y. Selective Cyclooxygenase-2 Inhibitors: Heteroaryl Modified 1,2-Diarylimidazoles Are Potent, Orally Active Antiinflammatory Agents. *J. Med. Chem.* **2000**, *43*, 3168−3185.

(27) Li, J. J.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Collins, J. T.; Garland, D. J.; Gregory, S. A.; Huang, H.; Isakson, P. C.; Koboldt, C. M.; Logusch, E. W.; Norton, M. B.; Perkins, W. E.; Reinhard, E. J.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y.; Reitz, D. B. 1,2-Diarylcyclopentenes as Selective Cyclooxygenase-2 Inhibitors and Orally Active Antiinflammatory Agents. *J. Med. Chem.* **1995**, *38*, 4570−4578.

(28) Gierse, J. K.; Hauser, S. D.; Creely, D. P.; Koboldt, C.; Rangwala, S. H.; Isakson, P. C.; Seibert, K. Expression and Selective Inhibtion of the Constitutive and Inducible Forms of Human Cyclo-oxygenase. *Biochem. J.* **1995**, *305*, 479−484.

(29) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; Americal Chemical Society: Washington, DC, 1979.

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

(31) 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.

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

(33) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.

(34) Eldred, D. V.; Weikel, C. L.; Jurs, P. C.; Kaiser, K. L. E. Prediction of Fathead Minnow Acute Toxicity of Organic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **1999**, *12*, 670−678.

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

(36) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(37) Xu, L.; 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.

(38) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.

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

(40) Hall, L. H.; Kier, L. B. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure−Property Modeling*; VCH Publishers: New York, 1991.

(41) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for All Self-Avoiding Paths for Molecular Graphs. *Computers Chem.* **1979**, *3*, 5−13.

(42) 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.

(43) Madan, A. K.; Gupta, S.; Singh, M. Superpendentic Index: A Novel Highly Discriminating Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 272−277.

(44) Kier, L. B.; Hall, L. H. The E-State as an Extended Free-Valence. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 548−552.

(45) 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.

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

(47) Broyden, C. G. The Convergence of a Class of Double-Rank Minimization Algorithms. *J. Inst. Math. Appl.* **1970**, *6*, 222−231.

(48) Fletcher, R. A New Approach to Variable Metric Algorithms. *Comput. J.* **1970**, *13*, 317−322.

(49) Goldfarb, D. A Family of Variable-metric Methods Derived by Variational Means. *Math. Comput.* **1970**, *24*, 23−26.

(50) Shanno, D. F. Conditioning of quasi-Newton Methods for Function Minimization. *Math. Comput.* **1970**, *24*, 647−656.

CI010073H