Prediction of the Normal Boiling Points of Organic Compounds from Molecular Structures with a Computational Neural Network Model

Eric S. Goll and P. C. Jurs*

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

Computational methods were used to link the molecular structures of diverse, industrially important, organic compounds from three different data sets to their normal boiling points. The data were provided by the Design Institute for Physical Property Data (DIPPR) Project 801 database. These data sets were composed of 298 hydrocarbons and heteroatom-containing structures including N compounds (data set I), 277 heteroatom-containing compounds excluding N compounds (data set II), and 104 halogen- and heteroatom-containing compounds, all of which contained at least 1 type of N-functional group (data set III). Each compound was represented by a set of calculated molecular structure descriptors. Genetic algorithms were used to select the best subsets of descriptors. Multiple linear regression and computational neural networks were employed to create the models best suited for the prediction of normal boiling points. This study used a nonlinear genetic algorithm program, for the first time on these data sets, to obtain the final models.

INTRODUCTION

The normal boiling points of liquids reflect the strength of the intermolecular forces (among other forces present) that hold them together. The stronger the intermolecular forces, the more tightly the atoms will be held together and, therefore, the higher the normal boiling point. Polar interactions, including hydrogen bonding, are important contributors to the intermolecular interactions. For compounds unable to engage in polar interactions, the size of the compound is related to the strength of the intermolecular forces. The boiling point can be directly correlated to the chemical structure of any molecule.[1] Normal boiling points can be measured relatively easily, but having predictive models can be a useful adjunct, especially for predicting boiling points for large numbers of compounds. It would seem arduous to first synthesize and then experimentally determine the boiling point of each; this task could prove costly, time-consuming, and, in some cases, dangerous to the researcher or to the environment if radioactivity or other hazardous properties are present. In this study, we apply the quantitative structure−property relationship (QSPR) method to predict the normal boiling points of three diverse sets of organic compounds.

In a QSPR study, the structures of a set of compounds are related to a specific physical property through use of a mathematical model. The method is inductive; the models are developed from a training set of compounds for which the property is known. The key assumption in any QSPR study is that the molecular structure is closely correlated with the physical property being studied. An advantage to such a study is that not only can a physical property be predicted, but also a better knowledge of how the physical property and molecular structure are correlated may be obtained and thus prove beneficial for future studies.

In previous work, these same data sets had been investigated.[1,2] The studies included only linear regression (type I)

models and some work on type II computational neural network (CNN) models. Nonlinear feature selection programs necessary for the development of fully nonlinear CNN (type III) models had not yet been created. The focus of the current work was to obtain type I models, perform in-depth type II studies, and explore type III work on these data for the first time. The three data sets have been kept separate in order to allow direct comparison to previous work.

QSPR methods have been used with a wide variety of properties in addition to the boiling point. Physicochemical properties such as chromatographic retention,[3,4] aqueous solubility,[5,6] and supercritical CO_2 solubility[7] are just a few of the other QSPR studies that have been reported in the literature.

QSPR is just one of the many methods used to predict physical properties such as boiling point. Other predictive methods exist, which have limitations on the types of compounds that can be analyzed. The Meissner method,[8] for example, is applicable to organic compounds that contain C, H, N, O, S, and halides. The Lydersen−Forman−Thydos method[9,10] is used for organic compounds containing C, H, N, O, and halides. The Ogata−Tsuchida method[11] is used for organic compounds that contain C, H, N, O, and halides of the form RX, where R is a limited number of hydrocarbon radicals with eight or less carbons. The Somayajulu−Palit method[12] can be applied to *n*-alkyl halides, acids, amines, ketones, aldehydes, benzenes, cyclohexanes, and cyclohex-1-enes. The Kinney method[13,14] is applicable to alkanes, alkenes, alkynes, cycloalkanes, and cycloalkenes. Stiel and Thodos[15] developed a method to predict the boiling points of alkanes only.

One other method of note was utilized for comparison in the previous work. The QSPR-predicted boiling points were compared with values determined by a different technique, the Joback[16] method. This method is classified as a group contribution technique. A compound's molecular structure

is divided into subgroups, and estimates of the values of its physical properties can be made. These estimates are made based on the type of functional groups included in the structure, as well as the number of times each functional group appears. It was shown that the values predicted by QSPR were better than those from the Joback method.

METHODOLOGY

QSPR studies, as presented here, are performed using the software system entitled Automated Data Analysis and Pattern recognition Toolkit (ADAPT).[17,18] It is used to study the complementary nature between molecular structures and physicochemical properties. After compiling a data set, ADAPT, which is composed of independent routines, serves three main purposes: descriptor generation, feature selection, and modeling. Descriptors are numerical values used to classify different characteristics about a certain structure in order to yield information about the property being studied. The closeness of descriptor values can be used to detect similarities or differences between the given structures.

Three types of models are generated in this work. A type I model is a linear model developed by traditional multiple linear regression analysis. It is the least computationally intensive of the three models. A type II model is generated through the use of a computational neural network (CNN). A type II model is a linear/nonlinear hybrid model because it is based on a set of descriptors chosen by multiple linear regression but a nonlinear CNN model is developed from these descriptors. It is important to note that the best set of descriptors chosen to support a linear relationship are not necessarily the best set for a nonlinear one. It is advantageous to create type II models to study the errors associated with various network architectures. In a type III model, the reduced descriptor pool is analyzed with a genetic algorithm to seek the best subset of descriptors to support a nonlinear CNN model. A CNN is used as a fitness evaluator for each subset of descriptors that is selected by the genetic algorithm. A small set of nonlinear models are reported, and further analysis is necessary to differentiate among them. The use of neural networks usually produces the most effective models; however, they are also the most computationally intensive of the three types. The feature selection routines to develop type III models are the most recently developed.

EXPERIMENTAL SECTION

All computations were performed using a DEC 3000 AXP Model 500 and a DEC Alpha Station 500 workstation running under UNIX. Descriptors were generated with ADAPT routines. Two evolutionary optimization procedures, genetic algorithm[19] and simulated annealing,[20] were used to screen subsets of descriptors. A CNN[21] developed at Penn State University was also utilized to develop satisfactory models.

Data Sets. Three diverse groups of industrially important organic compounds comprise the data sets. Data set I was composed of 298 hydrocarbons and heteroatom-containing structures including nitrogenous compounds. Data set II was composed of 277 heteroatom-containing compounds excluding nitrogenous compounds. Data set III was composed of 104 halogen- and heteroatom-containing compounds, all of which contained at least 1 type of nitrogen functional group.

The compounds and experimental boiling point data were provided by the Design Institute for Physical Property Data (DIPPR) Project 801 database. Each of these data sets has been investigated previously, so we used them again so that comparisons could be made. Listings of the compounds forming each of these data sets have been published, so they are not reproduced here.

Structure Entry and Molecular Modeling. The molecular structure of each of the compounds was entered via the drawing capabilities of the HyperChem Molecular Modeling software package. A compressed connection table, ring information, three-dimensional atom coordinates, and other information are generated for each structure drawn. Geometry optimization via a molecular mechanics routine in Hyper-Chem was used to form the initial three-dimensional representations. These initial representations were submitted to MOPAC,[22] a semiempirical molecular orbital modeling routine, and the three-dimensional representations of the compounds in their lowest energy conformations were generated. This was necessary because the generation of geometric and hybrid descriptors is based on an accurate molecular structure and geometry. If the structures were not in their lowest energy conformations, inaccurate values for the geometric and hybrid descriptors would be calculated. A PM3 Hamiltonian[23] was utilized for geometry optimization.

Descriptor Generation. Upon completion of molecular modeling, descriptor generation was performed using ADAPT routines. Four types of descriptors can be calculated with ADAPT that encode the topological data, geometric data, electronic data, and hybrid information of the molecular structures. Topological descriptors encode connectivity information within a compound. Such descriptors encode κ indexes,[24−26] path lengths of molecules,[27,28] count of atom types, types of bonds, hybridization of carbon atoms, molecular distance-to-edge indexes,[29,30] and molecular paths and connectivity.[31−37] Geometric descriptors yield information about the molecular shape. Such descriptors include the principal moments of inertia,[38] gravitational indexes,[39] shadow−area projections,[40,41] and solvent-accessible molecular surface area and volume.[42] Electronic descriptors encode charge information.[43,44] This charge information can be in the form of charges on the most positive or negative atoms, dipole moments, σ charges, electron densities, energies of the highest occupied molecular orbital and the lowest unfilled molecular orbital, electronegativity, and molecular hardness. There are some descriptors that encode information about more than just one of the aforementioned types; these have been classified as hybrid or combination descriptors. Hybrid descriptors which combine electronic and geometric information include the charged partial surface area[45] and hydrogen-bonding descriptors.

Descriptor Reduction. In most QSPR studies, more descriptors are calculated per compound than can be used for model construction. It is best to construct low-error models with as few descriptors as possible. Since one can calculate approximately 220 descriptors, it is important to reduce the pool of descriptors so that only information-rich descriptors remain. The ratio of descriptors to compounds in the data set should be no more than 0.6.[46]

The overall pool is reduced by two methods of objective feature selection. Objective feature selection is done without

the use of the dependent variable; only the descriptor values are used in this process. The first of the two methods is identical testing, which is used to remove any descriptor with greater than 90% identical values for each of the compounds in the set. This is important because if there is no discrimination between descriptor values, little useful information would be provided by that descriptor. The second is based on pairwise correlations, which are used to remove one of two descriptors that have $R > 0.95$.

Once the descriptor pool had been reduced as much as possible through objective means, subjective reduction was applied. This type of feature selection involves the use of the dependent variable to find subsets of descriptors that correlate best with the boiling point. A genetic algorithm with normal crossover mating was applied to reduce the pool of descriptors further and to allow the user to select the number of descriptors desired in the model. The routines use model builders to calculate the fitness of each descriptor subset. There are two types of fitness evaluation—multiple linear regression and computational neural network analyses, which are linear and nonlinear, respectively.

Type I Linear Models. A type I model is the simplest of the models that can be generated and follows traditional multiple linear regression analysis. The reduced pool was evaluated through the use of a genetic algorithm[19] and simulated annealing.[20] Once the most potentially useful subset of descriptors has been selected, multiple linear regression can be applied to create a linear model.

Six outlier validation tests were performed on type I models. They were residual, standard residual, studentized residual, leverage, DFFITS statistic, and Cooks distance. Although some of the compounds may reflect characteristics that could denote outliers, if the statistical difference between the models with the compounds present or absent is not significant, they need not be removed from the data set. In this case, the original models were considered valid and kept without change.

Computational Neural Networks. The computational neural networks used in this study were three-layer (input—hidden—output), fully connected, feed-forward networks. A sigmoidal transformation of the descriptor values is performed in the first layer, the input layer. The number of input layer neurons is determined by the number of descriptors chosen for the model. The transformed values are passed to the second layer, the hidden layer, where a sigmoidal transfer function is performed. The resulting values are passed to the third layer, the output layer. The output layer contains one neuron that produces the predicted boiling point of a given structure.

A quasi-Newton Broyden—Fletcher—Goldfarb—Shanno (BFGS) algorithm was used to train the network.[19] To do this, the data set had to be split into a training set and a cross-validation set. The cross-validation set was used to prevent overtraining of the network, which can occur after the minimization of the cross-validation set rms error. Training was stopped there because after this point the network was being trained to fit the individual training set compounds instead of the overall set of compounds. After the network was satisfactorily trained, an external prediction set of compounds not used in training was used to test the predictive ability of the network.

Type II and Type III Nonlinear Models. Type II and III models utilize computational neural networks. In a type II model, the same descriptors from the linear type I model are used (those descriptors which were selected by a linear regression method), but the combination of more adjustable parameters, in addition to the nonlinear nature of the neural networks, generally leads to a reduction in the rms error of the prediction and training sets. Since type II models are less computationally intensive to generate than type III models, they are used for testing different network architectures and eventually finding the optimum one for the given data set, which saves a great deal of time. The optimum type II architecture can then be applied to type III models and will generally provide the best results.

A type III model is one that is constructed by means of a nonlinear CNN fitness function evaluation within a genetic algorithm feature selection routine. The descriptors in such a model are chosen for their nonlinear relationship with the dependent variable. Due to their nonlinear relationship, type III models are generally the most computationally intensive models.

RESULTS AND DISCUSSION

Data Set I. The compounds in data set I had molecular masses ranging from 42 to 341 Da, and the experimental boiling points ranged from 225 to 648 K. A large fraction (38%) of the compounds was composed of saturated and unsaturated hydrocarbons, while others contained phenyl, halogen, vinyl, allyl, alcohol, ester, ether, carbonyl, or carboxylic functional groups. Furthermore, there were 34 N-containing compounds in data set I, including amino, cyano, and nitrile functional groups.

The 298-compound data set was split into a 268-member training set and 30-member prediction set. The training set was used to develop type I models and was further divided into a 241-member training set and 27-member cross-validation set for type II and type III model constructions. The prediction and cross-validation sets were chosen at random but were selected to be representative of the different types of compounds in the data set. Sixty descriptors remained in the final reduced descriptor pool.

The best type I model contained eight descriptors. Six of the eight descriptors were included in a previous[1] type I model—three topological, two hybrid, and one electronic. Two additional new descriptors are included in the present model. The model has a standard deviation of 10.62 K for the 268-member training set, with a multiple correlation coefficient of 0.990. Both values are slight improvements on the previously developed model.

It was surprising that five of the eight descriptors in the model were topological and none encoded geometric information. The five topological descriptors included the Weiner number,[27] the number of oxygen atoms in the molecule, the molecular ID,[33] the number of primary sp^2-hybridized carbons, and the molecular distance to edge[29] of tertiary to tertiary carbons. These latter two descriptors were not available prior to 1994, when the previously developed model was created. Therefore, with the availability of these new descriptors, better models can be generated.

The Weiner number is the sum of the graph-theoretical distances (lengths 0—45) between pairs of atoms on the

NORMAL BOILING POINTS OF ORGANIC COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **977**

Table 1. Descriptors Used in the Type III Nonlinear Computational Neural Network Model for Predicting Normal Boiling Points of Heteroatom-Containing Compounds

label	range	description[a]
NO	0 to 4	no. of oxygen atoms in compd
WTPT1	4.9 to 46.5	molecular ID
3SP3	0 to 2	no. of tertiary sp^3-hybridized carbons
DPOL1	2.7×10^{-4} to 4.61	dipole moment
FNSA2	-0.54 to -1.3×10^{-2}	fractional neg surf area 2
FNSA3	-0.14 to -1.9×10^{-3}	fractional neg surf area 3
RDTA	0 to 2	ratio of no. of donors to no. of acceptors
SCDH3	0 to 6.8×10^{-2}	H-bonding descriptor

[a] NO, the number of oxygen atoms in each compound; WTPT1, a weighted paths descriptor that calculates the sum of uniquely weighted paths;[33] 3SP3, the number of tertiary sp^3-hybridized carbon atoms in the compound; DPOL1, the dipole moment of the compound; FNSA2, fractional negative surface area 2 or $[Q_{T-}\sum(-SA_i)]$/total molecular surface area, where Q_{T-} is the sum of the total negative charges for the compound and $-SA_i$ is the ith negatively charged atom in the compound;[45] FNSA3, fractional negative surface area 3 or $[\sum(-SA_i)(Q^-_i)]$/total molecular surface area;[45] RDTA, the ratio of the number of H-bond donors to the number of H-bond acceptors; and SCDH3, $\sum$(surface area $\times$ charge)/total molecular surface area.

structure. Molecular ID is the sum of the unique weighted paths within a compound. For example, in a three-atom compound ABC, the molecular ID would be equal to the sum of the path lengths from A−B, B−C, and A−B−C.

Two hybrid descriptors were included in the model, fractional negative surface area and a descriptor designed to capture a compounds' tendency to engage in hydrogen bonding. Finally, one electronic descriptor was included, the average charge on carbonyl or cyano carbons. The standard deviation was 10.62 K, which is better than the previously published work by over 1 K.[1]

The eight descriptors in the type I model have pairwise correlation coefficients ranging from near 0 to a high of 0.835, with an average absolute value of 0.22.

The same eight descriptors forming the type I model were then used to generate an 8−7−1 computational neural network type II model. The errors for this model were 7.74 K for the training set, 6.97 K for the cross-validation set, and 8.47 K for the prediction set.

The third model type examined was a nonlinear type III model. The reduced pool of descriptors was provided to the genetic algorithm using a CNN to evaluate the fitness of each descriptor subset. The best model found was with an 8−3−1 CNN, which produced rms errors of 7.75 K for the training set, 7.17 K for the cross-validation set, and 8.69 K for the prediction set. If one outlier, propylene, is removed from the prediction set, the new error falls to 7.51 K. Several CNN architectures were examined, and the 8−3−1 network performed best. The eight descriptors supporting this model are shown in Table 1. Figure 1 shows a plot of the calculated versus observed boiling points for all of the 298 structurally diverse compounds for the type III models.

The first three descriptors are topological descriptors; they encode structural details of the compound. One descriptor is an electronic descriptor; it encodes information about the dipole moments of the compounds. The last four descriptors are hybrid descriptors; they encode combinations of topological, geometric, and electronic information about compounds.
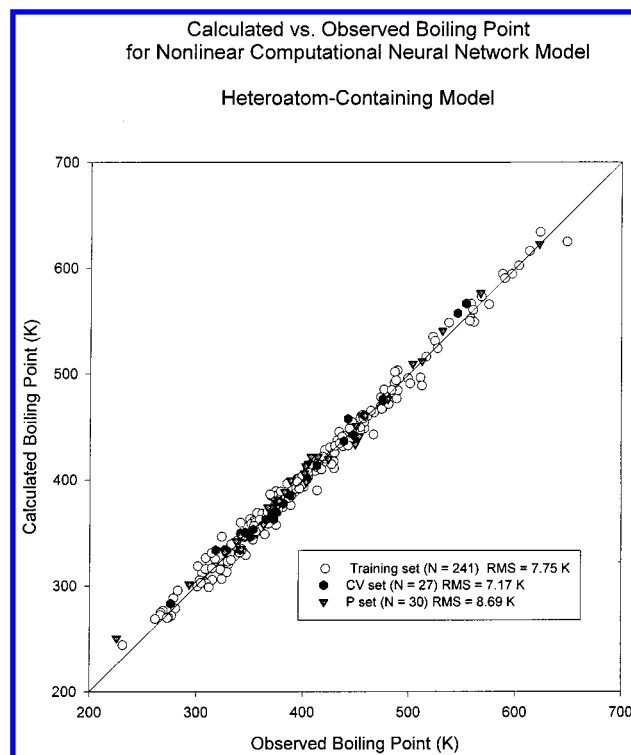
Figure 1. Plot of calculated/predicted versus observed boiling points for nonlinear computational neural network model. Heteroatom-containing model.

Of the eight descriptors in the model, three were topological, four were combination descriptors, and one was electronic. As in the type I model, there were surprisingly no geometric descriptors included in the final model. The topological descriptors included the number of oxygen atoms in the compounds, molecular ID,[33] and the number of tertiary sp^3-hybridized carbons in the compound. The electronic descriptor was the dipole moment of the compound. The hybrid descriptors were the ratio of the number of donor atoms to the number of acceptor atoms in the compound,[47] the sum of the product of surface area and charge of donatable hydrogens divided by the total surface area of the compound,[47] fractional negative surface area 2, and fractional negative surface area 3. The latter two are termed charged partial surface area (CPSA) descriptors, which combine molecular surface area and partial atomic charge information.[45] They combine negative atomic charge, negatively charged surface area, and total molecular surface area data for two unique calculations. Two of the combination descriptors directly encoded hydrogen-bonding effects; none were present in the type I model. This type of descriptor is important since a great many of the compounds in the data set contained alcohol, amine, and other functional groups that involve hydrogen bonding.

The 8 descriptors in the type III model have 28 pairwise correlation coefficients ranging from near 0 to a high of 0.90, with an average absolute value of 0.34.

Other CNN architectures were examined to see if they would produce better models. With a network architecture of 8−2−1, the best model found had a prediction set error of 9.18 K, a cross-validation set error of 8.82 K, and a training set error of 8.78 K. The difference between the three sets is a mere 0.4 K. An 8−4−1 architecture was also studied. The best model found had a prediction set error of

9.05 K, a cross-validation set error of 6.59 K, and a training set error of 6.83 K. The difference between the three sets was 2.46 K. Last, models with only seven descriptors were investigated via a 7−3−1 network. All of these models had high prediction set errors and a rather large difference between the errors of the three sets. The best model, however, had a prediction set error of 10.97 K, a cross-validation set error of 8.48 K, and a training set error of 8.10 K. The difference between the three sets was 2.87.

To ensure that the models described above were not found due to chance, a Monte Carlo experiment was performed. The experimental boiling point values for the 298 compounds in the study were scrambled randomly, thus eliminating any relationship between descriptors and boiling point. Training on such data should yield gross errors in the training, cross-validation sets, and prediction sets. This assumption proved to be correct. The errors were 66.8, 78.4, and 104.9 K, respectively. The enormous increase in errors for these models strongly supports the assertion that the results obtained for the actual boiling point data were not due to chance.

Studies of the same set of 298 compounds comprising data set I in this paper have been reported by 2 other research groups, Katritzky et al.[39] and Hall and Story.[48] The Katritzky paper reported a four-descriptor linear regression model with a standard deviation of 15.21 K for the entire set of compounds. No subset of the compounds was set aside for external validation in that work. Hall and Story developed a nonlinear 19−4−1 CNN model using electrotopological state (E-state) descriptors as the independent variables. They report an rms error of 5.30 K for the 268-member training set and 5.87 K for the 30-member prediction set. E-state descriptors were not available in our descriptor routines when the present work was performed. Judging from the Hall and Story paper and from the work on the other two data sets presented here, E-state descriptors encode molecular structure information that is useful for boiling point models.

Data Set II. The compounds in data set II had molecular masses ranging from 32 to 346 Da, and the experimental boiling points ranged from 145 to 653 K. All 277 compounds contain at least 1 heteroatom. A majority of functional groups were present in the data set. Phenyl, halogen, vinyl, allyl, alcohol, ester, ether, carbonyl, and carboxylic functional groups were all represented in the set. Nitrogen-containing compounds, however, were excluded from the data set.

The 277-compound data set was split into a 250-member training set and 27-member prediction set. The training set was used to develop type I models and was further divided into a 225-member training set and 25-member cross-validation set for type II and type III model constructions. The prediction and cross-validation sets were chosen at random but were selected to be representative of the different types of compounds in the data set.

After objective feature selection had been completed, the reduced descriptor pool contained 91 descriptors. A 10-descriptor linear type I model was generated. It was surprising to note that 9 of the 10 descriptors in the linear model were topological; the last was a hybrid descriptor. The topological descriptors included first- and second-order valence molecular connectivity, the number of chlorine atoms in the molecule, the number of single bonds in the molecule, the number of lone pairs in the molecule, the ratio of

molecular ID to the total number of atoms in the molecule, the average electrotopological state value over all hetero-atoms, and the difference between the maximum and minimum atomic electropological state values. The final descriptor in the model, a hybrid descriptor, was the ratio of the sum of the surface area of donatable hydrogens to the total molecular surface area.

The electrotopological state index is a measure of the reactivity of each atom, analogous to the concept of free valence. It is meant to encode information regarding inter-molecular interactions.[48−50]

Three of the descriptors, the number of chlorine atoms, the number of single bonds, and the number of lone pairs, came from the same program. These three descriptors probably came into the linear model because of their prominence in the data set. About 16% of the compounds contained at least one chlorine atom. About 99% of the compounds contained at least one single bond. About 66% of the compounds contained at least one lone pair. The number of single bonds in a compound is ultimately related to molecular size and molecular weight. When several single bonds are present, the size of the compound, and obviously the molecular weight, increases. Generally, high molecular weight compounds with a significant degree of branching have high boiling points. The number of lone pairs is a measure of the number of heteroatoms and halogens present in the compound. The more heteroatoms and halogens that are present in a compound, the higher the boiling point generally is.

The statistical properties for this linear model were good. The R value was 0.989, and the standard deviation was 12.61. This error was virtually the same as previously studied.[2] The 10 descriptors in the type I model have pairwise correlation coefficients ranging from 0.004 to 0.822. The average absolute value of the pairwise correlation coefficients was 0.288.

The same 10 descriptors were next sent to a computational neural network. All of the architectures examined had 10 input neurons and 1 output neuron, but the number of hidden neurons was varied from 3 to 10. The best type II model architecture found was a 10−9−1 computational neural network. The rms errors associated with this model were 8.92 K for the training set, 9.03 K for the cross-validation set, and 11.03 K for the prediction set.

After discovering the optimal 10−9−1 architecture, a type III model was then generated. After submitting the reduced descriptor pool to a genetic algorithm, seven topological, one geometric, and two combination descriptors were selected as the best model. Table 2 shows the descriptors chosen and the range of descriptor values. As in the type I model, the first- and second-order valence molecular connectivity, the number of single bonds in the molecule, the number of lone pairs in the molecule, and the difference between the maximum and minimum atomic electropological state values appeared in the model. There were two additional topological descriptors in this model—the total number of paths within specified limits in the compound and the superpendentic index.[37] The geometric descriptor that appeared in the model was the gravitation index of heavy atoms. The two hybrid descriptors were the ratio of the sum of the surface area of donatable hydrogens to the total molecular surface area,

NORMAL BOILING POINTS OF ORGANIC COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **979**

Table 2. Descriptors Used in the Type III Nonlinear Computational Neural Network Model for Predicting Normal Boiling Points of Nitrogen-Excluded Compounds

label	range	description[a]
ALLP1	3 to 660	total no. of paths within specified limits
V1	0.289 to 6.49	valence first-order molecular connectivity
V2	0 to 7.06	valence second-order molecular connectivity
NSB	0 to 13	no. of single bonds
NLP	0 to 8	no. of lone pairs
GRAV1	98.8 to 1410	gravitation index of heavy atoms
WNSA3	−14.27 to −1.1	surf-weighted charged partial surf area 3
SADH3	0 to 0.259	H-bonding descriptor
EDIF1	2.58 to 16.5	E-state descriptor
PND	1.41 to 228	superpendentic index

[a] ALLP1, the total number of paths in the structure within the specified lower and upper limits;[27,28] V1, valence first-order molecular connectivity; V2, valence second-order molecular connectivity;[31,32,35] NSB, the number of single bonds in the compound; NLP, the number of lone pairs in the compound; GRAV1, the gravitation index of heavy atoms, which is equal to $\sum(m_1 m_2)/r^2$, where m_1 and m_2 are the masses of atoms one and two in a given bond and r is the bond length;[39] WNSA3, surface-weighted charged partial surface areas or $\{[\sum(-SA_i)(Q^-_i)]$total molecular surface area$\}/1000$, where Q^-_i is the charge on the ith negatively charged atom in the compound and $-SA_i$ is the ith negatively charged atom in the compound;[45] SADH3, $\sum$(surface area of donatable hydrogen atoms)/total molecular surface area; EDIF1, an electrotopological state descriptor which calculates the difference between the maximum atomic E-state value and the minimum atomic E-state value;[48–50] and PND, which calculates the superpendentic index.[37]

Calculated vs. Observed Boiling Point
for Nonlinear Computational Neural Network Model

Nitrogen-Excluded Model

Figure 2. Plot of calculated/predicted versus observed boiling points for nonlinear computational neural network model. Nitrogen-excluded model.

which also appeared in the type I model, and the surface-weighted negatively charged partial surface area.

Since boiling points are related to the skeletal features of organic compounds such as the degree of branching and chain lengths, it would seem that descriptors that encode such information would be valuable. Thus, it is logical that an electropological state descriptor, the total number of paths within specified limits in the compound, the first- and second-order valence molecular connectivity, and the superpendentic index would appear in the model. CPSA descriptors combine the atomic charge and accessible surface area to characterize the ability to participate in intermolecular interactions and are specifically aimed at hydrogen-bonding donors and acceptors.[45] Since intermolecular forces are related to the ability of a molecule to engage in polar interactions, descriptors such as the weighted negative surface area are sensible members of the data set. The ratio of the sum of the surface area of donatable hydrogens to the total molecular surface area came from a descriptor program that encodes information about hydrogen bonding. The degree of hydrogen bonding is also proportional to the boiling point of a compound. The gravitation index takes into account masses of atoms and the distance between two atoms, which could also affect the boiling point.

As expected, the type III models gave the best results. The rms errors associated with the training, cross-validation, and prediction sets were 7.57, 8.07, and 8.26 K, respectively. These errors are nearly 1 K better than found in previous work.[2] Furthermore, in the previous work, two compounds were removed from the study, as they were believed to be outliers which caused skewed predictive values. In this current work, however, all compounds from the original data set were kept. So not only are the errors lower than in a
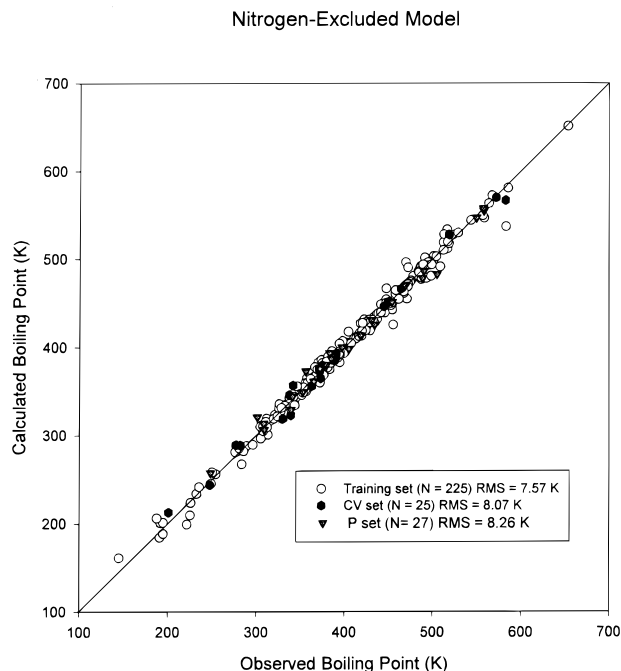
previous model, no justification had to be made in removing compounds from the data set. Figure 2 shows a plot of the calculated versus observed boiling points for all of the 277 compounds for the type III models.

It is apparent that topological descriptors were the dominant encoders of information in the models. Ninety percent of the type I model was composed of topological descriptors, while 70% of the type III model was composed of topological descriptors. Overall, 6 of the 10 descriptors in each model were the same.

Last, the results obtained have been tested for chance effects. A Monte Carlo experiment was performed in which the dependent variables were scrambled. As expected, the models generated produced high rms errors. The best model generated by the ADAPT programs had rms errors of 70.4, 89.7, and 102.0 K for the training, cross-validation, and prediction sets, respectively. These errors are an order of magnitude larger than the models predicted with the correct dependent variables. Thus, only the correct dependent variable can be used to generate reasonable models, and the results obtained were not due to chance.

Data Set III. The compounds in data set III had molecular masses ranging from 27 to 227 Da, and the experimental boiling points ranged from 252 to 628 K. The data set was not as diverse as the other two, as every structure contained at least one kind of nitrogen functional group. Thus, the structures were characterized by amines, nitro groups, pyrroles, and cyano groups.

The 104-compound data set was split into a 94-member training set and a 10-member prediction set. The training set was used to develop type I models and was further divided into an 86-member training set and 8-member cross-validation set for type II and type III model constructions. The prediction and cross-validation sets were chosen at

random but were selected to be representative of the different types of compounds in the data set.

After objective and subjective feature selection had been completed, the final reduced descriptor pool contained 80 descriptors. A 10-descriptor linear type I model was generated. All four types of descriptors were represented in the model. The two topological descriptors were the number of double bonds in each structure and the sum of E-state values over all heteroatoms. Two geometric descriptors were included in the model, standardized area projected onto the *XY* plane and the cube root of the gravitation index of heavy atoms. The charge on the most negative atom and the electric dipole moment encode the electronic information in the model. The remaining four descriptors are all hybrid. The difference in partial surface area is a charged partial surface area descriptor.[45] The three other hybrid descriptors are hydrogen-bonding descriptors. They include the sum of the product of surface area and charge of donatable hydrogen atoms divided by the total molecular surface area, the sum of the surface area of acceptor atoms divided by the total molecular surface area, and the sum of the charges on acceptor atoms divided by the total molecular surface area.

For this linear model, all of the *T* values were greater than four. The *P* values are all low. The *R* value was 0.987, and the standard deviation was 14.06 K. This error was more than a 2 K improvement over the error previously reported.[2] The 10 descriptors in the type model I model have pairwise correlation coefficients ranging from 0.007 to 0.794. The average absolute value of the pairwise correlation coefficients was 0.312.

Type II models were next generated as the same 10 descriptors were sent to a computational neural network. Due to the small size of the data set, the number of adjustable parameters in the network architecture was restricted. The number of adjustable parameters can be calculated by the following equation: $AP = IL \times HL + HL \times OL + HL + OL$, where AP is the number of adjustable parameters, IL is the number of neurons in the input layer, HL is the number of neurons in the hidden layer, and OL is the number of neurons in the output layer. To avoid overtraining the network, the ratio of training set compounds to adjustable parameters should be greater than 2.0. Recalling that there were only 94 members in the training set and 10-descriptor type I models performed best, there could be no more than 47 adjustable parameters. Thus, no more than three hidden neurons could be utilized. The best type II model architecture found was a $10-3-1$ computational neural network. The rms errors associated with this model were 9.17 K for the training set, 10.08 K for the cross-validation set, and 18.78 K for the prediction set. When nitrobutane was removed from the prediction set, the error dropped to 11.72 K.

After discovering the optimal $10-3-1$ architecture, a type III model was then generated. In previous work, 3 compounds were removed from the data set as outliers, cyanogen, malononitrile, and triamylamine.[2] Models were thus generated for both a 101-member data set after removing the three compounds flagged as outliers and also using the original 104-member data set.

After submitting the reduced pool of the 101-member data set to a genetic algorithm, 4 topological, 3 geometric, and 3 combination descriptors were selected as the best model. Interestingly, the number of double bonds, the sum of E-state

values over all heteroatoms, the cube root of the gravitation index, and the sum of the product of surface area and charge of donatable hydrogen atoms divided by the total molecular surface area appeared in both models, but the remaining descriptors encoded very different information. The other topological descriptors in the model were the molecular distance-to-edge of secondary to tertiary carbons[29] and the number of secondary carbons with sp[2] hybridization. Two different geometric descriptors also appeared in the model; they were related to the first major moment of inertia and the first major moment of inertia divided by the second major moment of inertia with different weighting schemes. The two different hybrid descriptors were the count of acceptor atoms and the ratio of donor atoms to acceptor atoms. The rms errors using this model were 7.58, 8.35, and 25.89 K, respectively. Removing one compound from the prediction set, nitrobutane, the rms error of the prediction set dropped dramatically to 8.59 K. The model previously reported for the 101-member data set had a training set error of 10.7 K and a prediction set error of 9.3 K.[2]

Surprisingly, a more improved model was generated from the 104-member data without removing the 3 outliers. This model contained five topological, one geometric, one electronic, and three combination descriptors. Three of the 10 descriptors—1 geometric, 1 electronic, and 1 combination descriptor—in this model also appeared in the type I model. They were the cube root of the gravitation index, the dipole moment, and the sum of the product of surface area and charge of donatable hydrogen atoms divided by the total molecular surface area, respectively. The topological descriptors included the number of secondary carbons with sp[2] hybridization, the cluster three molecular connectivity, the sum of atom IDs for nitrogen atoms, the molecular distance-to-edge of secondary to quaternary carbons, and the minimum atomic E-state value. The remaining combination descriptors included the partial positive surface area and fractional positive surface area.

As previously mentioned, the degree of branching and path lengths directly affect the boiling points, so such descriptors would be important in a boiling point predictive model. Thus, it is logical that the four topological descriptors above appeared in the model. Cluster three molecular connectivity occurs when a three-atom chain, ABC, has a fourth atom, D, branching off of the second atom, B. The other topological descriptors deal with hybridization and atom types, which play a part in determining the boiling points. The gravitation index descriptors deal with atom masses and bond lengths, which apparently also have an effect on the boiling point. The dipole moment probably was selected due to highly electronegative atoms in the model. All compounds in the data set had at least one nitrogenous functional group, some of which are of a more electronegative nature than others. The combination descriptors encode features of molecules that are responsible for intermolecular interactions and hydrogen bonding, which can also contribute to a characteristic boiling point. The descriptors from this model and the range of descriptor values appear in Table 3.

The errors associated with this model were 5.87, 7.60, and 22.11 K for the training set, cross-validation set, and prediction sets, respectively. If nitrobutane is removed from the prediction set, the error again dropped dramatically to 7.14 K. Figure 3 shows a plot of the calculated versus

NORMAL BOILING POINTS OF ORGANIC COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **981**

Table 3. Descriptors Used in the Type III Nonlinear Computational Neural Network Model for Predicting Normal Boiling Points of Nitrogen-Included Compounds

label	range	description[a]
MOLC7	0 to 1.54	cluster 3 molecular connectivity
WTPT5	2 to 8.76	sum of atom IDs for N atoms
2SP2	0 to 9	no. of secondary C's with sp^2 hybridization
MDE 24	0 to 3.14	molecular distance-to-edge of secondary to quaternary C's
GRAV3	4.27 to 10.8	cube root of the gravitation index of heavy atoms
DPOL1	1.65×10^{-3} to 5.25	dipole moment
PPSA3	8.76 to 36.2	partial positive surf area 3
FPSA3	2.87×10^{-2} to 0.164	fractional charged partial surf area 3
SCDH3	0 to 5.76×10^{-2}	H-bonding descriptor
EMIN1	2.58 to 16.5	E-state descriptor

[a] MOLC7, cluster three molecular connectivity;[34–36] WTPT5, the sum of all path weights starting from nitrogen atoms;[33] 2SP2, the number of secondary carbons with sp^2 hybridization; MDE 24, the molecular distance-to-edge of secondary to quaternary carbons;[29] GRAV3, cube root of the gravitation index of heavy atoms, which is equal to $\sum(m_1 m_2)/r^2$, where m_1 and m_2 are the masses of atoms one and two in a given bond and r is the bond length;[39] DPOL1, the dipole moment; PPSA3, partial positive surface area 3 or $\sum(+SA_i)$ where $+SA_i$ is the ith positively charged atom in the compound;[45] FPSA3, the fractional charged partial surface area 3 or $[\sum(+SA_i)(Q^+{}_i)]$/total molecular surface area;[45] SCDH3, $\sum$(surface area × charge)/total molecular surface area; and EDIF1, an electrotopological state descriptor which calculates the minimum atomic E-state value.[48–50]
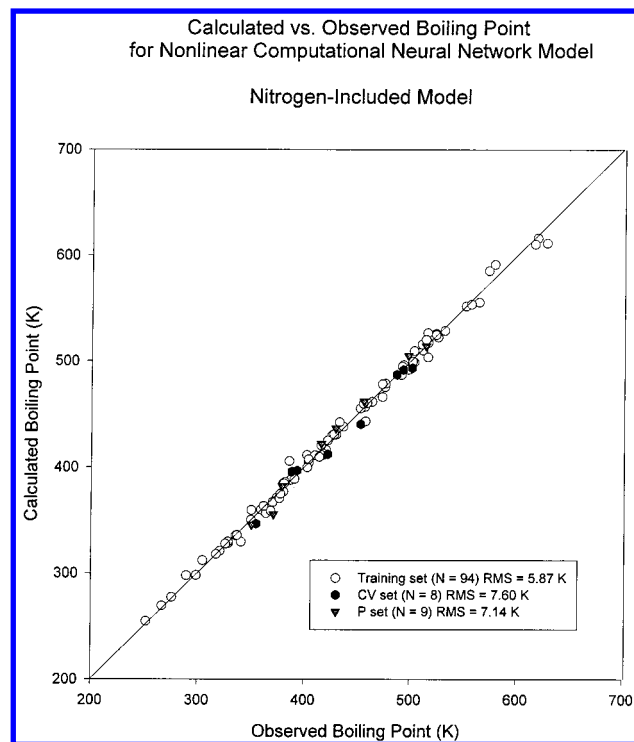
Figure 3. Plot of calculated/predicted versus observed boiling points for nonlinear computational neural network model. Nitrogen-included model.

observed boiling points for the 103 compounds for the type III models (excluding nitrobutane).

Last, the results obtained have been tested for chance effects. A Monte Carlo experiment was performed in which the dependent variables were scrambled. As expected, the models generated produced high rms errors. The best model generated by the ADAPT programs had rms errors of 20.60,

Table 4. Descriptors Used Multiple Times in the Six Models for Prediction of Normal Boiling Points

	data set I		data set II		data set III	
descriptor	type I	type III	type I	type III	type I	type III
NO	X	X				
ALLP	X			X		
WTPT	X	X	X			X
FNSA	X	XX				
MDE	X	X				X
SP HYBR	X			X	X	X
DPOL		X				X
SCDH3		X			X	X
V		XX		XX		
NSB	X	X				
NLP	X	X				
SADH3	X	X				
ESTAT	XX	X			X	X
GRAV				X	X	X

21.23, and 34.25 K for the training, cross-validation, and prediction sets. These errors are nearly triple the errors calculated when the dependent variable was not scrambled. Thus, only the correct dependent variable can be used to generate reasonable models, and the results obtained were not due to chance.

Comparison of the Chosen Descriptors. Six models are reported here: the heteroatom-containing type I and type III models, the nitrogen-excluded type I and type III models, and the nitrogen-included type I and type III models. Table 4 shows a comparison of the descriptor types that appeared in more than one model. Each row of Table 4 relates to a type of descriptor, e.g., the row labeled WTPT reports the number of weighted path descriptors that were used in more than one of the six models. A descriptor that was selected for a particular model is represented by an "X". In some cases, a specific type of descriptor appeared twice in a given model, and this is represented by "XX". Since only descriptors that appeared in more than one of the six models are presented here, the number of X's in each column do not necessarily sum to the total number of descriptors in the model corresponding to that column.

The weighted path descriptors, WTPT, appeared in four of the six models. The electrotopological state descriptors, ESTAT, appeared five times in four models. It should be noted that they were not calculated for the heteroatom-containing models, as they were not available when that data set was studied. The two hydrogen-bonding descriptors, SCHD3 and SADH3, appeared in three models (in data sets I and III) and two models (data set II), respectively. The gravitation index descriptors, GRAV, appeared in one nitrogen-excluded model and in both nitrogen-included models. It is interesting to note that two valence molecular connectivity descriptors appeared in both nitrogen-excluded models. The number of oxygen atoms in a given compound, NO, only came into the heteroatom-containing models. The all path descriptors, ALLP, appeared in one type I model and in one type III model of a different set. The fractional negatively charged partial surface area descriptors, FNSA, seem important in the heteroatom-containing data sets, as they appear three times in the two models. The number of single bonds, NSB, and the number of lone pairs, NLP, appeared only in nitrogen-excluded models. The molecular distance-to-edge descriptors, MDE, and the hybridization

descriptors, SP, appeared in three different models. The dipole moment, DPOL, appeared in two models. The results summarized in Table 4 show that there is relatively little overlap of the descriptors chosen among the three data sets studied. In general, the specific descriptors chosen during a QSPR study are very strongly dependent on the compounds comprising the data set. Thus, using different sets of compounds for the same property leads to the selection of very different sets of descriptors.

CONCLUSIONS

The ADAPT system has been used to generate models to predict the boiling points of compounds from three different data sets based solely on molecular structure. The use of neural networks coupled with a genetic algorithm feature selection routine has allowed for more sophisticated, strictly nonlinear models to be developed.

The ADAPT method has been validated with scrambling experiments. The models generated with scrambled dependent variables yielded inconclusive results. Thus, only the correct dependent variable can be used to generate reasonable models, and the results achieved with the correctly ordered dependent variable are quite unlikely to be due to chance.

Type III models of 8−3−1, 10−9−1, and 10−3−1 were optimum for data set I, data set II, and data set III, respectively. In each case, the models found showed a significant improvement upon the previous models generated. The four kinds of descriptors, topological, geometric, electronic, and hybrid, came into play in these new models, suggesting that many different characteristics play a part in the overall boiling point of a compound.

ACKNOWLEDGMENT

REFERENCES AND NOTES

(1) Egolf, L. M.; Wessel, M. D.; Jurs, P. C. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947−956.

(2) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841−850.

(3) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure−Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799−1807.

(4) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Alkylbenzenes. *Anal. Chim. Acta* **1997**, *342*, 113.

(5) Mitchell, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.

(6) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds Using a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.

(7) Engelhardt, H. L.; Jurs, P. C. Prediction of Supercritical Carbon Dioxide Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 478−484.

(8) Meissner, H. P. Critical Constants from Parachor and Molar Refraction. *Chem. Eng. Prog.* **1949**, *45*, 149−153.

(9) Forman, J. C.; Thodos, G. Critical Temperatures and Pressures of Hydrocarbons. *AIChE J.* **1958**, *4*, 356−361.

(10) Forman, J. C.; Thodos, G. Critical Temperatures and Pressures of Organic Compounds. *AIChE J.* **1960**, *6*, 206−209.

(11) Ogata, Y.; Tsuchida, M. Linear Boiling Point Relationships. *Ind. Eng. Chem.* **1957**, *49*, 415−417.

(12) Somayajulu, G. R.; Palit, S. R. Boiling Points of Homologous Liquids. *J. Am. Chem. Soc.* **1957**, *79*, 2540−2544.

(13) Kinney, C. R. A System Correlating Molecular Structure of Organic Compounds with Their Boiling Points. I: Aliphatic Boiling Point Numbers. *J. Am. Chem. Soc.* **1938**, *60*, 3032−3039.

(14) Kinney, C. R. Calculation of Boiling Points of Aliphatic Hydrocarbons. *Ind. Eng. Chem.* **1940**, *32*, 559−562.

(15) Stiel, L. I.; G., T. The Normal Boiling Points and Critical Constants of Saturated Aliphatic Hydrocarbons. *AIChE J.* **1962**, *8*, 527−529.

(16) Reid, R. C.; Joback, K. G. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Comm.* **1987**, *57*, 233−243.

(17) Jurs, P. C.; Chou, J. T.; Yuan, M. In Olson, E. C., Christofferson, R. E., Eds.; American Chemical Society: Washington, DC, 1979; pp 103−129.

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

(19) 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.

(20) 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.

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

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

(23) Stewart, J. P. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.

(24) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109−116.

(25) Kier, L. B. Shape Indices for Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 1−7.

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

(27) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

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

(29) 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*, 8, 387−394.

(30) Mitchell, B. E. Ph.D. Dissertation, The Pennsylvania State University, University Park, PA, 1997.

(31) Kier, L. B.; Hall, L. H. Molecular Connectivity VII: Specific Treatment to Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.

(32) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. Molecular Connectivity I: Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971−1974.

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

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

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

(36) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(37) 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*, 9,

(38) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; pp 144−156.

(39) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Organics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(40) 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.

NORMAL BOILING POINTS OF ORGANIC COMPOUNDS

J. Chem. Inf. Comput. Sci., Vol. 39, No. 6, 1999 **983**

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

(42) Pearlman, R. S. In *Molecular Surface Area and Volumes and their Use in Structure/Activity Relationships*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980; Chapter 10.

(43) Abraham, R. J.; Smith, P. E. Charge Calculations in Molecular Mechanics IV: A General Method for Conjugated Systems. *J. Comput. Chem.* **1987**, *13*, 288−297.

(44) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1992**, *18*, 492−504.

(45) 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.

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

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

(48) Hall, L. H.; Story, C. T. Boiling Point and Critical Temperature of a Heterogeneous Data Set: QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1004−1014.

(49) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.

(50) Hall, L. H.; Mohney, B. M.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.