Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes

DAVID T. STANTON and PETER C. JURS*

Chemistry Department, The Pennsylvania State University, University Park, Pennsylvania 16802

MARTIN G. HICKS

Beilstein Institute, Varrentrappstrasse 40-42, D-6000 Frankfurt/Main 90, Federal Republic of Germany

Received October 9, 1990

Computer-assisted methods are employed for the development of predictive equations which relate molecular-based structural features to the normal boiling points for a large number of compounds containing furan, tetrahydrofuran, and thiophene ring systems. Predictive models are described for a set containing all three ring classes, as well as for a set containing only furans and tetrahydrofurans, and another set comprised only of thiophene-related compounds. The fit error for the combined data set is 4.9% of the mean boiling point for the data set, while the fit error for the furan/THF subset was 5.8% and for the thiophene subset was 3.8% of the mean boiling points of the respective data sets. Similar results are obtained for the prediction of new external data sets for each model. The models developed here are examined to gain insight into the relationship between structural features and normal boiling points for the three ring classes described.

INTRODUCTION

The normal boiling point is an important property that is used to characterize organic compounds. The boiling point is often one of the first properties to be determined in the attempt to identify an unknown.[1] However, it may often be the case that appropriate reference data is not available for a given compound. It may also not be possible to determine the boiling point for a given compound due to a lack of available material. In such instances, predictive methods can be employed to obtain an estimate of the boiling point. Another important application of estimation methods is the validation of experimental data. In such cases, the estimation method is used to detect possible errors before the experimental data is used. Estimation methods could also be used to fill gaps in large data files where it is impractical to obtain experimentally determined values.

Several methods for the prediction of the boiling points of organic compounds have been described in the chemical literature over the years. Pearson[2] has described a reasonably effective and simple method for the prediction of the boiling points of simple organic compounds. Lyman et al. have summarized several more generalized methods for the prediction of boiling points based on specific group additivity rules.[3] Other methods employing group additivity schemes have also been reported.[4,5] In addition, some methods have been described which require experimentally determined quantities such as gas chromatographic retention indices in order to estimate normal boiling points.[6,7] In a series of papers, Cramer reported studies in which a number of physical properties, including boiling points, were interrelated by factor analysis.[8-10]

Computer-assisted methods have been developed in our laboratory which employ structural parameters (or *descriptors*) that are calculated directly from the structures of a given set of compounds. These parameters are then correlated to observed boiling points by using multiple linear regression analysis techniques to obtain linear predictive equations. Such equations are termed quantitative structure–property relationships (QSPR). Hansen and Jurs[11] have described the development of such a predictive equation based on a set of olefins, while Smeeks and Jurs[12] have described a similar equation based on a set of acyclic alcohols.

There are several advantages to the QSPR method. Group contribution methods require that all the structural fragments of a query compound exist in the table of fragment constants. The existence of only one fragment in the query molecule, which is absent in the stored-fragment table, causes the method to fail. Since parameters calculated in our method encode more generalized information, predictive equations developed for one set of compounds can often be applied to a new set of compounds which are related to but not specifically represented in the original training set. In addition, it is often possible to obtain insight into the relationship between the structure of a series of compounds and the property of interest. Implementation of the QSPR methodology using a computer allows for rapid and facile entry of structures and data, calculation and subsequent analysis of the descriptors, and generation and validation of the model equations.

The goals of the current work were twofold. First, it was of interest to develop equations to correlate structural features of a given class of compounds (furans, tetrahydrofurans, and thiophenes in this case) with their observed boiling points. Later, a more global model incorporating all three ring classes would be sought. The equations derived in this manner would then be employed for the purpose of predicting boiling points of similar compounds within the Beilstein Institute data files where experimentally determined values are unavailable. Also, the equations could be used in a quality control capacity to check existing data for possible errors. The second goal of this study is to determine if the empirically derived QSPR equations could lend insight to the relationship between structure and normal boiling point, in terms of the importance of the descriptors used, for the three classes of heterocycles involved.

METHODOLOGY

The general procedure used in this study is outlined in the flow diagram shown in Figure 1. All computations were performed on a Sun 4/110 workstation, running under the UNIX operating system. The programs used are part of the ADAPT software system.[13,14]

Data Sets. The structures and boiling point data for 254 furans and tetrahydrofurans and 223 thiophenes were received

0095-2338/91/1631-0301$02.50/0

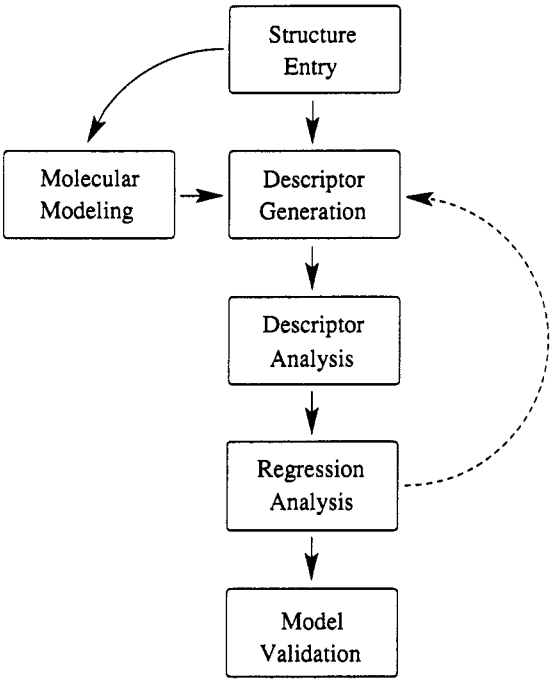

Figure 1. General flow diagram for the procedures used in an ADAPT structure–property relationship (SPR) study.

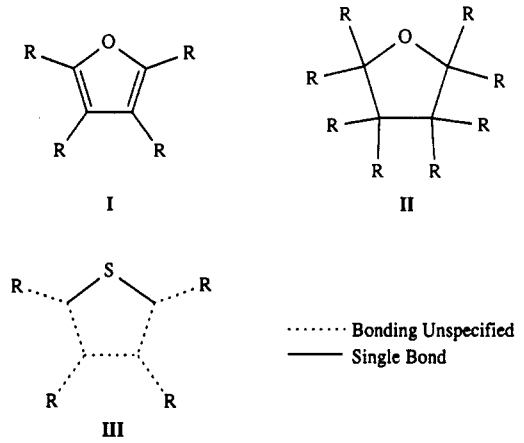

Figure 2. Generic diagrams for the structures of the compounds involved in the boiling point study. The unspecified bonding shown in **III** indicates sites of possible unsaturation or ring fusion. The R groups can be any of a variety of substituent types.

as ASCII files from the Beilstein Institute. The general structures for the three data sets are given in Figure 2. While the compounds in each data set were broadly classified as furans, tetrahydrofurans, and thiophenes, the data sets also contained other ring types (aliphatic, aromatic, and heteroaromatic), as well as a wide variety of other functionalities (e.g., esters, ethers, alcohols, amines, halogens).

Structure Entry and Processing. The structures of the compounds involved in this study were received in the form of machine-readable connection tables in the CAS Registry Services format.[15] A FORTRAN program was written which converted the data to the ADAPT connection table format which was then stored on disk. A separate program was written that converted the bond patterns for the aromatic ring systems from alternating single–double bond (Kekulé) structures to the ADAPT aromatic bond format shown in Figure 3. The exceptions to this process were the five-membered heterocyclic aromatic ring systems. These were retained in the alternating single–double bond format.

Once the structure data had been processed and stored in ADAPT format connection tables, the structures were placed in energy-minimized conformations using the MM2 algorithm.[16] In cases where such calculations were not possible due to the lack of necessary parameters, the bond lengths and bond angles were corrected to standard values, and the structures were placed in energy-minimized conformations by using alternative molecular mechanics calculations.[13] Three-dimensional molecular models were developed so that geometry-dependent structural descriptors could be used.

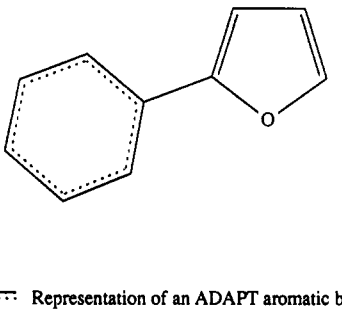

Figure 3. Diagram illustrating the representation used for aromatic systems in ADAPT.

Boiling Point Data. The boiling point data used in this study were received and used in units of °C and were reported to have been determined at 760 mmHg. The data for individual compounds were received in several forms. Some of the compounds had only a single reported value, while others had several reported boiling ranges. For the purposes of this study, it was necessary to obtain a single value for all compounds. Initially, the furan and tetrahydrofuran data sets were carefully screened, and a limit of 4.0 °C was placed on the usable boiling range for a single compound. The mean of the boiling range was then used as the boiling point. However, as the study progressed, it was noted that the inherent experimental error in the data set was greater than the restrictions which were arbitrarily imposed, causing the expectations for the results of the modeling process to be unrealistic. Therefore, the processing of the thiophene boiling point data was less rigid, and the mean of all the reasonable data available was used to obtain the boiling point value for a given compound.

Descriptor Generation. Many structure-based molecular descriptors were evaluated in this study. Most of these descriptors fall into one of three general classes that encode the topological, electronic, or geometric features of the molecules. Examples of topological descriptors include path counts and molecular connectivity indices.[17] Greatest positive and negative partial atomic charges[18] and the submolecular polarizability parameter[19] are examples of electronic feature descriptors. Geometry-based descriptors include length-to-breadth ratios and solvent-accessible surface areas. Some calculated physical properties such as molecular polarizability,[20] molar refractivity,[21,22] and molecular weight are also used. Finally, some descriptors combine two or more of the general types of information into a single descriptor. Examples of such descriptors are the CPSA descriptors which were recently developed in our laboratory.[23]

Descriptor Analysis. Once the desired set of descriptors had been calculated and stored, the process of descriptor analysis was begun. The goal of this analysis is to examine the pool of descriptors in an objective manner and to remove from further consideration those descriptors which are redundant or which do not contain enough discriminatory information to be of any significant value. All descriptors that contained identical values for 90% or more of the compounds in a given data set, including both zero and non-zero values, were removed. All possible combinations of remaining descriptor pairs were examined to identify those pairs which were highly correlated. As a rule of thumb, a critical value of 0.950 for the correlation coefficient (r) was used. If two descriptors were correlated at or above the critical value, one descriptor was

discarded. The decision of which one to retain was based on the possible physical interpretation of the descriptor, ease of calculation, or usefulness in past studies. The result of this analysis is a reduced pool of information-rich descriptors which can then be screened by using multiple linear regression analysis methods.

Regression Analysis. In most cases, linear regression models were developed by the method of multiple regression with progressive deletion.[24] The process involves forming a model through stepwise addition of terms (descriptors), where the inclusion of a given term is based on the F statistic values. A deletion process is then employed where each independent variable is held out in turn, and a model is developed by using the remaining pool of descriptors. Then all pairs and triplets are held out, and the model development process is repeated. This series of steps has the effect of uncovering potentially superior equations that may have been obscured by the presence of a descriptor which was highly correlated to the dependent variable. In the case where a more thorough analysis was desired, the method of best subsets regression was employed with the Minitab statistical package.[25]

In all cases, models were evaluated for statistical significance based on the overall F test. Individual descriptors in the model were examined for significance using the partial F test. Models were also examined for robustness and predictive ability through both internal and external validation methods as was appropriate. The specifics of these evaluations are included in the discussion below.

RESULTS AND DISCUSSION

Before any models could be developed, two problems which were specific to the data sets had to be addressed. These were the diversity of the structures involved and the large amount of experimental error contained in the data sets. The first and most difficult problem to solve was the issue of the experimental error.

Furans and Tetrahydrofurans. The initial focus of this study was the 254 member furan and THF data set. The generalized structures are given in Figure 2, compounds I and II, respectively. The average standard deviation of the boiling point data for replicate values in the data set was determined to be 7.4 °C. Examination of the structures suggested that the two ring classes could be combined because of the topological similarity of the parent ring systems. One of the most striking features of the data set was its diversity. Many of the common functional groups (e.g., esters, carboxylic acids, alcohols, double bonds) were present in the data set, and in many instances two or more functionalities were contained within a given structure. Only the halogens had been excluded during the initial search of the Beilstein database. This was done because it was thought that the halogens might be more difficult to properly represent with the molecular descriptors available.

Work was begun by attempting to model the data set as a whole. The results of this initial effort were poor, so several methods of automated subsetting were tried, but these yielded poor results as well. Finally, a small, manually selected, subset of 91 compounds which involved both ring classes and contained only simple, nonassociating (non-hydrogen bonding) functional groups yielded the first reasonable model. This particular subset of compounds was chosen because it is known that boiling point increases with increasing molecular weight for a set of similar compounds. The model obtained by using this small subset was then used to predict the boiling points of the remaining 163 compounds. It was hypothesized that compounds which were poorly predicted by this model would generally yield large positive prediction errors. The positive sign of the error would be due to the fact that some of the structural features of the molecules responsible for strong intermolecular interactions, and therefore higher boiling points, would not be properly encoded in the model, causing the boiling point to be underestimated. The magnitude of the error would then be a function of the amount of information missing from the model with respect to a given compound. While results of this nature were observed, a more notable result was the observation of large negative prediction errors. Many of the 163 compounds in the prediction set yielded prediction error values in excess of −100 °C. It was found, through examination of the literature, that the boiling points for the compounds yielding large negative prediction errors had actually been determined at pressures below 760 mmHg. Since this data set was chosen specifically so that all the boiling point data had been reported at 760 mmHg, the unexpected prediction results caused us to reexamine the basic assumptions we had made concerning the data set.

Furan/THF Training Set Selection. In a designed QSPR study, the data set upon which the relationship will be based is usually carefully chosen. Optimally, one would like to have all the values for the property of interest measured in a single laboratory under identical conditions for each compound in the data set. This would minimize the amount of experimental error and make the task of modeling the data set much easier. However, it is often not possible to use such carefully obtained data, and one must use the information in the data available. This study is a prime example of just such a case. The Beilstein database is comprised of data that has been extracted from the primary literature over many decades. Thus, for any given set of compounds, experimental values for a particular endpoint have been determined by many different people, in many different laboratories, under widely different conditions. These problems are often compounded by the omission in the primary literature of critical information such as the pressure at which the boiling point was determined. Therefore, some of the emphasis of this work was shifted to the problem of data analysis and the need to extract a good training set from a larger data set that contains a relatively large amount of error from a number of sources.

With the existence of the experimental error in mind, our attention turned to the task of selecting a training set for model development which is based on compounds with boiling points determined at atmospheric pressure. However, since there existed no a priori method of removing only those compounds with errors in the experimental data associated with boiling points being determined at reduced pressure, a variety of statistical methods were used to select the final training set.

Since it was assumed that the majority of original data set contained minimal error, the process of fitting a regression function to the majority of data should allow for the detection of those compounds exhibiting large levels of error. It was found that this approach, used in an iterative manner, yielded a good training set. The analysis must be done in an iterative fashion because the selection of the molecular descriptors to be used in the regression analysis is influenced by the existing outliers. As the outliers are removed, the data set is refined, which results in the selection of different variables in subsequent steps. Thus, multiple linear regression analysis was used, in conjunction with a variety of outlier detection methods to select the final furan/THF training set.

Of the original 254 compounds, 10 were removed from further consideration due to the very large negative prediction errors noted above and subsequent verification of the error in the literature. The remaining 244 compounds yielded a model with a reasonable fit ($R^2 = 0.801$) and standard deviation of regression ($s = 28.4$ °C). We decided to examine the model by using a method termed robust regression analysis (RRA) which has been described by Rousseeuw[26,27] and implemented in a program called PROGRESS. The PROGRESS program was

Table I. Description of the Six Tests Used for the Purpose of Outlier Detection

test	description[28,29]
(1) residual	difference between experimental and fitted boiling point
(2) standardized residual	residual divided by SD of the fitted regression eq
(3) studentized residual	residual divided by its own SD
(4) leverage	measure of the influence of a given point in determining the fitted eq
(5) DFFITS	describes the difference in fit of the regression eq when a given point is removed
(6) Cooks distance	describes the change in the model coefficients when a given point is removed

modified in order to access data from the ADAPT system.

The methods of regression generally used in ADAPT are variations of the method of linear least squares, which is very sensitive to the existence of outliers in the data set under study. Robust regression is similar, but is based on the concept of minimizing the median of squared residuals, and thus very insensitive to the existence of outliers. Generally, if a data set does not contain outliers, the results for least-squares regression and robust regression will be similar. However, the existence of a single outlier in the data set can cause models developed by these two methods to be quite different. Robust regression allows for the identification of the observations which are causing problems. The model described above, based on 244 observations, was submitted to the PROGRESS program for analysis. The results indicated that 27 of the 244 observations were outliers and yielded large negative residuals (fit errors) much like what was obtained for the 10 compounds described above. Since the type of result obtained for the outliers was the same as previously observed, these 27 compounds were set aside, yielding a training set of 217 compounds.

Molecular descriptor analysis was repeated in the new training set. The new 217 observation data set was then submitted to linear regression analysis, and much improved results were obtained. The fit of the model was increased (R^2 = 0.960), and the standard deviation of regression was reduced (s = 12.6 °C). These results were very encouraging. While there was generally good correlation between the fitted and observed boiling points, there were still a few compounds yielding relatively large negative residuals, and there seemed to be an overall negative skew to the residuals. These results suggested that there may still be additional error in the training set. Robust analysis could normally be applied again to find potential outliers, but the current model contains more descriptors than can be processed in the current implementation of the program PROGRESS. Because of this, more conventional tests involved in the process termed *residual analysis* were applied.

There are several tests one can apply in the process of residual analysis. While each acts as a specific measure of how much influence a specific observation has on the regression function, they are generally used in connection with least-squares regression methods, which are sensitive to the existence of outliers. Thus, if sufficient error exists in the data set, it may not be possible to rely on any one test to detect an outlier. For this reason, we chose to apply six different standard statistical tests and to define an outlier as an observation which failed three or more of the six tests. The six tests we chose to use are described in Table I.[28,29] Using this criterion, eight additional compounds were indicated as outliers, with all but one of the observations exhibiting relatively large negative residuals. These were also removed from the data set. It was felt that at this point that the majority of compounds which possessed the type of error described above had been detected and removed. The remaining set of 209 compounds was then collected as the and used as the final training set.

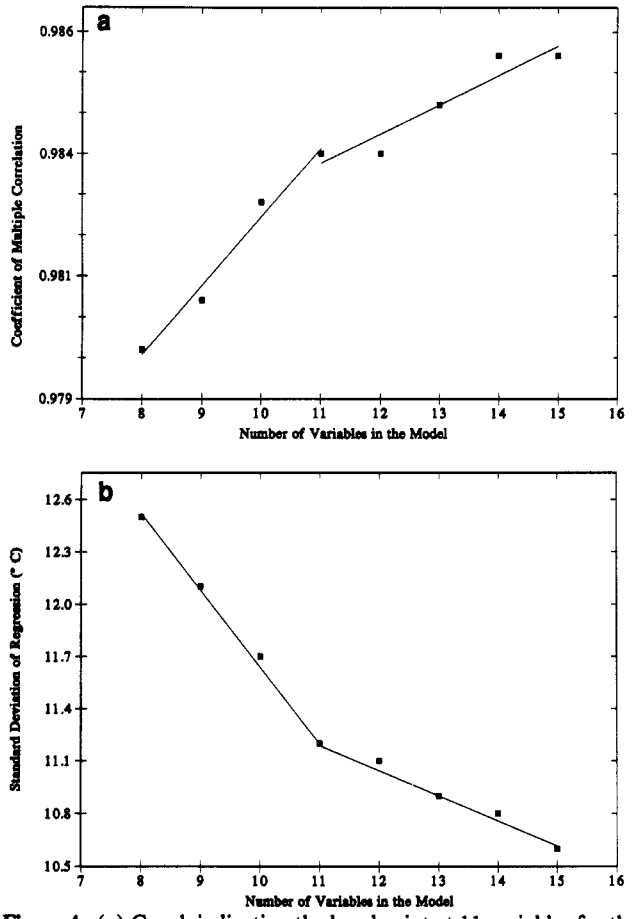

Figure 4. (a) Graph indicating the breakpoint at 11 variables for the choice of model size for the furan/THF class specific data set, based on the value of the multiple correlation coefficient. (b) Graph indicating the breakpoint of 11 variables for the choice of model size for the furan/THF class specific data set, based on the standard error of regression.

Table II. Details of Final Furan/THF Class Specific Model Based on the 209-Observation Training Set

R = 0.984 (R^2 = 0.969), s = 11.2 °C, n = 209

F value (for AOV) = 554.5

structural descriptor	regression coeff	SD of coeff
(1) path 1 simple molecular connectivity	61.61	4.46
(2) no. of single bonds	−25.36	1.11
(3) valence-corrected path 3 molecular connectivity	22.94	3.63
(4) av distance sum connectivity[a]	21.14	3.99
(5) square root of KAPPA-3[b]	47.15	7.31
(6) PPSA-1[c]	0.61	0.052
(7) PPSA-3[c]	10.78	1.04
(8) FNSA-3[c]	−1296.78	55.76
(9) WPSA-3[c]	−26.87	2.89
(10) RPCG[c]	145.05	13.00
(11) LUMO energy, simple Huckel method	11.52	2.97
intercept	−280.94	

[a] See Balaban.[30] [b] See Kier.[31] [c] See Stanton and Jurs.[23]

Model Development. The process of descriptor analysis and linear regression analysis was repeated once again with the final training set of 209 compounds. Several good models of various sizes were obtained, and it was necessary to choose the best model. The choice was made based on the rates at which the coefficient of multiple correlation (R) and the standard deviation of regression (s) changed with changing size of the model. The goal is to have the best model with the

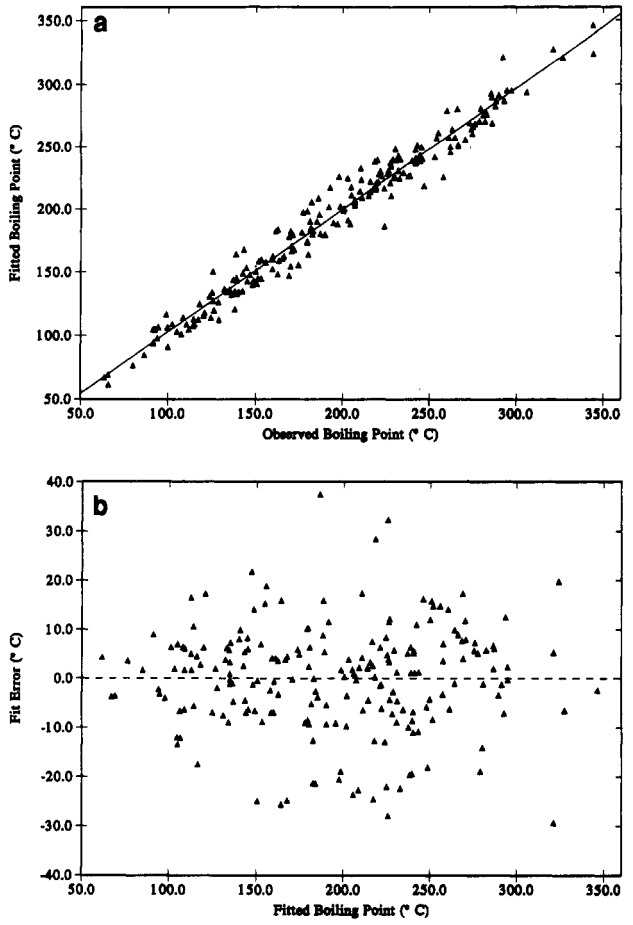

Figure 5. (a) Correlation of the fitted and observed boiling points for the compounds in the final furan/THF class specific model training set. (b) Distribution of the error involved with the fit for the furan/THF class specific model.

fewest terms. The plots given in Figure 4a,b indicate a breaking point at 11 descriptors. The 11 descriptor model gave a good fit to the experimental data (R^2 = 0.968) and a good standard deviation of regression (s = 11.2 °C). This model is summarized in Table II, while the results are shown graphically in Figure 5a,b. Examination of the graph in Figure 5b still indicates a slight skewing of the residuals to the negative side. However, the statistics for this model were quite acceptable, and it was not desirable to reduce the data set further in order to improve it.

The descriptors in the equation were then considered in order to determine what insight they might provide concerning the molecular features of the data set and how they affect observed boiling points. The model contains five descriptors (descriptors 1, 3, 7, 10, and 11) that are based on the topology of the molecules. These are related to the size and shape of the molecule. It is known that boiling point is related to the effective molecular volume, and these descriptors tend to encode that type of information. The model also includes four CPSA-type descriptors (descriptors 2, 6, 8, and 9), which we have found to be important when stronger intermolecular interactions are involved. Since the CPSA-type descriptors combine information concerning the size and shape of the molecule, as well as electronic information, they contribute information to the model concerning both dispersive and polar intermolecular interactions. The final two descriptors (descriptors 4 and 5) contribute information of an electronic nature, which is important for polar-type interactions. Thus, the descriptors involved in the model seem consistent with the understanding of what types of molecular features will have the greatest effect on the boiling points of the compounds involved.

It is necessary to remember that it is not possible to place too much emphasis on the physical interpretation for any single descriptor in this type of equation. This is true for a number of reasons. The diversity of the data set precludes the examination of any one particular molecular feature, and the statistical significance of any given descriptors may depend on any number or combination of molecular features in the data set. Also, because these descriptors are all derived from the same structure, there is a certain amount of information overlap (collinearity) which makes physical interpretation of a single descriptor more difficult. Finally, because of the nature of the regression methods used, the experimental error in the boiling points (as well as any error associated with the descriptors themselves) influences the choice of descriptors to a certain extent.

The development of the predictive model is only the first step of the process. Once one has chosen a model which appears to possess a reasonable fit to the training set compounds, the model itself must be analyzed to determine if any deficiencies exist and to assure that it is robust. This analysis is termed *model validation*, and it is comprised of various statistically based tests which examine properties of the model chosen. If problems are detected within a given model, another would be chosen and the validation process repeated. One advantage of the modeling process described above is that it yields a pool of potentially useful equations which can be drawn upon should problems be detected in one or more of the models chosen for validation.

Validation of the furan/THF predictive model reported here was begun by testing the descriptors (independent variables) for high collinearity. One measure of collinearity is the *variance inflation factor* (VIF), which describes how much the variance of a given descriptor coefficient is inflated above the level which would be present if no collinearity existed.[24] The test involves treating one of the descriptors in the model as a dependent variable, and regressing the one against the remaining descriptors by using a multiple linear regression program. This procedure is repeated for each descriptor in the predictive equation. The VIF values are then calculated as $(1 - R^2_k)^{-1}$, where R^2_k is the coefficient of multiple determination when the kth descriptor is regressed on all remaining descriptors. The objective is to determine the extent of information overlap among the descriptors. If significant overlap is detected, some problems may exist in the equation. Also, it may be possible to reduce the number of descriptors in the model. The highest VIF value obtained for the furan/THF model was 55.6 (descriptor WPSA-3) with a mean VIF value of 15.5. In general, VIF values above 10 are indicative of potential collinearity problems. While the high value obtained is above the usual limit, the calculated standard deviation values of the model coefficients are quite acceptable. Also, one must consider the possible reasons for information overlap in models developed in this fashion. Since all the descriptors are derived directly from the structures of the molecules, it is possible that one structural feature or property is correlated with another feature or property. Even though overlap exists, it is the difference in the information content that makes these descriptors useful in the model. Therefore, before this model was considered unsuitable, other validation tests should be explored.

Next, it is necessary to test the observations in the training set and determine if any of these express undue influence on the model. It is possible that a given point can shift the position of the regression function, and that such an influential observation may go unnoticed in a test of residuals due to an artificially good fit to this point. In order to determine the individual influence of a given observation, one removes it from the training set and recalculates the coefficients of the model.

The new model is then used to predict the observation held out. This process, known as jackknifing, is repeated for each observation. The individual prediction error values are then examined. Of interest are the observations which yield prediction errors much larger than the fit errors obtained when the observation was part of the training set. The results of this test indicated that none of the observations were exhibiting high influence on the regression function.

As a final test of the model, 10% or 21 of the 209 observations from the final training set were chosen at random and stored as a separate prediction set, while the remaining 188 observations were used to recalculate the model coefficients. This process was repeated so as to have five randomly chosen prediction sets. The five remaining sets of 188 observations were then used to recalculate the model coefficients, and the new subset equations were then used to predict the boiling points of the respective randomly chosen prediction compounds. The new models and the predictions were then examined. The model statistics (fit and coefficients) were found to remain stable, and the prediction error was larger but similar to the standard error of the original model. This approach, termed *internal validation*, indicated that the model is still stable when 10% of the original information is removed. Since the prediction sets were chosen at random, it also indicated that the models are indeed broad in scope and that structural features which influence the boiling points of the compounds in the prediction sets are already encoded in the remaining 188. Through the application of the tests described above, the chosen model has been found to be stable and robust. In addition, any problems associated with the large VIF value described above do not seem to degrade the model, and there seems to be no reason to discard it. However, it is important to be aware of the potential for problems associated with collinearity, especially in the case where a new observation for prediction is an extrapolation of the current data set.

The true test of the utility of the model chosen is in its application to the prediction of boiling points for a set of compounds which are similar to the training set, but which were not involved in the development of the model. Such a test is termed *external validation*. Through external validation it is possible to determine the scope of a model and to determine where improvements can be made. For this purpose, a new data set was derived from the Beilstein database by using similar but slightly expanded search criteria. Data for a total of 318 new compounds were collected in this manner.

The results of the prediction of the boiling points for the 318 compound data set, shown graphically in Figure 6a, were generally good. A total of 197 (62.3%) of the compounds yielded prediction errors in the range of ±3.0 standard deviations based on the prediction model (±33.6 °C), and these are shown in Figure 6b. Of the original 318 compounds, 67 yielded large negative prediction errors. This type of error is most probably due to the same type of experimental error which caused several compounds to yield large negative residuals in the early stages of model development. In addition, 53 compounds yielded larger than expected positive prediction errors. The usual cause of such an error (excluding experimental error) is that the new compounds contain some structural feature that is not present in the training set compounds, and which has a significant influence in the boiling point. Examination of the 53 external prediction set compounds that yielded large positive prediction errors showed that 46 of these contained some sort of a fused ring system. Since there were no compounds containing a fused ring system in the training set, this is one possible reason for the large positive prediction errors. There may be other features in the external prediction set which are not well represented in the training set, but these are difficult to detect due to the diversity of the data set and the possible interactions between different features which are not obvious at first glance. However these results are encouraging for two important reasons. First, 62.3% of the external data set was well predicted with this model, indicating that the model is indeed useful. Second, the deficiencies noted actually indicate the direction that future work should take. In order to improve the utility of this model, it is necessary to increase the kinds of functionalities represented in the training set upon which the model is based. It may be possible to develop a model that is general in scope and useful for predictive purposes by examining a wide variety of chemical types.

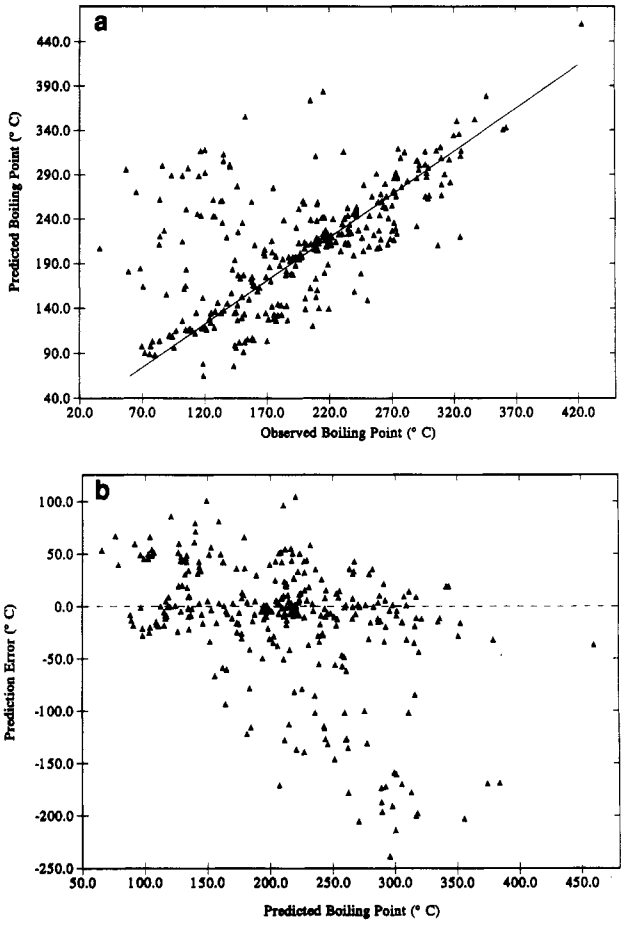

Figure 6. (a) Correlation of the predicted and observed boiling points for the 318 furan-related compounds, based on the use of the furan/THF class specific model. (b) Distribution of the error for the predictions of the 318 furan-related compounds, based on the application of the furan/THF class specific model.

Thiophenes. With the initial success obtained with the furan/THF data set, we decided to examine a new data set which included a greater diversity of compounds and functional group types. The thiophene data set is represented in general terms by structure **III** given in Figure 2. In addition to changing the heteroatom in the ring, halogens and fused ring systems were now included. The initial data set was composed of 223 compounds, which was subsequently trimmed to a set of 195 compounds for a variety of reasons relating to limitations in the software. All the structures were processed as described above. Examination of replicate data points in the experimental boiling point data to be used showed the standard deviation to be 5.2 °C.

Thiophene Training Set Selection. As a first step in examining the new data set, it was of interest to apply the furan/THF model to the prediction of the boiling points of the 195 thiophenes. The necessary descriptors were calculated and stored, and the boiling points were then calculated with the furan/THF model described above. The results of the pre-

Table III. Details of the Thiophene Class Specific Boiling Point Model

$R = 0.987$ ($R^2 = 0.974$), $s = 7.9$ °C, $n = 134$

F value (for AOV) = 697.5

structural descriptor	regression coeff	SD of coeff
(1) no. of single bonds	−6.44	0.86
(2) square root of ALLP-2[a]	32.01	1.82
(3) DPSA-3[b]	1.70	0.22
(4) WNSA-1[b]	−0.78	0.11
(5) mol wt	0.45	0.04
(6) dipole moment	6.11	0.87
(7) radius of gyration	36.97	2.84
intercept	−92.16	

[a] ALLP-2 is the ratio of the number of paths of all lengths (1–46) to the number of atoms in the molecule. [b] See Stanton and Jurs.[23]

dictions were quite good given that the data set involved was dissimilar to the furan/THF training set in many ways. As expected, many of the observations were poorly predicted. Most notable among the poorly predicted compounds were those containing halogens. However, it was also interesting to observe that as the halogen moved farther away from the thiophene ring, the prediction error decreased. This suggested that the furan/THF model may contain information which is related to the effects of halogen (without them being represented in the training set) but not the interaction of the halogens with the thiophene ring. This demonstrates one of the advantages of this methodology over a group contribution method. It may not be necessary to explicitly include all substituent types in all possible combinations, if such information can be included in a more general form by examining a broad variety of compounds.

Another observation from the prediction of the boiling points of the thiophenes was that many compounds exhibited large negative residuals. This is very similar to the observations made during the modeling of the furane/THF data set. Since it is now suspected that there is more error in the reported boiling points than first anticipated, due to some values being determined at a reduced pressure, the compounds yielding large negative prediction errors were set aside. The ability to detect these types of errors early in a study greatly decreases the amount of time necessary to generate good models and allows more effort to be focused on the task of producing good models. In later work, the 20 compounds removed at this step were reexamined. Of the 20, 15 were verified to be in error, two were incorrectly classified as being in error, and the remaining three yielded inconclusive results. These results were also encouraging given the differences in the data sets involved. While there is error in the predictions, these results indicate that the furan/THF model encodes a great deal more information than expected.

Further examination of the 175 remaining thiophenes was done in a manner similar to that used for the furan/THF data set. An iterative application of descriptor analysis and subsequent multiple linear regression and outlier detection was used to obtain a final training set of 134 observations. In all cases where outliers were detected, care was taken to justify their removal on the basis of chemical reasoning.

Thiophene Model Development. With the refined training set in place, it was now possible to begin the development of the predictive equation. Descriptor analysis was repeated and yielded a pool of 88 molecular descriptors to use in regression analysis. From a set of possible models obtained in regression, a reasonable model was selected in the same manner used for the furan/THF data set. The model obtained is given in Table III while the results are given graphically in Figure 7a,b. The correlation of fitted and observed boiling points is good, and

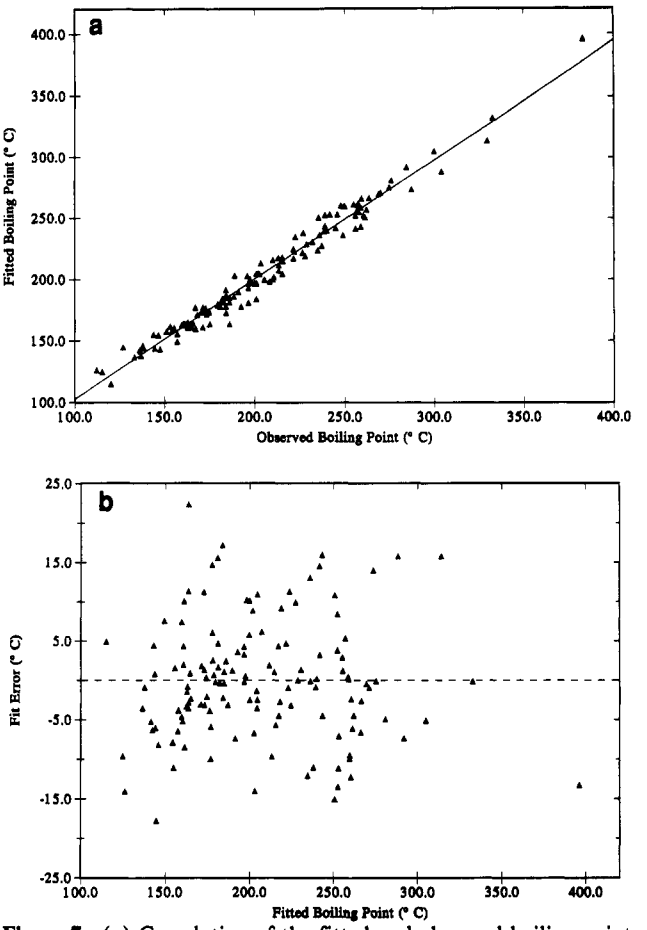

Figure 7. (a) Correlation of the fitted and observed boiling points for the thiophene class specific model. (b) Distribution of the fit error for the thiophene class specific model.

the fit error is evenly distributed over the temperature range with no apparent pattern. One point, which could be a potential outlier, exists alone at the high end of the temperature range, as is clear in Figure 7a. However, the regression function does not change significantly when the point is removed, and therefore it was retained.

The model equation exhibits some interesting characteristics. There are fewer descriptors in the thiophene model than were required for the furan/THF model, while exhibiting a similar fit to the observed boiling points. Two of the descriptors are of the CPSA type, which follows the trend noted in the furan/THF portion of the study. There is also a descriptor that is a transformed version of the ALLP-2 descriptor. The square root transformation was chosen through the examination of the correlation of the dependent variable with a number of the descriptors that survived the final phase of descriptor analysis. It was encouraging that two descriptors (molecular weight and dipole moment), which one would intuitively expect to see in such a model, are included. Size and shape information, features which also are known to affect the boiling point, is also supplied by the remaining descriptors. Thus, the physical relationship between boiling point and the descriptors in the model can be understood, and the model is appealing from that perspective.

Validation of the thiophene model was accomplished in a manner similar to that used for the furan/THF model. Jackknifing showed that none of the 134 observations in the training set were overly influential on the regression function and that the position of the regression line did not rely greatly on any one observation. Next, five randomly selected sets of 14 observations (roughly 10% of the final training set) were set aside for the purpose of internal validation, as previously described. Only one of the five models exhibited some deg-

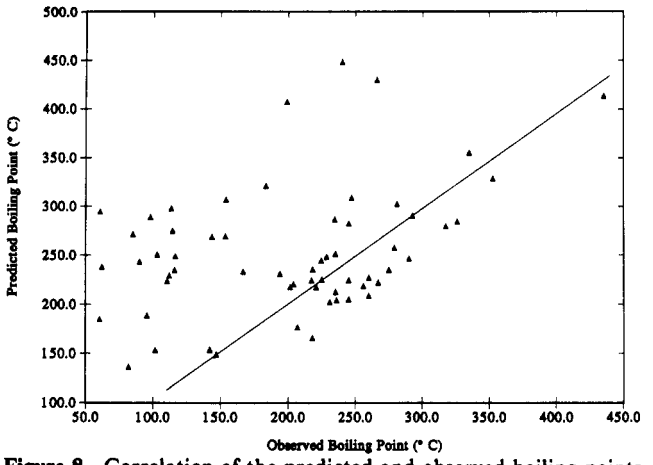

Figure 8. Correlation of the predicted and observed boiling points for the 61 thiophenes set aside during data analysis.

radation, indicating the model was fairly robust. As a final test of internal validation, the 134 observation training set was split in half, with one portion being used to recalculate the model coefficients while the other portion was set aside as a prediction set. The data splitting was accomplished using the DUPLEX method of Snee.[30] This particular test was applied because no external data set was available and it was desirable to demonstrate the validity of the final model.

The recalculated subset model yielded a good fit (R^2 = 0.972) and a standard deviation of regression of 7.9 °C, which is in good agreement with the 134 observation model. The subset model was then applied to the prediction subset. The predicted boiling points were in good agreement with the observed values, with a coefficient of simple correlation between the two of 0.984. The overall RMS prediction error for the subset was 8.8 °C, which is also in good agreement with the 134 observation model. These results demonstrate the final thiophene model to be stable and robust.

The final experiment to be done involved the reexamination of the 61 outliers that had been eliminated in various steps in the refinement of the data set. The 61 observations were collected into one set, and the boiling points were calculated by using the final thiophene equation. The results of the test are given graphically in the plot in Figure 8. Examination of the plot shows a large spread of points away from the line which represents the training set correlations. Many of the points exhibit large negative residuals, indicative of the boiling point being determined at reduced pressure. There are several points which roughly follow the line. These observations may be the result of some type of experimental or procedural error encountered when the measurement of the boiling point was made, such as an impurity in the sample or a poorly calibrated thermometer. Finally, there are a few points which fit the line rather well. These are most likely to be points that were removed along with true outliers. It is possible that the regression models developed along the way, and especially those obtained during the early part of the study, may have been influenced by the large number of outliers encountered and that such points could shift the regression function, thus making some valid points appear as outliers.

Combined Data Set Modeling. Once the individual furan/THF and thiophene data sets had been modeled, it was of interest to develop a more global model that involved compounds from both data sets. Since much of the data set refinement had already been accomplished, more attention could be paid to developing and validating models.

A total of 343 compounds were made available by combining the final training sets from both the furan/THF data set and the thiophene data set. Since it was desirable to have an external prediction set, a training set containing 240 of the 343 available compounds was chosen at random, leaving the remaining 103 compounds to act as an external prediction set. To make the work of predicting boiling points for the external set easier, descriptors were calculated as previously described for the entire set of 343 compounds. Modeling then focused only on the 240 observation training set. Descriptor analysis was performed as before, yielding a final pool of 77 descriptors. These 77 explanatory variables were then screened in regression analysis using interactive regression analysis (IRA). Within IRA, the user can build models by manually adding or deleting variables while the program reports the necessary statistics for the current model and the pool of remaining descriptors. The method of interactive regression analysis was chosen for two reasons. The first is related to the size of the variable pool that can be screened. The method of IRA allows the analysis of a pool of 50 descriptors, as opposed to the maximum of 34 allowed by routines used in the past. This has the advantage of allowing more combinations of different variables to be examined, which can lead to better models. The second reason for using the IRA method is that it places all the control of the regression analysis in the hands of the chemist. Thus, the variable addition and deletion steps are under user control, and decisions concerning the analysis can be made on the basis of both chemical intuition and statistical significance, rather than strictly a statistical basis as in the more automated regression methods.

A twelve variable model was produced in this fashion and was found to yield fitted boiling points that correlate well with the observed values for the 240 observation training set. As a first step in validating the combination data set model, the process of jackknifing was applied as described previously. In this manner, four compounds were shown to yield large jackknife residuals, suggesting they have a large influence on the regression function. It was also determined that these four compounds were all from the furan/THF training set. Since that data set was not as carefully screened as the thiophene data set was these compounds were set aside, and the coefficients of the model were recalculated on the basis of the remaining 236 compounds in the training set. The resulting model is detailed in Table IV, and the correlation of the fitted and observed boiling points is shown graphically in Figure 9a, while the distribution of the fit error is given in Figure 9b.

The results of this model are quite good. The standard deviation of regression (s) is 9.6 °C, which is 4.9% of the mean boiling point for the data set. Examination of the plots in

Table IV. Summary of the Furan/THF/Thiophene Combination Data Set Boiling Point Prediction Equation Based on 236 Observations

R = 0.987 (R^2 = 0.974), s = 9.6 °C, n = 236

F value (for AOV) = 709.9

structural descriptor	regression coeff	SD of coeff
(1) no. of single bonds	−17.2	1.03
(2) valence corrected path 1 molecular connectivity	29.2	2.71
(3) molecular ID no. of atoms in molecule[a]	159.0	36.55
(4) sum of molecular IDs for all heteroatoms[a]	3.2	0.49
(5) av sum distance connectivity[b]	16.3	4.59
(6) total solvent accessible surface area of molecule	0.69	0.065
(7) FPSA-2[c]	205.4	14.47
(8) FNSA-3[c]	−1236.6	55.80
(9) DPSA-2[c]	−0.35	0.031
(10) RPCG[c]	65.6	12.24
(11) dipole moment	5.1	0.67
(12) molecular polarizability	4.5	0.89
intercept	−551.1	

[a] See Randić.[33] [b] See Balaban.[30] [c] See Stanton and Jurs.[23]

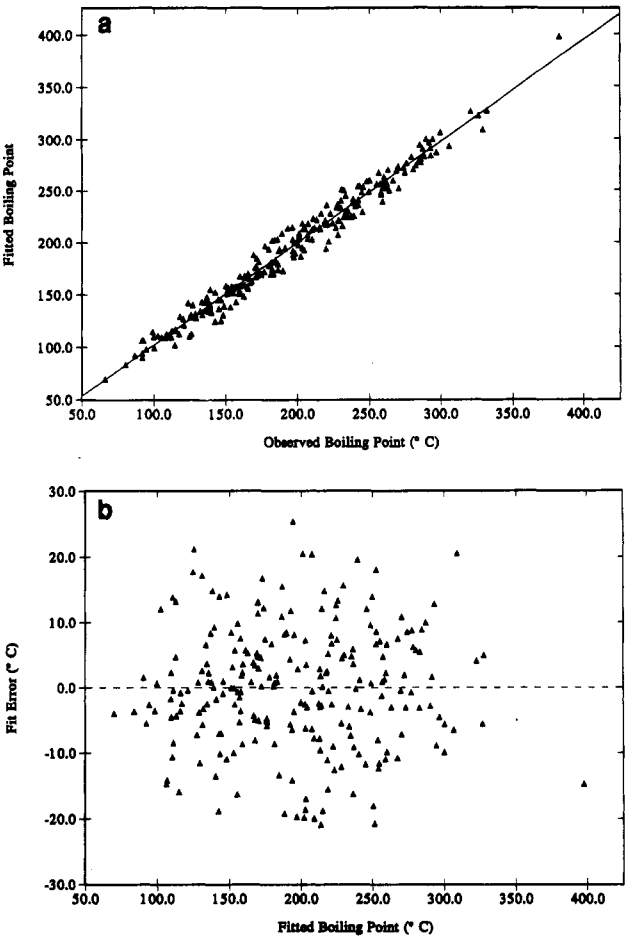

Figure 9. (a) Correlation of the fitted and observed boiling points for the compounds in the combined furan/THF/thiophene training set. (b) Distribution of the fit error for the combined furan/THF/thiophene data set model.

Figure 9 indicate a good correlation between the fitted and observed boiling points for the 236 compounds (R = 0.987), and the fit error is evenly distributed and shows no apparent pattern. One point (BRN-160370) is observed in the plots as having a potentially high influence on the regression function. However, the jackknife estimate and the fit estimate of the boiling point for this compound are not very different, and removal of the point has little effect on the resulting regression equation, so the point was retained.

Validation of the model had already been started with the calculation of the jackknifed estimates. Further examination of the model itself involved the calculation of the variance inflation factors of the descriptors. The mean VIF value was 14.9, with a high of 54.6 (DPSA-2) and a low of 2.0 (WTPT-2). These values suggested that a potential problem existed with regard to collinearity in the model. However, these results were very similar to those observed for the furan/THF class specific model. It was decided that the final test would be the application of the model to the 103 compounds which were set aside as an external prediction set.

The model was applied to the external data set, and good results were obtained. The RMS prediction error for the 103 compounds was 14.0 °C (6.8% of the mean boiling point for the prediction set), and the predicted and observed boiling points yielded a correlation coefficient of 0.969. The correlation of the predicted and experimental boiling points for the 103 compounds is shown graphically in Figure 10. These results suggest that the model is robust and that any collinearity which exists in the model is not causing severe problems.

It was interesting to note some of the characteristics of the fitted boiling points values for the 236 observation training

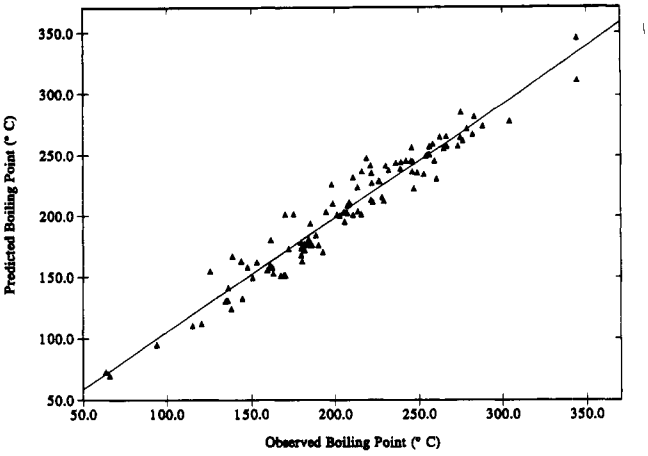

Figure 10. Correlation of the predicted and observed boiling points obtained for the 103 observation prediction set by using the combined class model.

set. The spread of the fit error for the thiophene portion of the training set was less than that for the furan/THF portion. The standard deviation of the thiophene fit error was 12.1 °C, while the standard deviation for the furan/THF portion was 15.3 °C. This indicates that there is more error associated with the furan/THF portion of the training set, and this is most likely to be related to the greater emphasis placed on the analysis of the thiophene data set. This is due to the increased awareness of the problems associated with the error in the experimental data and the improvement of the methodology used in the data analysis. Therefore, the greater amount of error noted for the fitted values of the furan/THF portion is not an indication of a bias in the model, but rather it reflects a difference in the quality of the respective data set portions.

As a final experiment, the combined data set model was applied to the prediction of the boiling points of the 318 observations in the large furan/THF prediction data set. This was done to determine if the predictions for the external data set improved with the addition of the information from the thiophenes to the model. The results of the predictions are shown graphically in Figure 11a,b. A noticeable reduction in the amount of positive error was observed, with little change in the amount of negative prediction error. These results are expected if the quality of the model has been improved. Since the negative error is related to boiling point values that were determined at reduced pressure, we would not expect these to change. However, since it is assumed that the large positive prediction errors noted previously are a result of deficiencies in the predictive equation, we expect to observe the degree of positive prediction error to diminish as the model is improved. This premise is supported by the results obtained. Thus, information concerning molecular features that effect the observed boiling points has been added to the predictive equation. Since the source of this additional information is the thiophene data set, features other than just the parent ring system must have been included.

CONCLUSIONS

Predictive regression equations for normal boiling points, based only on molecular structural descriptors, have been developed and have been found to yield good results for a wide variety of furan-, tetrahydrofuran-, and thiophene-containing compounds. The models developed indicate that the boiling point is a function of the size and shape of the molecule as a whole, as well as of the type of functionality present in the molecules, and do not show specific divisions based on the parent ring system. These predictive equations include structural descriptors which agree with the intuitive understanding of the features that affect the boiling points of these

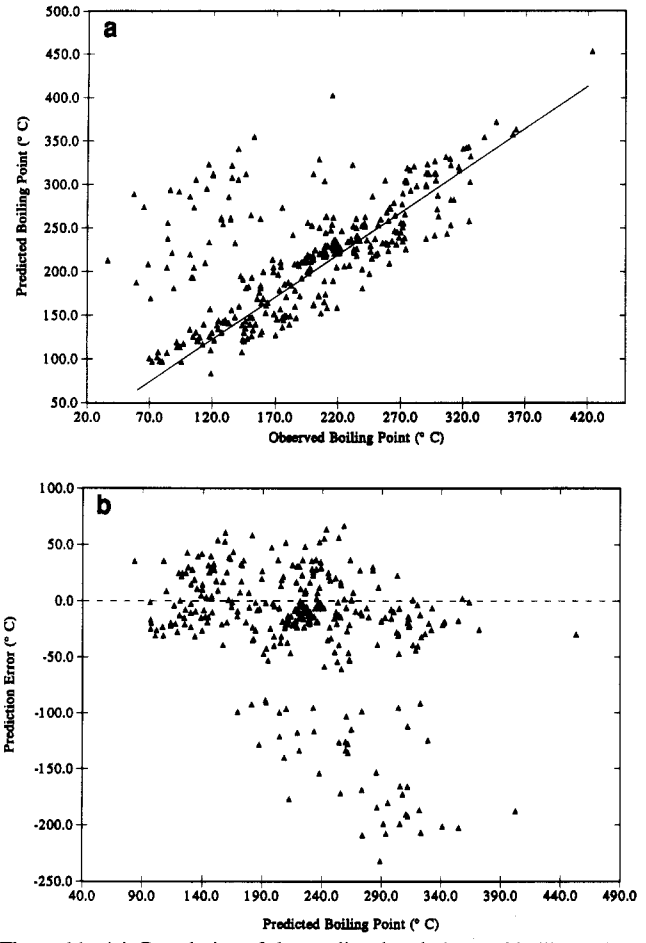

Figure 11. (a) Correlation of the predicted and observed boiling points for the 318 furan-related compounds, based on the furan/THF/thiophene combined class model. (b) Distribution of the prediction error obtained by using the combined furan/THF/thiophene model for the 318 furan-related prediction data set.

compounds, although additional work is necessary before a more complete understanding can be obtained. This can be accomplished by the examination of additional data sets containing different features than those represented in the three data sets examined here.

The results of the data analysis required to develop the models indicate the advantage of using this methodology for data sets of this type over methods which rely on group additivity principles. A data set containing significant amounts of error can be refined while allowing for the identification of observations exhibiting large amounts of experimental error. This form of quality control for experimental data can be useful, especially in very large databases where manual checking is impractical. In addition, new data can be screened and potential errors can be detected before being added to the database, thus preventing contamination of the database. Additional work is necessary to broaden the scope of the combined data set model in order to allow it to be of general use. However, the class-specific models are of value when it is necessary to have a more accurate prediction value for a new compound, when that compound is similar to a class for which a model has been developed.

ACKNOWLEDGMENT

The funding for this work was provided by the Beilstein Institute. Partial funding was also provided by the National Science Foundation for the purchase of the Sun 4/110 workstation.

Supplementary Material Available: Four tables showing combination model training set data and combination model prediction set data (including structures and Beilstein Registry Numbers), identification and data for furans and tetrahydrofurans in the final training set for the class-specific model, results of predictions of boiling points for furan external prediction set, and identification for thiophenes in the final class-specific training set (77 pages). Ordering information is given on any current masthead page.

REFERENCES AND NOTES

(1) Shriner, R. L.; Curtin, D. Y.; Fuson, R. C.; Morrill, T. C. *The Systematic Identification of Organic Compounds*, 6th ed.; John Wiley: New York, 1980.
(2) Pearson, D. E. *J. Chem. Ed.* **1957**, *28*, 60–62.
(3) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982.
(4) Lai, W. Y.; Chen, D. H.; Maddox, R. N. *Ind. Eng. Chem. Res.* **1987**, *26*, 1072–1079.
(5) Joback, K. G.; Reid, R. C. *Chem. Eng. Commun.* **1987**, *57*, 233–243.
(6) White, C. M. *J. Chem. Eng. Data* **1986**, *31*, 198–203.
(7) Bermejo, J.; Blanco, C. G.; Guillén, M. D. *J. Chromatogr.* **1985**, *331*, 237–243.
(8) Cramer, R. D., III *J. Am. Chem. Soc.* **1980**, *102*, 1837–1849.
(9) Cramer, R. D., III *J. Am. Chem. Soc.* **1980**, *102*, 1849–1859.
(10) Cramer, R. D., III *Quant. Struct.-Act. Relat.* **1983**, *2*, 7–12.
(11) Hansen, P. J.; Jurs, P. C. *Anal. Chem.* **1987**, *59*, 2322–2327.
(12) Smeeks, F. C.; Jurs, P. C. *Anal. Chim. Acta* **1990**, *233*, 111–119.
(13) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(14) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979; pp 103–129.
(15) *Machine-Readable Connection Tables for Input to CAS Registry Services*; Chemical Abstracts Service: Columbus, OH, 1986.
(16) Allinger, N. L.; Yuh, Y. H. *Molecular Mechanics, Operating Instructions for MM2 and MMP2 Programs, 1977 Force Field.* Quantum Chemistry Program Exchange, Program No. 395, 1980.
(17) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.
(18) Abraham, R. J.; Smith, P. E. *J. Comput. Chem.* **1988**, *9*, 288–297.
(19) Kaliszan, R. *Quantitative Structure-Chromatographic Retention Relationships*; Wiley-Interscience: New York, 1987.
(20) Miller, K. J.; Savchick, J. A. *J. Am. Chem. Soc.* **1979**, *101*, 7206–7213.
(21) Vogel, A. I. *Elementary Practical Organic Chemistry Part 2: Qualitative Organic Analysis*, 2nd ed.; John Wiley: New York, 1966.
(22) Hanch, C.; Leo, A.; Unger, S. H.; Kim, K. H.; Nikaitani, D.; Lien, E. J. *J. Med. Chem.* **1973**, *16*, 1207–1216.
(23) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323.
(24) Small, G. W.; Jurs, P. C. *Anal. Chem.* **1983**, *24*, 164–175.
(25) *Minitab Reference Manual*, release 7.2; Minitab: State College, PA, 1989.
(26) Rousseeuw, P. J. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880.
(27) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; John Wiley: New York, 1987.
(28) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1985; pp 391–393.
(29) Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; John Wiley: New York, 1980.
(30) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
(31) Kier, L. B. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109.
(32) Snee, R. D. *Technometrics*, **1977**, *19*, 415–428.