Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques

Leanne M. Egolf and Peter C. Jurs*

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

Received March 2, 1993

High quality models which relate structural descriptors to normal boiling points have been developed for large, diverse groups of heterocyclic compounds using both linear regression and neural network techniques. Parallel experiments were designed to compare the performance of these complementary modeling techniques on two different data sets. A formerly studied data set comprised of 299 tetrahydrofuran (THF), thiophene, furan, and pyran compounds was reexamined using neural networks. In addition, a new data set of 572 pyridine compounds was investigated to increase our understanding of the nitrogen-containing heterocycles. First, several new descriptors were developed to explore chemical principles which govern the boiling point process. In particular, descriptors that reflect hydrogen bonding and dipole–dipole interactions proved especially useful for improving the predictive models in the pyridine regression work. With each data set, neural networks were trained to predict boiling points with close to experimental accuracy using the back-propagation learning algorithm. Results from these boiling point investigations show that once the key structural features are indentified through traditional regression techniques, neural networks generally provide access to superior predictive equations. On the basis of this information, further studies were initiated to explore using neural networks as a tool to upgrade structural feature selection. Results from this phase of the study demonstrate that this methodology can be used to identify the most informationally rich descriptors.

INTRODUCTION

Being able to establish the normal (at 760 mmHg) boiling points of organic compounds proves valuable in a variety of practical applications. When one is trying to pinpoint the identity of an unknown chemical substance for instance, the boiling point is one of the first properties to be investigated.[1] Tables of normal boiling points are also frequently cited in hazard assessment publications to help define the risks associated with—and the regulations that should be imposed on—industrially important compounds. This property cannot only serve independently as a measure of a compound's volatility but, when used in conjunction with flash point data, can help determine flammability ratings for many combustible materials.[2] Finally, the property of boiling point has been recognized in the fields of both chemistry and engineering as a powerful parameter which can be used to predict a number of key physical and physicochemical properties. Properties successfully modeled when incorporating boiling point data include critical temperature,[3] molar volume,[4] enthalpy of vaporization,[5] and chromatographic retention indices.[6]

Because of the obvious utility of reliable boiling point data and the occasional unavailability of samples for physical analysis, researchers have made a concentrated effort to develop equations that yield high quality estimations. Group contribution methods and the methods of corresponding states have received the most attention.[5,7] While these methods have proven quite successful, they are generally limited in application. The group contribution methods are constrained by the size of the structure and the diversity of functionalities (i.e., atoms, bond types) which have been parametrized. The method of corresponding states relies on either calculated information or measured physical properties such as surface tension, liquid density, vapor density, molar refraction and critical volume, pressure, and temperature to serve as the variables in predictive equations. The calculated information, since it is commonly based on group contributions itself, may suffer from the same restrictions as those listed above. Meanwhile, the equations which use measured properties as variables are restricted to studying only those compounds which report the required physical property data. Not surprisingly, the propagation of error through the use of this method is also a very real concern. Reviews by Lyman et al.[5] and Reid et al.[7] summarize these and related approaches.

ADAPT[8] is a methodology which can sidestep many of the difficulties associated with the other techniques. In addition, the capabilities of this software package are continually being upgraded and expanded to encompass increasingly diverse compounds. Briefly, the strength of the ADAPT methodology rests in the use of what are known as descriptors. These descriptors encode topological, geometrical, and electronic information, all of which are derived directly from the molecular structure. Our methods are geared to uncover key structural features and then use the numeric representations of these features as variables to develop predictive equations. The property of normal boiling point is one of many properties which we have successfully explored using the ADAPT software methodologies.

Our studies into boiling point prediction were initiated by members of the Beilstein Institute who were interested in obtaining boiling estimates for structurally diverse organic compounds. These estimates would be used for three primary functions: to detect error in the Beilstein physical property database; to validate new experimental data before it is introduced into this database; and to fill the numerous data gaps where experimental values are unavailable.

Past work in this area is summarized in Figure 1. A primary goal of this research was to generate a single predictive equation which would be applicable to as wide a variety of compounds as possible. This objective necessitated a review of all of our previous boiling point studies in order to identify a common set of descriptors which could adequately tie all of the groups

0095-2338/93/1633-0616$04.00/0

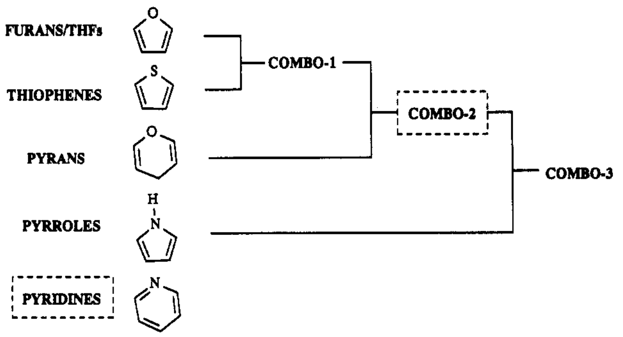


Figure 1. Beilstein boiling point data sets studied with the ADAPT methodology.

of compounds together. It was discovered that a set of 20 descriptors, referred to as the reduced descriptor pool, could provide the structural information necessary to effectively model all data classes examined up to this point.[9] This ability is evidenced in two of the combination models, Combo-1 and Combo-2 (see Figure 1), which represent stepwise additions of the more homogeneous compounds.[9,10]

Results obtained from the most recent combination model Combo-3, though, revealed that there is a definite limit to the predictive capabilities of this abbreviated descriptor pool.[9] The intent of the Combo-3 work was to determine if nitrogen heterocyclic compounds could be combined with the oxygen and sulfur heterocycles of the Combo-2 data set—using the reduced descriptor pool. Although a satisfactory model was obtained for the 752-compound training set, the model had a noticeable deficiency. The N-containing compounds could not be as successfully predicted as the O- and S-containing compounds. We feel this inadequacy is a direct consequence of the descriptor pool used. Since the reduced descriptor pool was compiled strictly from models previously developed on O- and S-containing compounds alone, these descriptors may not contain information that is vital to the successful representation of N-containing compounds. Thus, explorations of the larger descriptor pool have again become necessary.

To this end, the current study of 572 pyridine compounds has been initiated in an effort to increase our understanding of the N-containing heterocycles. It is hoped that this research will enable us to disclose descriptors that are specific to the chemistry associated with the normal boiling point processes of N-containing heterocycles. Because the pyridines are the largest class of heterocyclic compounds currently available from Beilstein, a pyridine model would also provide the means to critically review and fill in data gaps for a major portion of the Beilstein database. As in all previous studies, we will employ the methods of linear regression analysis to identify the most useful structural features for modeling the boiling points of these compounds.

A second goal of this study is to investigate the applicability of artificial neural networks, a methodology related to nonlinear regression techniques.[11,12] The advantage gained through this computational technique appears to hinge on its unique capacity both to learn the underlying rules inherent in a training set of data and to generalize using those rules to a new set of data.[12] The steadily growing interest in neural networks is beginning to become apparent within the chemistry community. Recent research efforts have focused on exploring the use of neural networks for classifying mass spectra,[13] identifying proton-NMR spectra,[14] interpreting IR spectra,[15–17] predicting ^{13}C chemical shifts,[18,19] estimating aqueous solubilities,[20] investigating structure–activity relationships,[21–23] and modeling chromatographic data.[24,25]

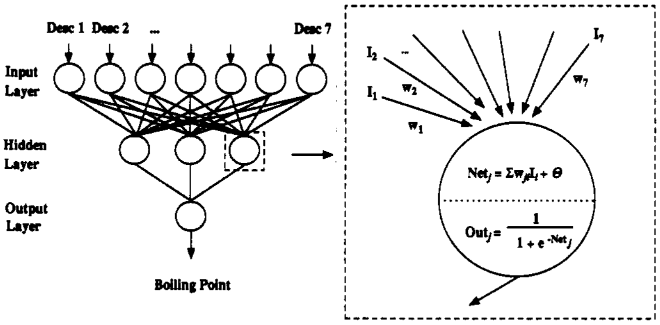


Figure 2. Architecture and function of a fully-connected, feed-forward computational neural network.

The study presented here is the first to introduce the use of neural networks as a means of obtaining normal boiling point estimates. Through our work direct examinations will be made comparing the results from neural networks versus those from linear regression on two data sets: Combo-2 (O- and S-containing compounds) and pyridines (N-containing compounds). The utility of using neural networks as a tool to direct feature selection will also be explored.

THEORY

Artificial neural networks were inspired by scientists' interpretation of the architecture and functioning of the human brain.[12] The way in which the computer-based network works is to accept a set of input facts, transform those facts, and generate an associated set of output facts. Through an iterative "learning" process, the network refines the information derived from the input values (descriptors) in order to reproduce an associated set of target values (experimental boiling points). Once a network has been trained to recognize the underlying theme for a given set of input/target pairs, it may be used to predict an output value corresponding to a new group of input values.

The architecture of a feed-forward, multilayer neural network is illustrated in the left half of Figure 2. A network is comprised of individual processing units known as neurons, represented here as circles. For many networks, as is the case with this one, the neurons are arranged in three layers, an input layer, a hidden layer, and an output layer, each of which is closely linked with its succeeding layer via a netlike array of adjustable connections.

The right half of Figure 2 illustrates how information is processed within the network. The highlighted neuron, located in the hidden layer, receives information from each of the seven input neurons, I_i, (descriptors) through the individual connections, denoted as arrows. Each of these connections is associated with its own weight, w_{ij}, which serves to regulate the amount of signal being passed from one neuron to the next. The hidden neuron assimilates all of the incoming information, applies a bias Θ term, and computes net$_j$. This net sum is next submitted to a nonlinear activation function which generates the output signal, out$_j$, for the neuron. This out$_j$, along with output signals from the parallel hidden neurons, is then passed downstream through the network to be again summed and transformed at the final output neuron.

The signal which exits the output neuron represents, in our system, the network's current boiling point estimate, o_{pk}, for the set of input descriptors. When this value is compared to the desired (target) boiling point value, t_{pk}, a measure of the network error can be computed. Here, the sum-squared error

E is used to quantify the effectiveness of the network parameter settings:

$$E = \sum_{1}^{p} (t_{pk} - o_{pk})^2 \qquad (1)$$

As the data are repeatedly passed through the network, the overall error will successively decrease as the network adjusts its weights and biases to reflect the structure–boiling point relationship.

While a number of training algorithms have been developed for neural network systems, the method of back-propagation is the one applied to our research.[11] On the basis of the steepest descent method, this technique optimizes the connection weights by stepping in a direction which will reduce the error of the estimate. Here, the connection weight, w_{jk}, is corrected by the amount Δw_{jk}, which is proportional to the network error E with respect to each weight.

$$\Delta w_{jk} = -k \frac{\partial E_p}{\partial w_{jk}} \qquad (2)$$

Once the error of estimate has been quantified, an error signal will be propagated back through the network, adjusting the weights and biases in order to minimize the input/target value discrepancy. Using what is known as the δ, or difference, rule[11] the network parameters are modified in the following manner. First, the connections associated with the output neuron alone are corrected. An error function is calculated from the equation

$$\delta_{pk} = (t_{pk} - o_{pk}) f'\mathrm{net}_{pk} = (t_{pk} - o_{pk}) o_{pk} (1 - o_{pk}) \qquad (3)$$

where δ_{pk} is the error term for compound p at the output neuron k, t_{pk} is the target output for compound p at neuron k, and o_{pk} is the actual output for compound p at neuron k.

Through a recursive process, an error function is next calculated for connections between the hidden and input layer neurons by computing an error term at hidden neuron j.

$$\delta_{pj} = f'\mathrm{net}_{pj}(\delta_{pk} w_{jk}) = o_{pj}(1 - o_{pj})(\delta_{pk} w_{jk}) \qquad (4)$$

Accordingly, the weight adjustment for an input/hidden connection is represented as

$$\Delta w_{ij}(n + 1) = \eta \delta_{pj} o_{pi} + \alpha \Delta w_{ij}(n) \qquad (5)$$

where Δw_{ij} is the change in weight between hidden layer neuron j and input neuron i. δ_{pj} is the error function for p at neuron j, and o_{pi} is the output from input neuron i for compound p.

The final two terms are the learning rate coefficient η and the momentum α. The learning rate coefficient is introduced in order to control the average step size of the weight changes, while the momentum term is used to ensure that the previous weight change, or step, is taken into account when the direction and size of the next weight change are calculated. Zupan and Gasteiger recommend the implementation of this term in order to encourage a more directed search of the error surface, resulting in a quicker error convergence.[12]

A neural network is trained, that is the neural network learns, from a predefined group of compounds known as the training set (tset). One potential problem associated with the use of neural network techniques is that it may learn a far too specialized relationship—in other words, overtrain.[26] This problem is averted by choosing a portion of the compounds to serve as what is termed as cross-validation set (cvset). Because this set of compounds has no direct influence on the actual learning process, it can be used to monitor the predictive capability of the network at regular intervals during a training run. By tracking the cross-validation set boiling point error, the optimal set of weights (and biases) to be used as the final predictive model can then be identified.

The separate functions of the training set and cross-validation set are evidenced during a training run. For instance, throughout the course of the training run, the error associated with the training set will continually decrease. This same error trend is not observed with the cross-validation set where the error will decrease to a certain critical point and then begin to steadily increase. The optimal network parameters are defined as those which are present at this point of minimal cross-validation set error.[26] The reason for this is as follows. Up to this transition point, the training set is learning the generalized structure–boiling point relationship. In other words, the relationship which is developed would be of general utility for the prediction of similar pyridine compounds. After the transition point, though, the network instead begins to learn how to fit the structural idiosyncracies of the individual training set compounds. In short, the network begins to develop a much more exclusive structure–property relationship. Consequently, what the cross-validation set allows us to do is avoid overtraining and, by doing so, ensures that the selected model will be effective for the pyridine data set as a whole.

METHODOLOGY

All computations were performed on a Sun 4/110 workstation using the ADAPT software system.[8] The Beilstein Research Institute provided both the structural information and the experimental boiling points for these studies. The structural data were supplied in the form of connection tables and the boiling point data in the form of ASCII text.

Combination-2 Data Set. The Combo-2 data set has been characterized previously. Briefly, this data set combines furan, tetrahydrofuran, thiophene, and pyran compounds. All procedures used to screen these compounds were described in earlier publications[9,27] but follow the general methods discussed below.

Pyridine Data Set. The Beilstein Institute predetermined which compounds would serve as the pyridine data set. It is important to note that although every one of the 572 members of the set does contain pyridine as a basis ring, the term "pyridine" is actually a very incomplete characterization of this diverse group. Here, any given pyridine ring can have a wide range of structural extensions including linear, branched, cyclic, and/or aromatic architecture. In addition, many compounds contain one or more functionalities including ester, ether, carbonyl, nitro, cyano, halogen, sulfur, alcohol, amine, and amide moieties. Nine representative structures are shown in Figure 3 to illustrate the structural diversity of the data set.

To ensure that all of the structures were compatible with our software capabilities, the compounds were examined to identify infrequently encountered moieties. A total of 22 compounds had to be removed from further consideration since they contained functionalities (e.g., Se, Si, deuterium, N-oxides) not currently parametrized in the ADAPT system. All remaining structures were then submitted to molecular mechanics routines in order to place them in energy-minimized conformations.[28]

Prior to assigning each structure a normal boiling point, the experimental data had to be filtered to remove extraneous as well as statistically inconsistent data points. The experimental data provided by Beilstein consist of boiling points (or ranges) and the pressures at which the boiling points were measured. In those cases where no pressure data are listed,

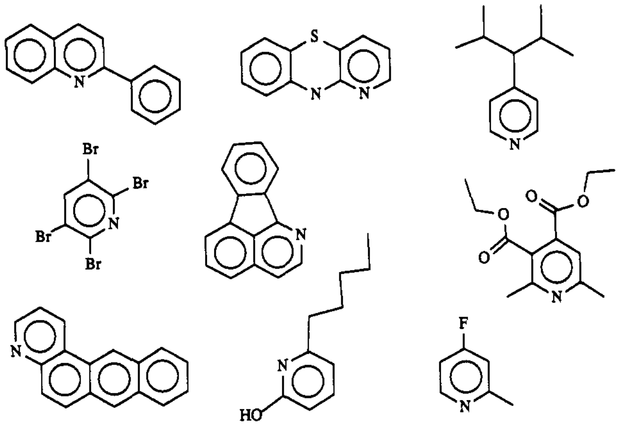


Figure 3. Nine example structures from the pyridine data set.

Table I. 20 Members of the Reduced Descriptor Set

descriptor	classification
1. no. of atoms	topological
2. no. of single bonds	topological
3. molecular weight	topological
4. 1st order molecule connectivity	topological
5. valence-corrected 1st order molecular connectivity	topological
6. valence-corrected 3rd order path molecular connectivity	topological
7. average distance sum molecular connectivity	topological
8. molecular ID/no. of atoms	topological
9. sum of atomic ID's for heteroatoms	topological
10. radius of gyration	geometrical
11. dipole moment	geometrical/ electronic
12. greatest positive partial atomic charge	electronic
13. Hückel LUMO energy	electronic
14. PPSA-1	CPSA
15. PPSA-3	CPSA
16. DPSA-3	CPSA
17. FNSA-3	CPSA
18. WPSA-3	CPSA
19. WNSA-2	CPSA
20. RPCG	CPSA

a boiling point is assumed to have been taken at 760 mmHg. Since our work centers on modeling and predicting only normal boiling points, we first eliminated all data which failed to fall in the range 760 ± 10 mmHg. Because in many cases there are multiple observations (up to 50 for this data set) reported for any one compound, the mean normal boiling point and associated standard deviation for each compound with multiple citings were calculated. Compounds showing an unusually high standard deviation were analyzed in an attempt to discover the source of the variance. A compound in question was retained if the data discrepancy could be rectified by removing an experimental observation that was clearly in error; otherwise, the compound was dropped from the data set. Only two compounds had to be removed due to unresolvable conflicts, leaving a total of 548 pyridines. Lastly, due to storage limitations, a random subset of 499 of these pyridines was selected for use as our final working set of compounds.

DESCRIPTOR GENERATION AND OBJECTIVE FEATURE SELECTION

Combination-2. The Combo-2 portion of this work focuses solely on the previously refined reduced descriptor set which is summarized in Table I.

Pyridines. After each molecular structure was stored along with its associated normal boiling point, descriptor generation began. As discussed earlier, descriptors encode the topological, geometrical, and electronic characteristics of molecular structure. Many of the descriptors used in this study were selected because of their ability to reflect structural features that are intrinsic to boiling point processes. Specifically, shape, branching, surface area, molecular weight, and flexibility index descriptors were introduced since these features contain information concerning van der Waals interactions, molecular flexibility, etc.—qualities that directly influence a compound's tendency to boil. Atom and bond counts and atomic charges were also incorporated as potentially useful descriptors. Finally, the CPSA descriptors which quantify the charges associated with a molecule's solvent-accessible surface area have proven powerful in previous boiling point studies[9,29] and thus were added to the descriptor pool.

A number of new descriptors were developed in order to investigate various aspects of boiling point theory more thoroughly. These descriptors are grouped under three main headings: hydrogen bonding, dipole–dipole interactions, and dielectric constants. The impetus for—and specific nature of—these descriptors will be explained in context within the Discussion section of this paper.

Prior to actual model development, a robust pool of descriptors was identified through the methods of objective feature selection. The first step in this process was to discard any descriptor which contained ≥90% identical values since it offered very limited structural discrimination among the compounds in this particular data set. When only informationally rich descriptors remained, it was then necessary to examine the group as a whole. One descriptor was eliminated from each pair exhibiting a high pairwise correlation ($r \geq 0.95$), thereby greatly reducing information overlap. Often this final feature selection is based on the the descriptor's ease of calculation or its physical interpretability.

Software limitations and/or statistical considerations placed further restrictions on the number of descriptors that could be used in model-building routines. Thus, the goal was to identify those subsets of descriptors which encompassed the most boiling point information. In some cases select pools of descriptors were built on chemical intuition. A more objective, descriptor-ranking approach which also proved invaluable was vector-space descriptor analysis (vsda), a program which utilizes the Gram–Schmidt orthogonalization procedure.[30] Briefly, vsda proceeds as follows. After defining a two-dimensional descriptor space with an initial basis vector and user-selected vector (descriptor), the remaining descriptor pool is scanned to identify that descriptor which contains the largest component that lies outside the current space. This descriptor thus offers the most unique structural information. That feature is then added to the descriptor space, thereby defining a space of one higher dimension. This ordering process is repeated, successively adding the most informationally rich descriptor, until the entire pool of descriptors has been exhausted.

REGRESSION ANALYSIS

Once the reduced descriptor pool had been chosen, regression analysis was begun. Equations which link the structural features with the normal boiling points were developed following the general form

$$\mathrm{BP}_j = b_0 + \sum_{i=1}^{n} b_i X_{ij} \qquad (6)$$

where BP_j is the normal boiling point of compound j, b_0 is the y-intercept of the regression line, b_i is the coefficient of the *i*th descriptor x_i, and *n* is the number of descriptors in the final model.

Leaps-and-bounds[31] and interactive regression analysis (IRA) were the linear regression routines utilized in this study. Leaps-and-bounds is a highly automated program which self-sufficiently identifies top descriptor subsets based on the R^2 criteria. The IRA algorithm, by contrast, is designed to be used interactively. This routine allows the researcher to watch how a model responds as new descriptors are added and old ones removed. Capitalizing on this capability, IRA can be used to analyze and refine models suggested by the more automated but less internally critical programs like leaps-and-bounds.

DISCUSSION

Combination-2. As discussed, the 20-member reduced descriptor set proved effective in former regression work on this 299-compound data set. The top model identified through regression techniques yielded a solid 11-variable model with an R = 0.981 and s = 11.8 °C. Therefore, these 20 informationally rich descriptors were implemented as the initial inputs for a neural network consisting of 20 input neurons, 10 hidden neurons, and 1 output neuron.

This 20:10:1 network was successfully trained using a randomly selected 270-compound training set and 29-compound cross-validation set. Then the connections between the input and hidden layers were systematically clipped to decrease the number of parameters available to the network. It was found that a 20:5:1 network could predict boiling points with the same degree of accuracy as the 20:10:1 network, although the smaller network had half as many adjustable parameters. Furthermore, experiments showed that the particular 270/29 data subdivision did not markedly affect the overall root mean square (RMS) errors achieved.

As introduced earlier, two important terms related to neural network performance are the momentum and the learning rate. These were optimized by monitoring the RMS errors in prediction. Momentum was varied over the range 0.3–0.9, while the learning rate was varied over the range 0.01–2.5. The best values were found to be 0.75 and 0.1, respectively.

Next, the concept of using the neural networks to direct feature selection was explored. Inspection of the final weights assigned to the input-to-hidden connections disclosed that some of the inputs (descriptors) were being essentially ignored. Consequently, descriptors 7, 13, 14, and 19 of Table I were dropped in successive network investigations. As the number of descriptors were decreased, the number of hidden layers were also decreased. Ultimately, the network architecture studied most intensively was a 16:3:1 system, which has 55 adjustable parameters.

Figure 4 is a representative training curve for the Combo-2 16:3:1 network. The symbols illustrate the general decrease in boiling point prediction errors for both the tset and cvset. The horizontal dashed line shows the error associated with the previously reported regression model, 11.8 °C.[9] The RMS error for the cross-validation set reached its minimum at 14 000 epochs with a value of 8.49 °C. Thus the weights at this point were saved as the best values for this network. The fitted vs observed boiling points for the tset/cvset compounds are displayed in Figure 5. The distribution of points along the ideal (1:1) line is excellent, and only a few outliers are seen.

The results obtained with the 16-descriptor neural network should be compared with the results obtained given a parallel tset/cvset regression system. To this end, a model was constructed using the methods of regression analysis. The fitted vs observed results of this boiling point model are illustrated in Figure 6. Here, the RMS error for the cvset compounds was 11.8 °C. A direct comparison revealed that a 30% improvement in prediction accuracy was gained through the use of neural network techniques. This improvement is probably due to the neural network's ability to utilize interrelations among descriptors as well as nonlinearities, whereas linear regression analysis cannot.

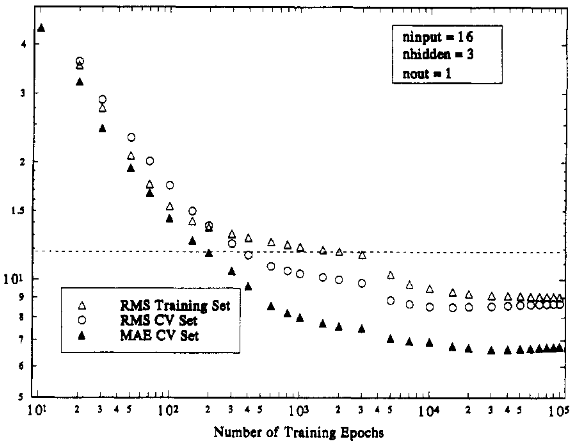


Figure 4. Training curves for the Combo-2 16:3:1 neural network.

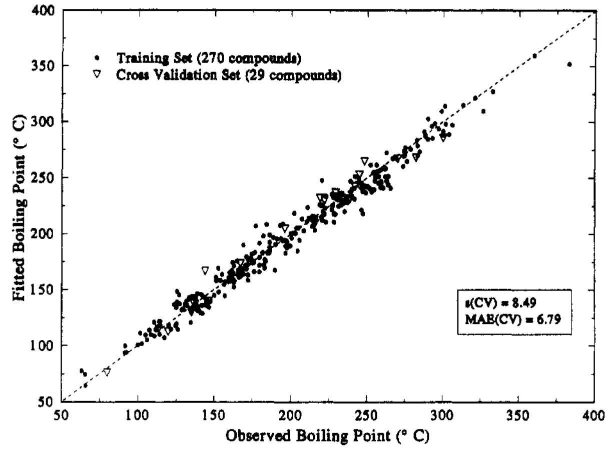


Figure 5. Combo-2 fitted vs observed results using neural network techniques.

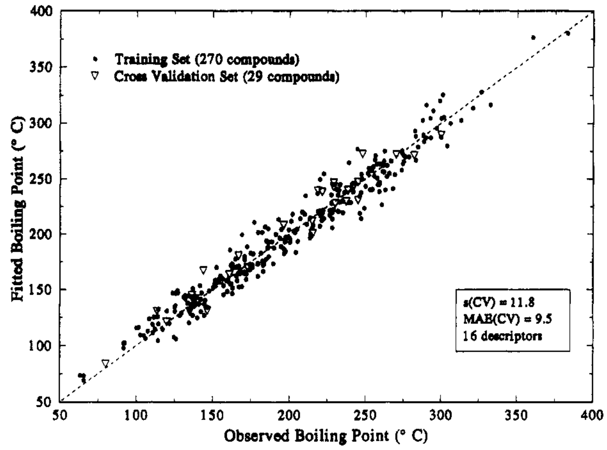


Figure 6. Combo-2 fitted vs observed results using linear regression techniques.

Pyridines. As was discussed earlier, it is not uncommon for the Beilstein database to contain boiling points which were measured at reduced pressures but not flagged as such. In the Data Set section of this paper, we described how much of this data was eliminated after using statistical tests to

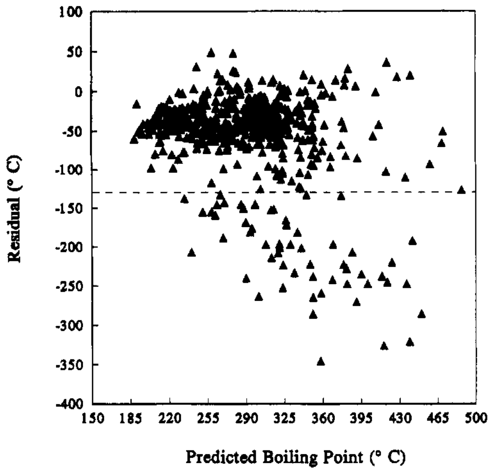


Figure 7. Identifying the questionable normal boiling point data in the Beilstein database.

identify inconsistencies within a set of boiling point values associated with a given compound. It was at this point in the study that the data set was examined in order to spot any obvious boiling point inconsistencies that were evident across the compounds. Unless the majority of the compromised data can be identified and removed prior to equation development, model quality will be severely reduced.

A screening procedure which was proven effective in the past requires using a normal boiling point model (developed through the ADAPT methodology) to predict rough boiling points for the data set of interest. Therefore, the model encompassing the most structural diversity to date, the Combo-3 model, was used to estimate the boiling points for the pyridine compounds. Figure 7 is the residual plot for these predictions, where residuals are defined as

$$\text{residual} = \text{BP}_{\text{expt}} - \text{BP}_{\text{calc}} \qquad (7)$$

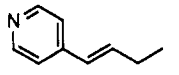
with BP_{expt} ideally being the experimental normal boiling point and BP_{calc} the calculated normal boiling point.

Residual plots are normally used to investigate the distribution of prediction error over the physical property range. Here, it was used to quickly identify any remaining boiling point data which may have been acquired at reduced pressures. Since a compound would be able to boil much more readily at a reduced pressure, a lower than expected boiling point would be reported for BP_{expt}. Consequently, any boiling point that was measured at a reduced pressure will result in an excessively large negative residual as shown through eq 7.

Excessively large negative residuals were defined as those which were 20 °C more negative than the largest positive residual. Examinations of the residual plot show that the mean residual falls at approximately -30 °C. Since the worst positive residual is located just under 50 °C, the error spread in the positive direction is 80 °C. The equivalent cutoff in the negative direction would fall at -110 °C, therefore the actual cutoff was made to be -130 °C, illustrated in Figure 4. Finally, since there were some concerns that the Combo-3 model might be falsely skewing the predictions of these nitrogen-based compounds, the pyrrole model was subsequently used to predict the boiling points of the pyridines. This model gave virtually identical results. Hence, all previously established cutoffs were retained.

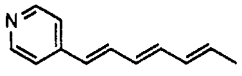
Of the 59 compounds that fell below the cutoff, all could be justifiably removed. For instance, a number of compounds had boiling points that fell significantly below the normal boiling point for their base compound, where the base compound was pyridine ($n_{\text{bp}} = 115$ °C), quinoline ($n_{\text{bp}} = 238$ °C), or isoquinoline ($n_{\text{bp}} = 242$ °C). Chemical trends suggest that the compounds with higher molecular weights should have, in general, higher boiling points. Furthermore, the data set contained some compounds whose boiling points were in severe disagreement with similar compounds. One set of example compounds is pictured below, where it appears that the boiling point assigned to Compound B is probably not its normal boiling point.

compound A



$n_{\text{bp}} = 200$ °C

compound B



$n_{\text{bp}} = 140$ °C

Several boiling points also exhibited clear inconsistencies in the data associated with a compound (shown below).

BP (°C)	pressure (mmHg)
140	10
132	2
129	1
133	– (assumed 760)

The source of this inconsistency is identical to the one responsible for the large standard deviations which originally accompanied some compounds' mean normal boiling points—except that the compounds cited here had only one "normal" boiling point value from the start. Finally, it was deemed acceptable to remove two final compounds with excessively negative residuals, 2,6-bis(trifluoromethyl)pyridine and 2,4-bis(trifluoromethyl)pyridine, since the structural features inherent in these molecules are very unusual and would not be adequately represented within the context of this particular population of pyridines. Just prior to model development, the remaining compounds were finally randomly subdivided into a test set (300 compounds) and prediction set (140 compounds).

Modeling began by turning to the reduced descriptor set. Since a conclusion from the Combo-3 work was that this group of features might not be able to sufficiently characterize the structure–boiling point relationships for nitrogen-containing compounds, we decided to test this hypothesis on the pyridine data set. To summarize, the main statistics for the reduced descriptor set were $R = 0.936$ and $s = 21$ °C, which fall far short of the statistics we have previously reported in the boiling point area. These results confirm our suspicions that a new set of descriptors must be identified which can more effectively encode the structural idiosyncracies of the pyridine compounds.

Efforts directed toward this end began by exploring the set of over 160 core descriptors currently housed in the ADAPT software system. The best model developed from these descriptors was a six-variable model with an $R = 0.936$ and $s = 21$ °C. While the R and s values are identical to those above, the three fewer variables shows this to be a more statistically desirable model. The larger partial F-values associated with the descriptors not only demonstrate their greater statistical significance but also exemplify the increased relative certainty which can be attributed to each individual feature in the model.

Outlier detection schemes, including robust regression analysis (RRA)[32] and data diagnostics generation (DDG),[33] played an integral part in helping to disclose the best models and in helping to point out the specific weaknesses inherent in those models. Consequently, it was through our analysis of these outliers that we were able to recognize one glaring theme. Neglect in encoding hydrogen bonding interactions

Table II. Best Pyridine Regression Model Using New Descriptors

descriptor	coeff ± SD
atomic charge weighted partial negative SA	–1.900 ± 0.140
valence-corrected 3rd order path molecular connectivity	21.33 ± 2.10
no. of single bonds	–6.143 ± 0.681
molecular ID/no. of atoms	341.6 ± 27.67
molar refractivity	3.398 ± 0.183
charge on carbon connected to oxygen	99.60 ± 10.88
$[\lvert(q_{\mathrm{don}})(\mathrm{SA}_{\mathrm{don}})\rvert + \lvert(q_{\mathrm{acc}})(\mathrm{SA}_{\mathrm{acc}})\rvert]/\mathrm{SA}_{\mathrm{tot.}}$	472.2 ± 30.85
intercept	–645.7 ± 52.71
$n = 291$ $R = 0.966$ $s = 15$ °C	

was obviously limiting our model-building progress. First, 19 compounds were cited to be outliers. All of the top models showed similar outlier problems. Next, a breakdown of the outliers shows that 58% were capable of hydrogen bonding. This fact is especially significant when seeing that only 13% of all of the compounds in the data set could hydrogen bond with themselves intermolecularly. Finally, all of the outliers capable of this hydrogen bonding were associated with large positive residuals. This follows again from eq 7 since any model which fails to account for the increased resistance to boiling introduced via hydrogen bonds will yield calculated boiling points which are far too low.

While we hesitate to draw conclusions concerning outlier trends and dipole–dipole interactions, it was interesting to note that five of the outliers contained carbonyl functionalities. Compounds with these moieties can align their dipoles, permitting intermolecular attractions. Although these forces are not nearly as strong as those of hydrogen bonding, the effects of dipole–dipole attractions have been clearly evidenced in the literature[34] through the slightly elevated boiling points associated with simple carbonyl-containing compounds.

New descriptors were developed in attempts to characterize these and other structure–boiling point relationships. Descriptors encoding hydrogen bonding and dipole–dipole interactions included simple counts of donatable hydrogens, acceptor groups, and carbonyl moieties. The charge on donatable hydrogens, acceptor groups, and any carbon attached to a carbonyl or hydroxy oxygen were three descriptors which introduced relative charge information. Charge information was combined with accessible surface area information in hopes of representing the true potential for atomic interaction. Dielectric constants have also been linked with boiling point processes.[35] Consequently, descriptors which featured the polarity of a molecule versus its overall size were generated. Finally, a count of the number of methylenes and Kier's flexibility index[36] were introduced. These descriptors were found to be essential in a related study[37] where hydrocarbon make–up, branching, and flexibility information had similarly been deemed important in theory.

This enlarged pool of structural features immediately proved more powerful. Table II reveals the best pyridine model uncovered through linear regression. The statistics show the dramatic improvement given two of the newly developed descriptors which entered the model as strong variables. The standard error, for instance, has decreased by 33%, yielding a 5% error at the mean of the boiling point range. Furthermore, a mere eight compounds were cited as outliers using RRA and DDG. This indicates that the structural information supplied through this model has drawn in over 50% of the previous outlying compounds. These results suggest that the new descriptors have effectively addressed the key structural features inherent in these nitrogen-based compounds and their specific boiling point processes. (Note that one additional compound was eliminated from consideration prior to generating this model since it had been flagged as a habitual outlier, and we felt it was hindering many model-building attempts.)

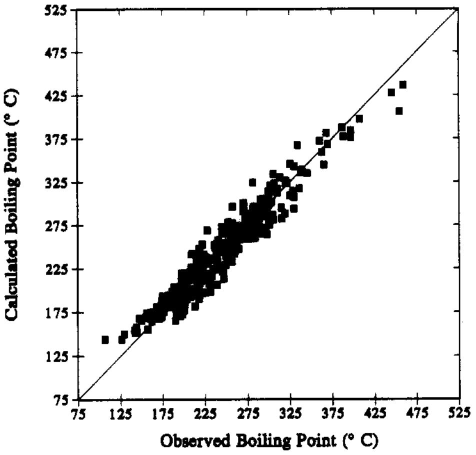


Figure 8. Pyridine calculated vs observed results using linear regression techniques.

Before this model could gain final acceptance, its quality had to be assessed. Validation began with a visual examination of the correlation and residual plots. Figure 8 displays the calculated versus observed data distribution relative to a 1:1 correlation line. Although there appears to be a tendency for the lowest few boiling points to be estimated a bit high, we feel that the overall picture shows the results to be sound throughout the boiling point range.

Next, the results were subjected to internal validation procedures which included jackknifing[33] and examinations of variance inflation factors (VIFs).[33] Jackknifing investigates how dependent the model is on any one compound while variance inflation factors give a measure of undesirable information overlap which may degrade the stability of the model.

With jackknifing, one compound is held out of the model and the model coefficients are recomputed. The boiling point of this compound is then estimated on the basis of the revised model. This process is repeated for each compound that was used in the development of the original model. A model is deemed more statistically robust when no individual compound shows a large difference between its average absolute regular residual and its average absolute jackknife (estimated) residual. The average absolute residuals at the mean of the boiling point range differed by only 0.1%, an amount judged insignificant.

Variance inflation factors provide insight into the multicollinearity among descriptors. Given that a VIF ≥ 10 implies that there may be too much interrelatedness between variables,[33] the descriptors proved to be fairly independent. Our mean VIF was 2.9, with a high of only 4.7.

While internal validation allows one to evaluate the statistical quality of a model, the ultimate test of a model's utility is to see how well it performs in external prediction. Figure 9 illustrates the boiling point prediction of our external data set. The R of 0.940 and standard error s of 20 °C represent acceptable results. The predictions are fairly well distributed on either side of the 1:1 correlation, although there may be some bad data at the higher end of the boiling point range. The skew in that data indicates it may have been measured at reduced pressures. It is obvious that the statistics would be significantly improved if even one of these data points were removed. For instance, by eliminating the most ques-

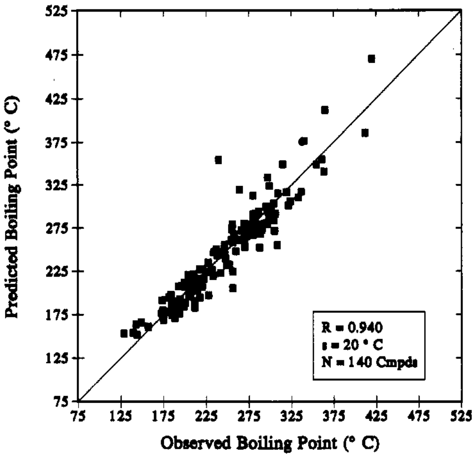


Figure 9. Boiling point predictions for the external prediction set using the pyridine regression model.

tionable data point, experimental n_{bp} = 240 °C, the statistics would jump to $R = 0.960$ and $s = 18$ °C.

Once the model's validity and applicability were verified, an examination of the specific descriptors revealed information concerning molecular structure and boiling point processes. The number of single bonds encodes size information. The valence χ[38] and weighted path[39] descriptors give a measure of the size and relative branching within a molecule. The partial negative surface area-3 descriptor[29,40] encompasses both the amount of negative charge as well as the accessibility of that charge. Perhaps in an attempt to quantify key aspects of the molecular polarizability, molar refractivity has entered into the model as the strongest descriptor. Molar refractivity is defined as

$$R_m = (M_m/\rho)[(n_r^2 - 1)/(n_r + 2)] \tag{8}$$

where M_m is the molar mass, ρ is the density of the material, and n_r is the refractive index.[41] The Lorentz–Lorenz equation (below) shows how molecular polarizability and molecular refractivity are directly related.[42]

$$R_m = 4/3 \times \pi \times \text{Avogadro's no.} \times \text{molecular polarizability} \tag{9}$$

Since compounds which are polarizable are subject to instantaneous dipoles, attractions may form between neighboring molecules, temporarily stabilizing a compound in the liquid state. This stabilization, however slight, will ultimately be reflected in the compound's normal boiling point. The final two descriptors were developed strictly for this data set. The partial atomic charge on the carbon attached to an oxygen may incorporate information concerning dipole–dipole interactions. The hydrogen bonding descriptor not only reveals whether a molecule could hydrogen bond, but it also describes the accessible donor and acceptor surface areas relative to the total molecular surface area.

Although it was hoped that the implementation of new descriptors would enable the pyridine boiling points to be estimated with as much accuracy as was seen with previous heterocyclic data sets, this did not happen. One explanation for this may have been that this data set encompassed more structural diversity. Up to this point, for example, no data sets had ever included cyano moieties and only one, Combo-3, included a compound with a nitro group. These features introduced an additional level of structural complexity. While we could have continued to develop new descriptors to tackle these concerns, we instead decided to branch out and further investigate the utility of artificial neural networks.

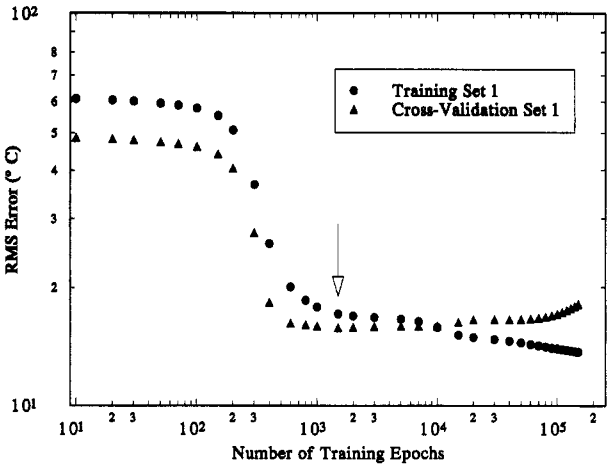


Figure 10. Pinpointing the optimized set of weights for a pyridine 7:3:1 neural network.

Since one of the goals of this work was to make a direct comparison of the capabilities of traditional linear regression versus neural networks, experiments were designed to meet these needs. The 291 compounds that survived the initial regression work were randomly divided into two groups: a training set of 261 compounds and a cross-validation set of 30 compounds. A total of five different random tset/cvset pairs were generated. Using the seven descriptors identified through linear regression, regression coefficients were recalculated for each of the training sets, in essence yielding five new models. These models were then used to predict the normal boiling points for the compounds contained in each corresponding cross-validation set. Because the main objective of our work is to develop models which can estimate boiling points with experimental accuracy, the statistic of interest was determined to be the standard error. Consequently, this was the quantity which was most closely monitored, as will be seen later.

Our parallel experiment with an artificial neural network necessarily used the same seven descriptors to serve as the input neurons in a 7:3:1 network. Figure 2 pictorially represents this specific architecture. Since the output of all neurons in our neural network fall in the interval between 0.0 and 1.0, the descriptors and target values (observed boiling points) were scaled to fall in the same range. The 261 compounds in our first training set were submitted to the network, and training was initiated using the back-propagation algorithm discussed earlier. For all experiments, the momentum was held at 0.75 and the learning rate at 0.01. The network was trained for 200 000 epochs during this stage of our investigations, where one epoch constitutes one presentation of each of the 261 training facts to the network.

During training the weights and biases were iteratively adjusted to reduce the sum squared error of estimation. The optimum network parameters were identified through monitoring the cross-validation set error as originally discussed in the Theory section of this paper. The variation of the RMS error during one tset/cvset training run is illustrated in Figure 10. The results show that the optimized set of conditions fell at 1500 epochs, where the cvset error reached a minimum of 15.80 °C (see arrow). It is important to re-emphasize that any training after this point only reduced the overall utility of the neural network since the network began to fit the structural idiosyncracies of the training set compounds alone. Therefore, the parameters generated at 1500 epochs were those stored for later use.

Table III. Linear Regression vs Neural Network Performance

	training set error (°C)		cross-validation set error (°C)	
test	regression	network	regression	network
1	15.05	17.12	18.90	15.80
2	14.85	14.21	20.82	17.43
3	15.21	14.01	17.65	14.88
4	15.10	14.01	18.35	14.20
5	14.64	13.51	22.41	20.18
mean	14.97	14.57	19.63	16.50

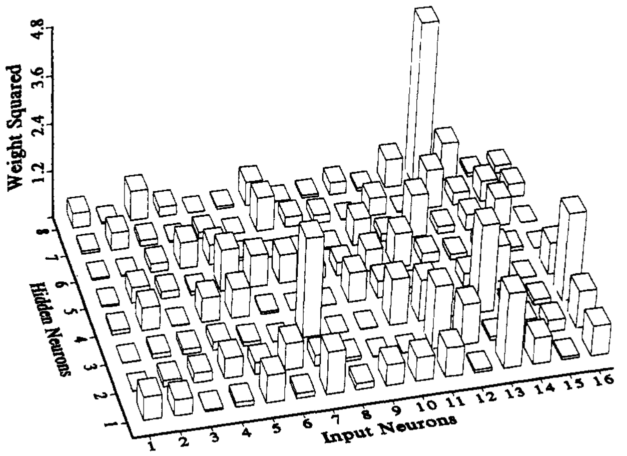


Figure 11. Identifying the least informative descriptors through examinations of the input-to-hidden weights.

Network parameters for the remaining four tset/cvset pairs were optimized in an identical manner. A summary comparing linear regression versus neural network results is presented in Table III. While the difference in RMS error associated with the training set appears to be negligible, the neural network yielded superior results for the cross-validation set. The 3 °C improvement is substantial, especially considering this is relative to our 15–20 °C error scale. Finally, the fitted errors associated with the individual compounds were examined for each of the five neural networks. These results showed that three compounds were poorly represented in all five optimized systems. Unfortunately, visual inspection revealed no structural explanation for why these compounds were not well fit with the rest of the pyridine class. Because these compounds were cited as potential problems so consistently though, they were removed from the data set before proceeding with any further tests.

The final experiment on the pyridine compounds was to determine if a different set of descriptors would provide a better model. Consequently, as with the Combo-2 work, we investigated the feasibility of using neural networks to aid in feature selection. Referring back to the model development stages using linear regression techniques, we identified a set of 16 descriptors which proved important in the strongest regression models. These structural features, which included the seven descriptors used in the original pyridine network, therefore served as the inputs for a new 16:8:1 neural network.

This test began by pinpointing the optimum set of weights and biases for three randomly selected tset/cvset (259/29 compound) pairs. Plots illustrating the optimal input-to-hidden squared weights were then generated. On the basis of the belief that the magnitude of the weight reflects the relative influence of a connection and thus the ultimate importance of the information feeding into that connection (i.e., the descriptor), we assessed the usefulness of each of the input descriptors. Figure 11, for instance, served as one of three tset/cvset plots which indicated that descriptors 1 and 8 were providing the least informational content to the optimized 16:8:1 networks. As a result, all connections leading to these input neurons were severed. One of the hidden layer neurons was also removed in order to further reduce the number of adjustable parameters.

This process was repeated three additional times, successively removing two descriptors (and hidden neurons) per pass using the same criteria. It is important to note that descriptor removal was accomplished in small steps to ensure that we were not removing too much crucial information all at once. After inputs had been eliminated, it was of interest to observe how the network would react and if it could compensate for a loss of structural information.

While feature selection progressed, the network RMS error was also closely monitored. Although the error was increased as the total information content was reduced, the increase was very small (≤1 °C). It was determined that the reduction in the number of adjustable parameters as well as the savings in overall run time far outweighed this minimal elevation.

At this point, eight descriptors remained. It was encouraging to note that all seven of our original descriptors had survived the weeding out process. The choice of which final descriptor to eliminate came down to two descriptors: the charge on the carbon attached to an oxygen and the difference in the atomic charge weighted partial surface areas (DPSA-3).[33] The utility of each of these descriptors was individually explored using two (parallel) 7:3:1 networks.

Two experiments were run. Up to this point, the same random initial starting weight had always been associated with a given descriptor in all of the tested networks. Therefore, while the six unchallenged descriptors would have the same initial weights in each 7:3:1 network, the two descriptors of interest had been assigned different initial weights. In experiment one, these assigned weights were retained as the initial weights. In experiment two, the 7:3:1 networks were made completely parallel by assigning identical initial weights. The results of these investigations showed that (1) the initial starting weights had no significant influence here on model accuracy and (2) the utility of each of these descriptors was high and virtually the same. Because of the slightly lower fit error reported when using the original descriptor, this descriptor was the one retained as the last descriptor in our seven-variable neural network model.

CONCLUSIONS

The normal boiling points of heterocyclic organic compounds comprising large, heterogeneous data sets have been successfully predicted using both linear regression and neural network techniques. Specific to the pyridine work, a more effective set of descriptors, which extended beyond the reduced descriptor set, was identified for modeling nitrogen-containing compounds. Descriptors reflecting hydrogen bonding and dipole–dipole interactions proved to be of immediate utility in helping to develop a more sound and accurate boiling point model. Additional nitrogen-containing groups of compounds will have to be explored to test the more general applicability of these structural features.

Once a set of key descriptors had been pinpointed through regression analysis, neural networks were used to create superior (or equivalent) models for the combination-2 and pyridine data sets. Neural network techniques also proved to be well-suited for feature selection. Examinations of the weighted neuron-to-neuron connections allowed the less useful descriptors to be identified and eliminated, yielding a more refined and ultimately more powerful descriptor pool. Finally, although this technique cannot yet be used to replace linear regression, the results of this work confirm the potential of using neural network methodologies to supply complementary information in these and other structure–property relationship investigations.

ACKNOWLEDGMENT

Steve Dixon is gratefully acknowledged for supplying much of the FORTRAN code which was used to calculate the hydrogen bonding descriptors. Dave Stanton is gratefully acknowledged for his guidance in accessing and managing the Beilstein database.

REFERENCES AND NOTES

(1) Shriner, R. L.; Curtin, D. Y.; Fuson, R. C.; Morrill, T. C. *The Systematic Identification of Organic Compounds*, 6th ed.; Wiley: New York, 1980.

(2) National Fire Protection Agency. *Fire Protection Guide on Hazardous Materials*, 10th ed.; NFPA: Quincy, MA, 1991.

(3) Fisher, C. H. Boiling Point Gives Critical Temperature. *Chem. Eng.* **1989**, *96*, 157.

(4) Sladkov, B. Estimation of Molar Volume of Inorganic Liquids at Boiling Points and Critical Points. *J. Appl. Chem. USSR* **1991**, *64*, 2273.

(5) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982.

(6) Hérberger, K. Discrimination between Linear and Non-Linear Models Describing Retention Data of Alkylbenzenes in Gas-Chromatography. *Chromatographia* **1990**, *29*, 375.

(7) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987.

(8) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979; p 83.

(9) Stanton, D. T.; Egolf, L. M.; Hicks, M. G.; Jurs, P. C. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306.

(10) Stanton, D. T. Ph.D. dissertation, The Pennsylvania State University, University Park, PA, 1991.

(11) McClelland, J. L.; Rumelhart, D. E. *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*; MIT Press: Cambridge, MA, 1988.

(12) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase. *Anal. Chim. Acta* **1991**, *248*, 1.

(13) Curry, B.; Rumelhart, D. E. MSnet: A Neural Network which Classifies Mass Spectra. *Tetrahedron Comput. Methodol.* **1990**, *3*, 213.

(14) Meyer, B.; Hansen, T.; Nute, D.; Albersheim, P.; Darvill, A.; York, W.; Sellers, J. Identification of Proton-NMR Spectra of Complex Oligosaccharides with Artificial Neural Networks. *Science* **1991**, *251*, 542.

(15) Tanabe, K.; Tamura, T.; Uesaka, H. Neural Network System for the Identification of Infrared Spectra. *Appl. Spectrosc.* **1992**, *46*, 807.

(16) Robb, E. W.; Munk, M. E. A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta [Wien]* **1990**, *I*, 131.

(17) Munk, M. E.; Madison, M. S.; Robb, E. W. Neural Network Models for Infrared Spectrum Interpretation. *Mikrochim. Acta [Wien]* **1991**, *II*, 505.

(18) Kvasnička, V. An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts. *J. Math. Chem.* **1991**, *6*, 63.

(19) Anker, L. S.; Jurs, P. C. Prediction of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts by Artificial Neural Networks. *Anal. Chem.* **1992**, *64*, 1157.

(20) Bodor, N.; Harget, A.; Huang, M.-J. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480.

(21) Aoyama, T.; Suzuki, Y.; Ishikawa, H. Neural Networks Applied to Quantitative Structure–Activity Relationship Studies. *J. Med. Chem.* **1990**, *33*, 2583.

(22) Aoyama, T.; Ichikawa, H. Obtaining the Correlation Indices between Drug Activity and Structural Parameters Using a Neural Network. *Chem. Pharm. Bull.* **1991**, *39*, 372.

(23) Aoyama, T.; Ichikawa, H. Basic Operating Characteristics of Neural Networks When Applied to Structure-Activity Studies. *Chem. Pharm. Bull.* **1991**, *39*, 358.

(24) Long, J. R.; Mayfield, H. T.; Henley, M. V.; Kromann, P. R. Pattern Recognition of Jet Fuel Chromatographic Data by Artificial Neural Networks with Back-Propagation of Error. *Anal. Chem.* **1991**, *63*, 1256.

(25) Peterson, K. L. Counter-Propagation Neural Networks in the Modeling and Prediction of Kovats Indices for Substituted Phenols. *Anal. Chem.* **1992**, *64*, 379.

(26) Hecht-Nielson, R. *Neurocomputing*; Addison-Wesley: Reading, MA, 1990; p 115.

(27) Stanton, D. T.; Hicks, M. G.; Jurs, P. C.; Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301.

(28) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; ACS Monograph 177; American Chemical Society: Washington, DC, 1982.

(29) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors for Quantitative Structure–Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323.

(30) Ciarlet, P. G. *Introduction to Numerical Linear Algebra and Optimisation*; Cambridge University Press: Cambridge, U.K., 1989; p 11.

(31) Furnival, G. M.; Wilson, R. W. Regressions by Leaps and Bounds. *Technometrics* **1974**, *16*, 499.

(32) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; Wiley: New York, 1987.

(33) Neter, J.; Wasserman, W.; Kuntner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1985.

(34) Streitwieser, A., Jr.; Heathcock, C. H. *Introduction to Organic Chemistry*, 2nd ed.; MacMillan: 1981; p 361.

(35) Streitwieser, A., Jr.; Heathcock, C. H. *Introduction to Organic Chemistry*, 2nd ed.; MacMillan: 1981; p 236.

(36) Keir, L. B. In *QSAR*: *Quantitative Structure–Activity Relationships in Drug Design*; Fauchère, J. L., Ed.; Alan R. Liss: New York, 1989; p 105.

(37) Egolf, L. M.; Jurs, P. C. Estimation of Autoignition Temperatures of Hydrocarbons, Alcohols, and Esters from Molecular Structure. *Ind. Eng. Chem. Res.* **1992**, *31*, 1798.

(38) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.

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

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

(41) Atkins, P. W. *Physical Chemistry*, 2nd. ed.; Freeman: New York, 1982.

(42) Miller, K. J.; Savchik, J. A. A New Empirical Method To Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206.