# Prediction of Inhibition of the Sodium Ion−Proton Antiporter by Benzoylguanidine Derivatives from Molecular Structure

Gregory W. Kauffman and Peter C. Jurs*

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

Received October 5, 1999

The use of quantitative structure−activity relationships to predict $IC_{50}$ values of 113 potential $Na^+/H^+$ antiporter inhibitors is reported. Multiple linear regression and computational neural networks (CNNs) are used to develop models using a set of information-rich descriptors. The descriptors encode information about topology, geometry, electronics, and combination hybrids. A five-descriptor CNN model with root-mean-square (rms) errors of 0.278 log units for the training set and 0.377 log units for the prediction set was developed. Examination of data set subclasses showed that systematic structural variations were also well-encoded resulting in 100% accuracy of prediction trends. An experiment involving a committee of five CNNs was also performed to examine the effect of network output averaging. This showed improved results decreasing the training and cross-validation set rms error to 0.228 log units and the prediction set rms error to 0.296 log units.

## INTRODUCTION

Heart disease remains the leading cause of death, contributing to approximately 30% of the total mortality rate in the United States.[1] Various forms of cardiovascular disease, including high blood pressure, cardiac ischemia, and angina pectoris may not initially be fatal; however, severity often elevates to more life-threatening conditions. This has spawned multimillion dollar research efforts to find cures for these diseases over the past several decades. Despite tremendous advances, this area of research has remained a high priority among the science community.

The NHE−1 $Na^+/H^+$ antiporter, which exchanges intracellular $H^+$ with extracellular $Na^+$, plays a vital role in various biological functions, including regulation of cellular volume[2] and intracellular pH.[3] This exchanger belongs to a family of protein isoforms, NHE−1 to NHE−4, which are expressed within specific organs. The NHE−1 isoform, however, is the only one expressed throughout the entire body, and particularly in the heart.[4,5] Selective inhibition of this isoform has been proven to have implications as a cardioprotective agent.[6] The selectivity of inhibition is required to avoid undesirable side effects, such as gastrointestinal or renal complications, which are a result of inhibition of the NHE−2 and NHE−3 antiporters, respectively.

Screening libraries of compounds for potential drug candidates that possess superior selectivity and potency can be an expensive and time-consuming process. To that end, the use of computational methods in drug design is becoming an increasingly useful tool.[7−9] Quantitative structure−activity relationships (QSARs) have successfully been used to predict a variety of biological activities including human intestinal absorption,[10] binding affinity,[11,12] and toxicity.[13,14] Furthermore, QSAR inhibition studies for an array of biological functions have been reported. These include MAO[15] and acetylcholinesterase[16] inhibitors, HIV-1 reverse transcriptase inhibitors,[17] and multiple drug resistance modulators.[18,19] These methods help reduce experimental time and may help elucidate structural features of a compound class that enhance the property of interest.

This paper presents the results from a quantitative structure−activity relationship study aiming to find a correlation between molecular structure of a class of 113 benzoylguanidine derivatives and their respective logarithmic inhibitory concentration ($IC_{50}$) values. Important structural features of each drug candidate are encoded by numerical descriptors. Small subsets of information-rich descriptors are selected to form models that may best predict the activity of interest. Multiple linear regression (MLR) analysis and computational neural networks (CNNs) are employed to evaluate the ability of a model to accurately relate the molecular structure and $\log(IC_{50})$ values.

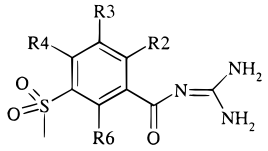
## EXPERIMENTAL SECTION

The 113 benzoylguanidine derivatives used for this study were taken from the literature.[20−25] A complete list of all compounds used in this study and their respective $\log(IC_{50})$ values are presented in Table 1. The reported $IC_{50}$ values range from 2 to 9500 nM. The standard deviation for each $IC_{50}$ value (measured in triplicate) was reported as ±14% on average. $\text{Log}(IC_{50})$ was used as the dependent variable in this study, and it ranged from 0.301 to 3.978. In vitro antiporter inhibition was measured as the uptake of $^{22}Na^+$ into rabbit red blood cells by following the $\beta$-decay with a scintillation counter. Details of experimental conditions and assays used for data collection may be found in ref 20.

The 113-member data set was randomly split into a 102-member training set and an 11-member external prediction set for development of linear models. Models were evaluated by statistical strength and minimization of the training set

10.1021/ci9901237 CCC: $19.00 
Published on Web 03/07/2000

**Table 1.** The 113 Benzoylguanidine Derivatives Used in This Study




| compd no. | $R_2$ | $R_3$ | $R_4$ | $R_6$ | observed log($IC_{50}$) | MLR model | CNN model[a] | ref |
|---|---|---|---|---|---|---|---|---|
| 1 | Me | | F | | 1.643 | 1.979 | 1.667 | 20 |
| 2[c] | Me | | Cl | | 1.398 | 1.888 | 1.512 | 20 |
| 3 | Me | | Br | | 1.079 | 1.544 | 1.434 | 20 |
| 4 | Et | | Cl | | 1.799 | 2.084 | 1.490 | 20 |
| 5 | Et | | Br | | 1.613 | 1.715 | 1.451 | 20 |
| 6 | F | | F | | 2.799 | 1.983 | 2.707 | 20 |
| 7 | Cl | | Cl | | 2.146 | 2.127 | 2.074 | 20 |
| 8 | Me | Cl | Cl | | 1.462 | 1.900 | 1.259 | 20 |
| 9 | Me | Me | Cl | | 1.763 | 1.525 | 1.805 | 20 |
| 10 | Me | $NO_2$ | Cl | | 1.643 | 2.495 | 1.323 | 20 |
| 11 | Me | 1-pyrrolyl | Cl | | 1.763 | 1.493 | 1.455 | 20 |
| 12[b] | H | | Cl | Me | 3.114 | 1.855 | 2.145 | 20 |
| 13 | $NH_2$ | | Cl | | 2.114 | 2.302 | 2.253 | 20 |
| 14 | $CH_2NEt_2$ | | Br | | 3.978 | 2.965 | 3.963 | 20 |
| 15 | Me | | Me | | 1.041 | 1.677 | 1.181 | 20 |
| 16[c] | Me | | Et | | 0.681 | 1.423 | 0.974 | 20 |
| 17 | Me | | Pr | | 0.756 | 1.209 | 1.012 | 20 |
| 18 | Me | | i-Pr | | 0.301 | 1.193 | 0.811 | 20 |
| 19 | Me | | 2-Bu | | 0.903 | 1.228 | 0.854 | 20 |
| 20[b] | Me | | *n*-pentyl | | 1.431 | 1.286 | 1.084 | 20 |
| 21 | Me | | cyclobutyl | | 0.914 | 1.782 | 1.505 | 20 |
| 22 | Me | | cyclohexyl | | 1.342 | 1.729 | 1.482 | 20 |
| 23[c] | Me | Me | Me | | 1.740 | 1.246 | 1.071 | 20 |
| 24 | Br | | Me | | 1.845 | 1.695 | 1.695 | 20 |
| 25 | H | | *p*-tolyl | | 2.000 | 1.531 | 1.451 | 20 |
| 26 | Me | | *p*-tolyl | | 1.398 | 1.376 | 1.337 | 20 |
| 27 | Me | | 4-chlorophenyl | | 1.653 | 1.716 | 1.495 | 20 |
| 28 | Me | | 2,4-dichlorophenyl | | 1.623 | 1.351 | 1.959 | 20 |
| 29 | Me | | 4-fluorophenyl | | 1.556 | 1.862 | 1.386 | 20 |
| 30 | Me | | 3,5-bis(trifluoromethyl) | | 3.447 | 2.598 | 3.401 | 20 |
| 31 | Me | | 3,5-dichlorophenyl | | 2.114 | 2.450 | 1.976 | 20 |
| 32 | Me | | phenyl | | 1.279 | 1.577 | 1.294 | 20 |
| 33 | Me | | 2-furyl | | 1.204 | 1.520 | 1.064 | 20 |
| 34 | Me | | SMe | | 0.778 | 1.198 | 0.954 | 20 |
| 35 | Me | | SEt | | 0.771 | 0.986 | 1.010 | 20 |
| 36[b] | Me | | SPr | | 0.833 | 0.903 | 1.068 | 20 |
| 37[b] | Me | | S−i-Pr | | 1.079 | 0.945 | 0.885 | 20 |
| 38 | Me | | S−t-Bu | | 1.342 | 0.756 | 1.405 | 20 |
| 39 | Me | | S−(3-chlorophenyl) | | 1.568 | 1.214 | 1.408 | 20 |
| 40 | Me | | S−(2-chlorophenyl) | | 1.146 | 0.883 | 1.234 | 20 |
| 41[b] | Me | | S−(4-chlorophenyl) | | 1.279 | 1.346 | 1.373 | 20 |
| 42[c] | Me | | S−Ph | | 1.146 | 1.149 | 1.576 | 20 |
| 43 | Me | | S−(4-pyridyl) | | 1.114 | 1.957 | 1.342 | 20 |
| 44 | Me | | OBn | | 1.447 | 1.706 | 1.276 | 20 |
| 45 | Me | | OH | | 2.342 | 1.805 | 2.155 | 20 |
| 46[c] | Me | | OMe | | 1.204 | 1.651 | 1.365 | 20 |
| 48 | Me | | OEt | | 1.279 | 1.498 | 1.399 | 20 |
| 49 | Me | | O−i-Pr | | 1.279 | 1.204 | 0.868 | 20 |
| 50 | Me | | O-cyclopentyl | | 1.431 | 1.701 | 1.438 | 20 |
| 51 | Me | | O-cyclohexyl | | 1.568 | 1.962 | 1.240 | 20 |
| 52 | Me | | O−t-Bu | | 0.973 | 1.098 | 0.787 | 20 |
| 53[c] | Me | | OPh | | 1.000 | 1.698 | 1.537 | 20 |
| 54 | Me | | O−(2-chlorophenyl) | | 1.146 | 1.162 | 1.115 | 20 |
| 55[b] | Me | | O−(3-chlorophenyl) | | 1.322 | 1.094 | 1.302 | 20 |
| 56 | Me | | O−(4-chlorophenyl) | | 1.114 | 1.196 | 1.223 | 20 |
| 57 | Cl | | OMe | | 1.602 | 2.088 | 1.873 | 20 |
| 58 | Cl | | O−(3-pyridyl) | | 1.716 | 1.696 | 1.897 | 20 |
| 59 | Me | Me | O−(3-pyridyl) | | 1.362 | 1.236 | 1.347 | 20 |
| 60 | H | | O−(3-pyridyl) | Me | 3.342 | 1.715 | 3.412 | 20 |
| 61 | Me | | O-pyrrolidinyl | | 1.041 | 1.373 | 1.035 | 20 |
| 62[b] | Me | | 1-piperidinyl | | 1.322 | 1.707 | 1.157 | 20 |
| 63 | Me | | 3-hydroxy-1-piperidinyl | | 1.477 | 1.597 | 1.330 | 20 |
| 64 | H | | 3-hydroxy-1-piperidinyl | | 1.813 | 1.817 | 1.609 | 20 |
| 65 | $NH_2$ | | 1-piperidinyl | | 2.279 | 2.073 | 2.198 | 20 |
| 66 | $NH_2$ | | 3-hydroxy-1-piperidinyl | | 2.204 | 2.135 | 2.316 | 20 |
| 67 | 1-piperidinyl | | 1-piperidinyl | | 3.000 | 3.136 | 2.962 | 20 |

**Table 1.** (Continued)

| compd no. | $R_2$ | $R_3$ | $R_4$ | $R_6$ | observed log($IC_{50}$) | MLR model | CNN model[a] | ref |
|---|---|---|---|---|---|---|---|---|
| 68 | Me | | 4-$NH_2$−1-piperidinyl | | 0.991 | 1.394 | 1.265 | 20 |
| 69 | OMe | | 4-$NH_2$−1-piperidinyl | | 1.863 | 1.482 | 1.506 | 20 |
| 70 | Cl | | 4-$NH_2$−1-piperidinyl | | 1.301 | 1.464 | 1.571 | 20 |
| 71 | Cl | | 3-hydroxy-1-piperidinyl | | 1.778 | 1.968 | 2.071 | 20 |
| 72 | Me | | 4-(2-pyrimidyl)piperazinyl | | 1.431 | 1.286 | 1.319 | 20 |
| 73 | Me | | 1-imidazolyl | | 1.362 | 1.884 | 2.039 | 20 |
| 74 | Me | | 1-benzimidazolyl | | 2.000 | 1.476 | 1.980 | 20 |
| 75 | Me | | 2-methyl-1-imidazolyl | | 1.462 | 1.773 | 1.935 | 20 |
| 76[b] | H | | 1-pyrazolyl | | 2.204 | 1.840 | 2.557 | 20 |
| 77 | Cl | | 1-imidazolyl | | 2.114 | 2.213 | 2.268 | 20 |
| 78 | Me | Me | 1-imidazolyl | | 2.462 | 1.663 | 2.350 | 20 |
| 79 | Me | | 4-methyl-1-imidazolyl | | 1.447 | 1.640 | 1.940 | 20 |
| 80[c] | Me | | 1-pyrazolyl | | 1.398 | 1.631 | 1.629 | 20 |
| 81[b] | H | | 1-imidazolyl | | 2.176 | 2.115 | 2.369 | 20 |
| 82 | Me | | 1-pyrrolyl | | 0.903 | 1.827 | 1.628 | 20 |
| 83 | Me | | $NH_2$ | | 1.204 | 1.564 | 1.490 | 20 |
| 84[c] | Me | | NHPh | | 1.322 | 1.538 | 1.698 | 20 |
| 85 | Me | | NH-2-pyridyl | | 0.964 | 1.227 | 1.104 | 20 |
| 86 | Me | | NH−(3-chlorophenyl) | | 1.519 | 1.767 | 1.336 | 20 |
| 87 | Me | | NH-2-pyrimidyl | | 0.919 | 1.306 | 1.345 | 20 |
| 88 | Cl | | NH-2-pyrimidyl | | 1.690 | 1.306 | 1.345 | 20 |
| 89 | Me | 1-pyrrolyl | H | | 2.000 | 1.602 | 1.846 | 20 |
| 90 | H | | CN | | 2.204 | 2.231 | 1.993 | 20 |
| 91 | Me | | CN | | 1.845 | 2.319 | 1.681 | 20 |
| 92 | Me | Me | CN | | 2.568 | 2.014 | 1.920 | 20 |
| 93 | Me | | $CF_3$ | | 1.568 | 1.920 | 1.715 | 20 |
| 94 | H | | $CF_3$ | | 2.146 | 2.061 | 2.471 | 20 |
| 95 | H | | OMe | | 2.431 | 1.637 | 2.301 | 21 |
| 96 | H | | O−(3-pyridyl) | | 1.881 | 1.811 | 2.056 | 22 |
| 97 | H | | cyclohexyl | | 2.531 | 1.635 | 2.427 | 23 |
| 98[c] | H | | Me | | 2.322 | 1.564 | 2.304 | 21 |
| 99 | H | | Cl | | 2.580 | 1.886 | 2.428 | 24 |
| 100 | H | | Br | | 2.255 | 1.743 | 2.291 | 21 |
| 101[b] | Me | | H | | 2.176 | 1.740 | 2.128 | 25 |
| 102 | H | | 1-piperidinyl | | 2.114 | 1.826 | 2.169 | 24 |
| 103 | H | | i-Pr | | 1.415 | 1.154 | 1.396 | 21 |
| 104 | Et | | O−(3-piperidinyl) | | 1.431 | 1.812 | 1.578 | 20 |
| 105[c] | Me | | O−(3-piperidinyl) | | 0.944 | 1.849 | 1.412 | 20 |
| 106[b] | Me | | O-5-(3-chloro)piperidinyl | | 1.146 | 1.823 | 1.444 | 20 |
| 107 | H | | O-5-(3-chloro)piperidinyl | | 1.875 | 1.722 | 1.611 | 20 |
| 108 | H | | 1-(4-amino)piperizinyl | | 2.204 | 1.512 | 1.961 | 20 |
| 109[c] | Et | | 1-piperizyinyl | | 1.398 | 1.658 | 1.720 | 20 |
| 110 | Et | | 1-(3-hydroxy)piperizinyl | | 1.613 | 1.729 | 1.510 | 20 |
| 111 | H | | N−(4-pyridinol) | | 2.851 | 2.599 | 2.415 | 20 |
| 112 | Et | | N−(4-pyridinol) | | 2.968 | 2.350 | 1.969 | 20 |
| 113 | Me | | N−(4-pyridinol) | | 1.875 | 2.396 | 2.164 | 20 |

[a] Type III model with 5−4−1 architecture. [b] Member of cross-validation set. [c] Member of external prediction set.

root-mean-square (rms) error. Validation of the best model was then performed using the prediction set, which was not used during descriptor selection or model development.

Nonlinear models were built using computational neural networks. A cross-validation set of 11 compounds was randomly removed from the 102-member training set used for linear model generation, thus generating a 91-member training set, an 11-member cross-validation set, and an 11-member prediction set. Cross-validation is used in model building to prevent overtraining of the CNN. Its use will be further elaborated upon in the discussion section.

All computations were performed on a DEC 3000 AXP model 500 workstation running the Unix operating system. The automated data analysis and pattern recognition toolkit (ADAPT) software package,[26,27] simulated annealing,[28] genetic algorithm,[29] and CNN routines,[30] written independently at Penn State, were used to develop the QSAR models for this study. QSAR development can be described in four steps: structure entry and optimization, descriptor generation and objective feature selection, linear model formation and validation, and nonlinear model formation and validation.

**Structure Entry and Optimization.** All compounds were sketched and preliminary models were built using HyperChem (Hypercube, Inc., Waterloo, ON) on a Pentium PC. This provided information on atom types and bonding arrangements to be used for geometry optimization. Accurate three-dimensional representations of all structures were obtained by the semiempirical molecular orbital package MOPAC[31] using the PM3 Hamiltonian.[32] Charge information for descriptor generation was obtained using the AM1 Hamiltonian.[33]

**Descriptor Generation and Objective Feature Selection.** The descriptors calculated using ADAPT encode important structural information about topology, geometry, electronics, and combination hydrids. Topological descriptors describe molecular connectivity[34,35] features, such as shape,[36] branching,[37] specific atom counts, and hybridization. Geometric descriptors include information such as moments of inertia

**Table 2.** Descriptors Used for the Type I Multiple Linear Regression Model

| descriptor | type | coefficient | error | range | explanation[a] |
|---|---|---|---|---|---|
| constant | | −5.895 | 1.018 | | |
| MOLC-5 | top. | 0.590 | 0.132 | 2.05−5.50 | Weighted 3rd order paths |
| MOLC-8 | top. | −1.519 | 0.332 | 0.96−2.07 | Weighted 4th-order path/clusters |
| GEOM-2 | geo. | 0.257 | $6.4 \times 10^{-2}$ | 1.56−5.14 | 2nd major moment |
| SHDW-4 | geo. | 11.79 | 1.761 | 0.44−0.58 | normalized 2-D projection on the *XY* plane |
| DIPO-0 | elec. | 0.164 | $2.5 \times 10^{-2}$ | 3.47−14.4 | dipole moment |

[a] MOLC-5, molecular connectivity for all third-order paths;[36] MOLC-8, molecular connectivity for all fourth-order paths and clusters;[36] GEOM-2, the second major geometric moment of the molecule; SHDW-4, normalized two-dimensional shadow projection on the *XY* plane when the molecule is aligned with its first two moments of inertia along the *X* and *Y* axes;[46] DIPO-0, dipole moment of the molecule obtained from MOPAC[31,32] using the AM1 Hamiltonian.[33]

and solvent-accessible surface areas.[38,39] Electronic descriptors provide information such as homo and lumo energies, dipole moment, and partial charges.[40,41] Combination descriptors encode topological, electronic, and geometric information. Charged partial surface areas[42] and hydrogen-bonding interactions[43,44] are two examples of such descriptors. Calculation of topological descriptors requires no prior geometry optimization; however, geometric and electronic descriptors require minimum energy 3-D conformations for accurate calculations. A total of 140 topological, 30 geometric, 10 electronic, and 48 combination descriptors were calculated for the entire data set.

The 228-member pool of descriptors was submitted to objective feature selection routines to eliminate descriptors that contain little or redundant information. Descriptors with identical values for greater than 90% of the training set observations were removed. Additionally, any one of two descriptors that have pairwise correlations greater than 0.85 was eliminated from the pool. Finally, the mutual orthogonality among descriptors in the pool was maximized using the Gram−Schmidt method.[45] This generated a reduced pool of 61 descriptors for use in model development.

**Linear Feature Selection and Model Formation.** Subsets of descriptors were selected from the reduced pool using simulated annealing[28] feature selection with a multiple linear regression fitness evaluator for development of a type I linear model. The descriptor subsets were evaluated for their ability to link the property of interest to molecular structure by way of descriptor T-values and rms error. Subsets with low rms errors and T-values greater than four are favorable. Furthermore, subsets with fewer descriptors that did not compromise rms error were preferred.

Once the best model was chosen, diagnostics to test for the presence of compound outliers and multicollinearities between descriptors in the model were performed. A variance inflation factor ($VIF = 1/(1 - R^2)$) less than 10 indicated that the model contained no multicollinearities. Finally, the model was validated using the external prediction set and was saved for use in development of nonlinear models.

**Linear Feature Selection and Nonlinear Model Formation**. The descriptors that comprised the best type I model were used to generate a type II nonlinear model using a three-layer, fully connected, feed-forward computational neural network. Network training was achieved using the BFGS (Broyden−Fletcher−Goldfarb−Shanno) quasi-Newton method.[46] The three layers consist of an input layer, a hidden layer, and an output layer. The number of neurons in the input layer is determined by the number of descriptors from the type I model. The number of neurons in the hidden layer is appropriately adjusted such that the network neither overgeneralizes nor memorizes the training data, either of which could result in poor predictive ability. One precaution taken when adjusting the number of hidden neurons was to ensure that the ratio of observations (compounds in the training set) to adjustable parameters was greater than two. This reduced the risk of overtraining and chance correlations among training set data. Only one neuron is used in the output layer, which represents the learned response, or in this case, the activity of interest $\log(IC_{50})$. A variety of network architectures were examined to find the optimal number of hidden neurons required to minimize the training and cross-validation sets. Final validation of the model was done using the prediction set.

**Nonlinear Feature Selection and Model Formation.** The final model, type III, was developed using solely nonlinear methods. The reduced pool of 61 descriptors was submitted to a nonlinear feature selection routine that utilizes simulated annealing and a genetic algorithm with a neural-network fitness evaluator to find subsets of descriptors. The best subsets were used to train a CNN in an analogous fashion to that described for a type II model. Selection of an optimal network architecture was again based on training set and cross-validation set rms errors, and the external prediction set was used for validation.

## RESULTS AND DISCUSSION

**Type I.** Descriptor subsets with T-values greater than four were considered for type I model development. Sets ranging from 3 to 10 descriptors were examined, with only 1 5-descriptor set fulfilling the criterion. This subset of descriptors, listed in Table 2, was considered for further model development. No statistical outliers were found, and variance inflation factors were well below 10 for this model, indicating that no multicollinearities were present among the five descriptors. The model had a training set rms error of 0.473 log units ($r^2 = 0.455$). The prediction set rms error was 0.546 log units ($r^2 = 0.009$), showing that this linear model has no ability to generalize. Among the five descriptors, pairwise correlations ranged from 0.033 to 0.729 with an average value of 0.229.

Of the five descriptors used in the linear model, two were topological, two were geometric, and one was electronic. The two topological descriptors incorporate information on weighted path counts. MOLC-5 measures the degree of third-order path fragments, and MOLC-8 encodes fourth-order path/cluster fragments.[36] These descriptors are potentially encoding information about the level of branching, which

**Table 3.** Descriptors Used for the Nonlinear Type III CNN Model

| descriptor | type | range | explanation[a] |
|---|---|---|---|
| MDE−14 | top. | 1.77−12.7 | distance-edge between all 1° and 4° carbon atoms |
| GEOM-2 | geo. | 1.56−5.14 | the 2nd major moment |
| DIPO-0 | elec. | 3.47−14.4 | dipole moment |
| PNSA-3 | comb. | −119.0−69.0 | atomic charge weighted partial negative surface area |
| RNCS-1 | comb. | 2.21−6.25 | relative negative charged surface areas |

[a] MDE−14, molecular distance edge term between all primary and quaternary carbon atom pairs;[48] GEOM-2, the second major geometric moment of the molecule; DIPO-0, dipole moment of the molecule obtained from MOPAC[31,32] using the AM1 Hamiltonian;[33] PNSA-3, expressed by $\sum(-SA_i)(Q^-_i)$, where $-SA_i$ and $Q^-_i$ are the surface area and charge contributions of the *i*th negatively charged atom, respectively;[42] RNCS-1, expressed by $SA_{MNEG} \times (Q_{MNEG} \div \sum Q_{TNEG})$, where $SA_{MNEG}$, $Q_{MNEG}$, and $Q_{TNEG}$ are the surface area of the most negatively charged atom, the charge of the most negatively charged atom, and the total negative charge, respectively.[42]

in turn contributes to steric interactions of the molecule. GEOM-2 contains information about the second-major geometric moment of the molecules, and SHDW-4 represents the normalized two-dimensional projection of a molecule on the *XY* plane.[39] These descriptors encode information about the size, shape, and bulk properties of each molecule, again representing steric interactions between the target compound and the active site. DIPO-0 provides information about the dipole moment of the molecules. This most likely describes favorable electronic interactions of the drug candidate and active site.

**Type II.** The descriptors used to form a type I model were then used to develop a type II CNN model. Network training was directed by optimization of weights and biases and was terminated upon minimization of the cross-validation set rms error. It is at this minimum that the network ceases training and begins memorizing features of the training set data, thereby decreasing the ability to generalize to external data. All possible architectures ranging from 5−2−1 (15 adjustable parameters) to 5−6−1 (43 adjustable parameters)[47] were evaluated for training and prediction ability. A 5−4−1 network architecture (29 adjustable parameters) was found to provide the most flexibility.

As expected, the nonlinear capabilities of CNNs resulted in improved training and prediction rms errors compared to those produced using linear regression. The training set rms error improved to 0.363 log units ($r^2 = 0.677$), a 23% decrease from the type I model. The prediction set rms error was reduced by 23% to 0.418 log units ($r^2 = 0.438$). The cross-validation set rms error was also 0.418 log units ($r^2 = 0.697$). Successful use of a CNN for model development using linear feature selection suggests that nonlinear feature selection coupled with a CNN may produce a quality model with high predictive ability.

**Type III.** A final type III model was developed using a nonlinear feature selection routine described above. Again, all possible network architectures were tested using the top three subsets of descriptors identified by feature selection. A 5−4−1 architecture was determined to be optimal, using the five descriptors listed in Table 3. Of these descriptors, one was topological, one was geometric, one was electronic, and two were combination hybrids. Among the five descriptors, pairwise correlations ranged from 0.020 to 0.657 with an average value of 0.219. The descriptors GEOM-2 and DIPO-0 were the only two identified as information-rich using both linear and nonlinear feature selection routines. MDE-14 is a topological molecular distance edge descriptor that encodes information about distances between all primary and quaternary carbon atom.[48] The two combination descriptors belong to a family of charged partial surface area (CPSA) descriptors[42] that encode information about atomic partial charges and solvent-accessible surface areas. Details of atomic charge weighted partial negative surface areas are contained in the PNSA-3 descriptor. The RNCS-1 descriptor describes information about relative negative charged surface areas. Interestingly, both CPSA descriptors include information about negatively charged regions of the compounds' surfaces. This might imply a strong dependence on areas of higher electron density for good activity.

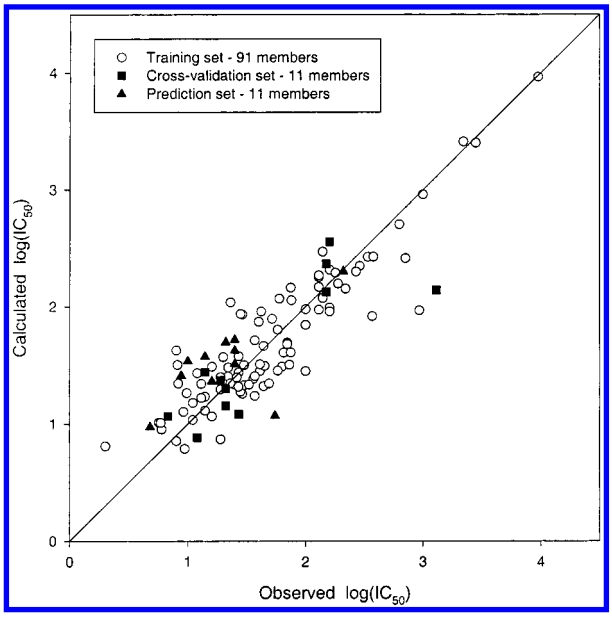



**Figure 1.** Plot of calculated/predicted versus observed log($IC_{50}$) for the training, cross-validation, and prediction set compounds used to develop the type III model described in Table 3. Data set members are listed in Table 1.

As expected, nonlinear feature selection and model development resulted in an overall improvement from the type II model. Training set rms error for the type III model was 0.278 log units ($r^2 = 0.812$), a 24% improvement over type II. An rms error of 0.362 log units ($r^2 = 0.697$) for the cross-validation set corresponded to a 13% improvement. Finally, prediction set rms error was 0.377 log units ($r^2 = 0.438$), a 10% improvement from the type II model. While the margin of rms error between the training and cross-validation sets for this model is greater than desired, the overall quality of the model can be realized by comparison to the type I and type II models. A plot of calculated/predicted versus observed log($IC_{50}$) values for the type III model is shown in Figure 1.

Visual inspection of Figure 1 revealed the presence of several compound outliers. Among these were one in the cross-validation set, one in the prediction set, and four in the training set. The cv set member (**12**) was one of only two compounds with substitution at the six-position. The prediction set member (**23**) was one of only a few with substitution at the three-position. More training data with

substitution at these positions may help improve these values. For the members of the training set (**73**, **82**, **112**), there is no obvious reasoning for the high deviation between calculated/predicted and experimental values. One training set member (**92**), though, did have substitution at the three-position. Again, a larger number of compounds with similar substitution patterns would help to better train the network to encode these minute structural features.

The only measure of error reported for the reference data is an average standard deviation of ±14% for each $IC_{50}$ value (measured in triplicate).[20] With only this information available, it is difficult to quantitatively compare the overall error associated with the results of this study with the error for the experimental data. Error lines demonstrating 14% deviation along the 1:1 correlation plot shown in Figure 1 were too narrow to provide valuable information for comparison. However, examination of our results shows that only 23% of the predicted values lie within 1 standard deviation of the experimental results. Despite this observation, the ability of the five-descriptor nonlinear model to accurately predict $\log(IC_{50})$ for a data set of such structural diversity is quite good. Clearly, improvements to this model can be made with the incorporation of more descriptors or the addition of more structures.

A minor concern of this study is the presence of three training set compounds (**14**, **30**, and **60**) whose experimental $\log(IC_{50})$ values are clearly larger than the bulk of the data. To monitor the influence of these three compounds on model development and assess whether their presence in model training artificially overrated the results, an additional experiment was performed that excluded their participation. The cross-validation set and prediction set members were held constant; however, the training set was reduced to 99 members for the type I model and 88 members for the type II and type III models. For the type I model, the training set had an rms error of 0.408 log units ($r^2 = 0.462$) and a prediction set rms error of 0.488 log units ($r^2 = 0.052$). The type II model produced rms errors of 0.274 log units ($r^2 = 0.745$), 0.345 log units ($r^2 = 0.467$), and 0.834 log units ($r^2 = 0.008$) for the training, cross-validation, and prediction sets, respectively. Finally, the type III nonlinear model resulted in rms errors of 0.265 log units ($r^2 = 0.762$) for the training set, 0.265 log units ($r^2 = 0.678$) for the cross-validation set, and 0.530 log units ($r^2 = 0.352$) for the prediction set. Compared to the models developed using the full 113-member data set, the training and cross-validation set rms errors in this experiment are marginally better. The prediction set error for the type I model was improved by removal of the three compounds, while the superior type II and type III models demonstrated substantially weaker predictive abilities. Furthermore, $r^2$ values of the prediction set for all three model types indicated the original experiment utilizing all 113 compounds produced better results overall.

A major concern with the development of QSARs is the risk of chance correlations between structure and activity. To ensure the models developed were not susceptible to chance effects, randomization experiments were performed for the type I and type III models. These experiments entail random scrambling of the dependent variable such that each structure is assigned a $\log(IC_{50})$ value from another member of the data set. If the QSAR model accurately predicts based upon systematic structural variations, then the correlation coefficient and rms errors for training and prediction will considerably worsen. For the type I model, a training set rms error of 0.538 log units ($r^2 = 0.210$) and a prediction set rms error of 0.818 log units ($r^2 = 0.050$) were produced. For the type III model, five individual scrambling experiments were conducted using varied training and cross-validation sets, however final validation was performed with a constant 11-member prediction set. The range of training set rms errors was 0.393−0.532 log units ($r^2 = 0.305-0.578$, average = 0.473). For the cross-validation set, rms errors ranged from 0.461 to 0.838 log units ($r^2 = 0.042-0.355$, average = 0.156). Finally, prediction set rms errors ranged from 0.494 to 0.938 log units ($r^2 = 0.0004-0.114$, average = 0.031). On the basis of these results, it is clear to see that the rms errors and correlation coefficients for these scrambling experiments are significantly worse than those of the type III CNN model developed with the real data. Thus, one can confidently conclude that the five-descriptor model developed was not based on chance. Finally, concerns about two of the CNN-based scrambling models producing lower rms errors and higher correlation coefficients than the type I linear model are not seen as meaningful because the two types of models were developed by entirely different methods (one linear and one highly nonlinear) and thus rms errors and correlation coefficients associated with them are not comparable to one another.

The quality of a QSAR model is dependent upon its ability to encode systematic structural variations within the entire training set and ultimately to utilize that information to account for similar trends in external data. Substituent effects on biological activity for several of the benzoylguanidine derivatives were analyzed in ref 20. Without exception, introduction of a methyl group at the $R_2$ position resulted in enhanced activity over the unsubstituted counterpart (**2** vs **99**, **3** vs **100**, **15** vs **98**, **18** vs **103**, **22** vs **97**, **26** vs **25**, **46** vs **95**, **62** vs **102**, **63** vs **64**, **68** vs **108**, **73** vs **81**, **80** vs **76**, **91** vs **90**, **93** vs **94**, **105** vs **96**, **106** vs **107**, **113** vs **111**). Examination of the predicted outputs for each pair of compounds in comparison with their corresponding experimental values verifies that the 5-descriptor type III model accounts for this subtle variation in structure for all 17 cases. Of the 34 compounds, 25 were training set members, 4 were cross-validation set members, and 5 were prediction set members.

Furthermore, ref 20 reported that compounds that were $R_2$-substituted with an ethyl group showed a reduced effect in comparison to the methylated derivative, but enhanced activity over the unsubstituted analogue (**4** vs **2** vs **99**, **5** vs **3** vs **100**, **47** vs **46** vs **95**, **104** vs **105** vs **96**, **109** vs **62** vs **102**, **110** vs **63** vs **64**, **112** vs **113** vs **111**). For the seven cases, the observed trend between ethyl and methyl substitution was correctly predicted five times and between the ethyl and unsubstituted compounds was correctly predicted six times. For the 21 examined, 17 were training set members, 1 was a cross-validation set member, and 4 were prediction set members. Finally, a limited comparison of two pairs of compound analogues substituted at either the $R_6$ or $R_2$ position indicated that substitution at the $R_6$ position led to a drastic reduction in potency (**2** vs **12**, **105** vs **60**). The type III model clearly predicted these trends well.

One final consideration of the data set was the mention of molar refraction as a classification criterion for good

**Table 4.** Committee CNN Rms Errors

| | training set | cross-validation set | prediction set |
|---|---|---|---|
| trial 1 | 0.278 | 0.362 | 0.377 |
| trial 2 | 0.274 | 0.337 | 0.422 |
| trial 3 | 0.299 | 0.310 | 0.283 |
| trial 4 | 0.319 | 0.300 | 0.514 |
| trial 5 | 0.356 | 0.355 | 0.309 |
| committee | | 0.228[a] | 0.296 |

[a] The rms error is combined owing to values of the training set and cross-validation varying from trial to trial.

inhibitors in ref 20. Benzoylguanidine derivatives with molar refraction values in the range of 10−26 (using a quadratic programming method)[49] were determined to possess the highest inhibitory effect experimentally. Interestingly enough, the ADAPT molar refraction descriptor (using Vogel's atomic contribution method)[50] was not selected as a member of the reduced descriptor pool during feature selection.

**Committee CNNs**. Improved results can often be achieved for problems in computational chemistry, such as the prediction of biological activity of complex molecules, by using better methods. Clearly, nonlinear methods of descriptor feature selection and model development surpass linear methods in quality. The next logical step is to improve the existing nonlinear methods to provide more accurate results. The approach of a committee of neural networks was considered to see if an averaging effect would improve the quality of the nonlinear type III model. The experiment was designed to include a committee of five neural networks, including the original type III CNN model. Furthermore, five different, randomly chosen training and cross-validation sets were used, holding only the external prediction set constant. Only the best five-descriptor subset identified by nonlinear feature selection was used in each trial. Each network was then trained in an identical fashion to the type III model described above, preserving the 5−4−1 architecture.

The rms errors for each CNN model are reported in Table 4. For the training set these values range from 0.274 to 0.356 log units, for the cross-validation set they range from 0.300 to 0.362 log units, and for the prediction set they range from 0.283 to 0.514 log units. Then, the predicted log($IC_{50}$) values for each compound from each of the five trials were averaged. These values were plotted against the observed log($IC_{50}$) value to generate the committee correlation plot shown in Figure 2. The rms error for the training and cross-validation committee data was 0.228 log units, an improvement over any one of the five individual CNNs. The prediction set rms error was 0.296 log units, a value comparable to the best prediction of the five trials.

The descriptors included in all five models varied nominally. The GEOM-2 and DIPO-0 descriptors described above were each in four of the five models. The distance edge descriptor, MDE-14, was included in three models; however, the similar MDE-13 descriptor also appeared in one of the models. Other descriptors that were incorporated into the five models on a more limited basis were carbon types, fragment descriptors, and shadow descriptors. Among the four additional five-descriptor models, the pairwise correlations ranged from 0.001 to 0.529 with average correlations for each of the four models being 0.120, 0.139, 0.144, and 0.223.

Finally, the correlation of the six compound outliers described earlier for the type III model were greatly

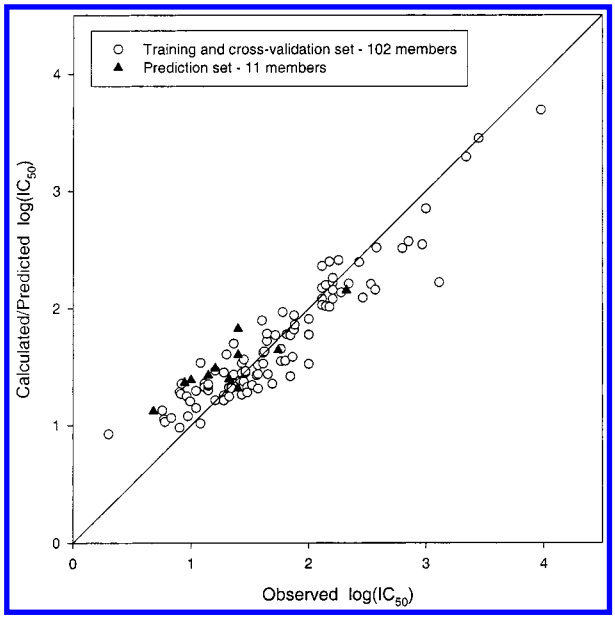



**Figure 2.** Plot of calculated/predicted versus observed log($IC_{50}$) for the training/cross-validation set and the prediction set for the committee CNN experiment. Calculated values are the average of the log($IC_{50}$) values from each of the five trials for each compound.

improved. This demonstrates the potential of committee neural networks to enhance predictive ability of QSAR models. In a real-world application of this methodology, one would chose the single model or committee of models to use for prediction without having access to any other measure of quality beyond the errors associated with the training sets and cross-validation sets. The committee CNN correlation plot shown in Figure 2 clearly illustrates an overall improvement of model quality over the individual trial shown in Figure 1.

## CONCLUSIONS

In this study, five-descriptor models were developed using multiple linear regression and computational neural network models. These models estimate the log($IC_{50}$) values for a class of 113 benzoylguanidine derivatives solely on the basis of descriptors derived from molecular structure. Table 2 lists the descriptors for the type I linear model, and Table 3 lists those identified in the type III nonlinear CNN model. The nonlinear type III model produced rms errors of 0.278 log units for the training set, 0.362 log units for the cross-validation set, and 0.377 log units for the prediction set. The descriptors used in these models suggest that steric interactions, dipole moment, and surface charges play an important role in the inhibitory effect of $Na^+/H^+$ antiporter mechanism for the benzoylguanidine class of compounds.

Models of this origin may potentially offer insight into the development of potent drugs. At a minimum, screening or design of compound libraries may be easily facilitated by development of a sound QSAR model. Demonstration of the nonlinear type III model to accurately generalize minute structural idiosyncrasies, such as methylated versus demethylated analogues, demonstrates the model developed in this QSAR study to be a useful one. Furthermore, use of a committee of neural networks can help improve the quality of a model by an averaging effect. The quality of such a model could be further improved by incorporation of a larger

data set, or perhaps by use of an entirely different set of compounds as the external prediction set, as long as compatible experimental conditions and assays are employed.

## REFERENCES AND NOTES


(1) Hoyert, D. L.; Kochanek, K. D.; Murphy, S. L. Deaths: Final Data for 1997. *Nat. Vital Statistics Rep.* **1999**, *47*, 1−105.

(2) Demaurex, N.; Grinstein, S. $Na^+/H^+$ antiport: modulation by ATP and role in cell volume regulation. *J. Exp. Biol.* **1994**, *196*, 389−404.

(3) Kaila, K.; Vaughn-Jones, R. P. Influence of sodium−hydrogen exchange on intracellular pH, sodium, and tension in sheep cardiac Purkinje fibres. *J. Physiol. (London)* **1987**, *390*, 93−118.

(4) Fliegel, L.; Frolich, O. The $Na^+/H^+$ exchanger: an update on structure, regulation and cardiac physiology. *Biochem. J.* **1993**, *296*, 273−285.

(5) Levitsky, J.; Gurell, D.; Frishman, W. H. Sodium Ion-Hydrogen Ion Exchange Inhibition: A New Pharacological Approach to Myocardial Ischemia and Reperfusion Injury. *J. Clin. Pharmacol.* **1998**, *38*, 887−897.

(6) Sholz, W.; Albus, U. Potential of Selective Sodium−Hydrogen Exchange Inhibitors in Cardiovascular Therapy. *Cardiovasc. Res.* **1995**, *29*, 184−188.

(7) *Neural Networks in QSAR and Drug Design*; Devillers, J., Ed.; Harcourt Brace: San Diego, CA, 1996.

(8) *Introduction to the Prinicples of Drug Design and Action*; Smith, J., Ed.; Harwood Academic: Amsterdam, 1998.

(9) *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997.

(10) 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.

(11) Robert, D.; Amat, L.; Carbo-Dorca, R. Three-Dimensional Quantitative Structure−Activity Relationships from Tuned Molecular Quantum Similarity Measures: Prediction of the Corticosteroid-Binding Globulin Binding Affinity for a Steriod Family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.

(12) Leonardi, A.; Motta, G.; Boi, C.; Testa, R.; Poggesi, E.; De Benedetti, P. G.; Menziani, M. C. Synthesis, Pharmacological Evaluation, and Structure−Activity Relationship Studies on Novel Derivatives of 6,7-Dimethoxyquinazoline 1-Adrenoceptor Antagonists. *J. Med. Chem.* **1999**, *42*, 427−437.

(13) Urrestarazu-Ramos, E.; Vaes, W. H. J.; Verhaar, H. J. M.; Hermens, J. L. M. Quantitative Structure−Activity Relationships for the Aquatic Toxicity of Polar and Nonpolar Narcotic Pollutants. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 845−852.

(14) 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.

(15) Medvedev, A. E.; Ivanov, A. S.; Veselovsky, A. V.; Skvortsov, V. S.; Archakov, A. I. QSAR Analysis of Indole Analogues as Monoamine Oxidase Inhibitors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 664−671.

(16) Hasegawa, K.; Kimura, T.; Funatsu, K. Application of GA-Based Region Selection to a 3D-QSAR Study of Acetylcholinesterase Inhibitors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 112−120.

(17) 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*, 4257−4264.

(18) Hiessböck, R.; Wolf, C.; Richter, E.; Hitzler, M.; Chiba, P.; Kratzel, M.; Ecker, G. Synthesis and in Vitro Multidrug Resistance Modulating Activity of a Series of Dihydrobenzopyrans and Tetrahydroquinolines. *J. Med. Chem.* **1999**, *42*, 1921−1926.

(19) Ecker, G.; Chiba, P. Structure−Activity Relationship Studies on Modulators of the Multidrug Transporter P-glycoprotein−An Overview. *Wien. Klin. Wochenschr.* **1995**, *107*, 681−686.

(20) Baumgarth, M.; Beier, N.; Gericke, R. (2-Methyl-5-(methylsulfonyl)-benzoyl)guanidine $Na^+/H^+$ Antiporter Inhibitors. *J. Med. Chem.* **1997**, *40*, 2017−2034.

(21) Lang, H. J.; Weichert, A.; Kleeman, H. W.; Englert, H.; Scholz, W.; Albus, U. Benzoylguanidine, Verfahren zu ihrer Herstellung, sowie ihre Verwedung als Antiarrhythmika. (Preparation of benzoylguanidines as drugs, e.g. antiarrythmic agents.) Eur. Pat. Appl. EP 589,336, 1994. *Chem. Abstr.* **1994**, *121*, 300597x.

(22) Lang, H. J.; Kleeman, H. W.; Scholz, W.; Albus, U. Substituierte Benzoylguanidine, Verfahren zu ihrer Herstellung, ihre Verwendung als Medikament oder Diagnostikum sowie sie enthaltendes Medikament. (Substituted benzoylguanidine cardiovascular and cardioprotective agents.) Eur. Pat. Appl. EP 602,523, 1994. *Chem. Abstr.* **1994**, *121*, 157133r.

(23) Weichert, A.; Lang, H. J.; Scholz, W.; Albus, U.; Lang, F. 3,4,5-Substituierte Benzoylguanidine, Verfahren zu ihrer Herstellung, ihre Verwendung als Medikament oser Diagnostikum sowie sie enthaltendes Medikament. (Preparation of benzoylguanidines as proton-sodium exchange inhibitors.) Eur. Pat. Appl. EP 577,024, 1994. *Chem. Abstr.* **1994**, *120*, 244385s.

(24) Englert, H. C.; Lang, H. J.; Linz, W.; Schölkens, B.; Scholz, W. Benzoylguanidine, Verfahren zu ihrer Herstellung, ihre Verwendung als Medikament sowie sie enthaltendes Medikament. (Preparation of benzoylguanidines as antiarrhythmics.) Eur. Pat. Appl. EP 416,499, 1991. *Chem. Abstr.* **1991**, *115*, 71158m.

(25) Lang, H. J.; Weichert, A.; Englert, H.; Scholz, W.; Albus, U.; Lang, F. Ortho-substituierte Benzoylguanidine, zur Behandlung und zur Prophylaxe von Herz-Rhythmus-Störungen und von ischämisch induzierten Schäden sowie zur Inhibition der Proliferation von Zellen. (Preparation of 3-(substituted sulfo)-benzoylguanidines as antiarrhythmics, antiischemics, and cell proliferation inhibitors.) Eur. Pat. Appl. EP 556,673, 1993. *Chem. Abstr.* **1994**, *120*, 8337g.

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

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

(28) 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.

(29) 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.

(30) 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.

(31) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange*; Indiana University, Bloomington, IN; Program 455.

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

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

(34) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(35) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press, Wiley: Letchworth, England, 1986.

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

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

(38) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.

(39) 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.

(40) Abraham, R. J.; Smith, P. E. Charge Calculations in Molecular Mechanics IV: A General Method for Conjugated Systems. *J. Chem. Inf. Comput. Sci.* **1987**, *9*, 288−297.

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

(42) 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.

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

(44) Pimentel, G. I.; McClellan, A. L. *The Hydrogen Bond*; Freeman: San Francisco, CA, 1960.

(45) Bradley, G. L. *A Primer of Linear Algebra*; Prentice-Hall: Englewood Cliffs, NJ, 1975.

(46) 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.

(47) Adjustable parameters are calculated based on the number of neurons used in each layer of the architecture as (input × hidden) + (hidden × output) + hidden + output. For a 113-member data set, the number of adjustable parameters should be kept below 56 to reduce the risk of chance correlations.

(48) 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, l. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.

(49) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physiochemical Parameters for Three-Dimensional Structure Directed Quatitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersice Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

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

CI9901237