Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles

David T. Stanton,† Leanne M. Egolf, 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, Germany

Received July 20, 1991

Computer-assisted methods are applied to the development of predictive models for the normal boiling points of diverse sets of pyrans and pyrroles. The models developed employ molecular structure based parameters or *descriptors* to encode the features of the compounds which determine the boiling point. A set of 20 descriptors is identified that allows for the development of good quality models for the pyrans and for sets of furans, tetrahydrofurans (THFs), and thiophenes, which have been studied previously. A model is presented which yields good predictions for a combined set of pyrans, furans, THFs, and thiophenes. The scope of this work is expanded to include nitrogen-containing heterocycles through the study of a diverse set of pyrroles. As part of this work, a new set of descriptors is developed for the purpose of capturing information concerning the molecular features responsible for intermolecular hydrogen-bonding interactions. Finally, the pyrrole dataset is combined with a large set of furans, THFs, thiophenes, and pyrans for the purpose of producing a more general boiling point prediction equation. The results of these studies are examined to determine their impact on future work.

INTRODUCTION

In a previous paper,[1] the results of studies involving the development of boiling point prediction equations for sets of heterocyclic organic compounds (furans, tetrahydrofurans, and thiophenes) using quantitative structure–property relationship (QSPR) methods and the ADAPT[2,3] system were described. The purpose of those studies was to provide the means of filling gaps and detecting errors in the Beilstein Institute physical property database. In addition, a clearer understanding of the structure–property relationship for boiling point was sought. Because of the diversity of the compounds in the database and the nature of the goal of the work, the experimental data necessary for the studies was drawn directly from the Beilstein database. Also discussed were the results of our first attempt to produce a boiling point prediction model based on a dataset that contained observations from both of the individual classes of compounds. The objective of that study was the development of a more global model, useful for estimating the boiling points of a wider range of compounds. The current study represents the continuation of the pursuit of a more global model through the study of a diverse set of pyrans.

While the datasets studied to date have included several compounds involving nitrogen-containing heterocyclic ring systems in addition to the ring systems described above, such compounds represent only a small fraction of the datasets studied. The scope of this work is further expanded here to include a diverse set of five-member nitrogen-containing heterocyclic compounds which will be loosely referred to as *pyrroles*. These compounds will also be combined with the set of furans, tetrahydrofurans, thiophenes, and pyrans examined previously in order to produce a model which incorporates the important features of all these datasets.

Intermolecular hydrogen bonding was expected to be a significant factor contributing to the observed boiling points of the pyrroles included in the new dataset. Strong (polar) intermolecular interactions have historically been more difficult to account for in QSPR studies because the available molecular structure parameters (or descriptors) failed to encode sufficient information concerning the structural features that were responsible for these interactions. In our laboratory, we have developed a set of charged partial surface area (CPSA) molecular structure descriptors that have been found to be useful for modeling properties which are influenced by polar intermolecular interactions.[4] However, these descriptors were developed to be general in nature and not designed for any particular type of polar interaction. Due to the composition of the pyrrole dataset, it was of interest to develop a new set of molecular structure parameters which would capture information specific to those structural features participating in intermolecular hydrogen bonding. A new set of hydrogen-bonding specific parameters is presented which are calculated in a manner similar to that used for the CPSA descriptors. Their use in modeling the normal boiling points of the pyrroles is discussed.

The goals of this study, which are similar to that for the furan/THF and thiophene datasets, are twofold. First, it was necessary to screen the available structural and experimental data in order to produce a good quality training set. This is necessary due to the existence of error involving some of the normal boiling point values obtained from the Beilstein database (boiling points determined at reduced pressure). Also, because of the broad criteria used to extract the initial pyran and pyrrole subsets from the Beilstein database, many compounds were included which would not strictly be classified as similar to the bulk of the dataset. These have to be detected and removed before modeling can begin.

The second goal of the study is to produce a good quality model for the individual pyran and pyrrole datasets and then to model combined datasets of furans, tetrahydrofurans, thiophenes, pyrans, and pyrroles. It is then necessary to establish the stability and robustness of these models. From the results of these studies, we hope to learn more about the connection between molecular structure and boiling point for these compounds.

† Present address: Procter & Gamble Pharmaceuticals, P.O. Box 191, Norwich, NY 13815.

0095-2338/92/1632-0306$03.00/0

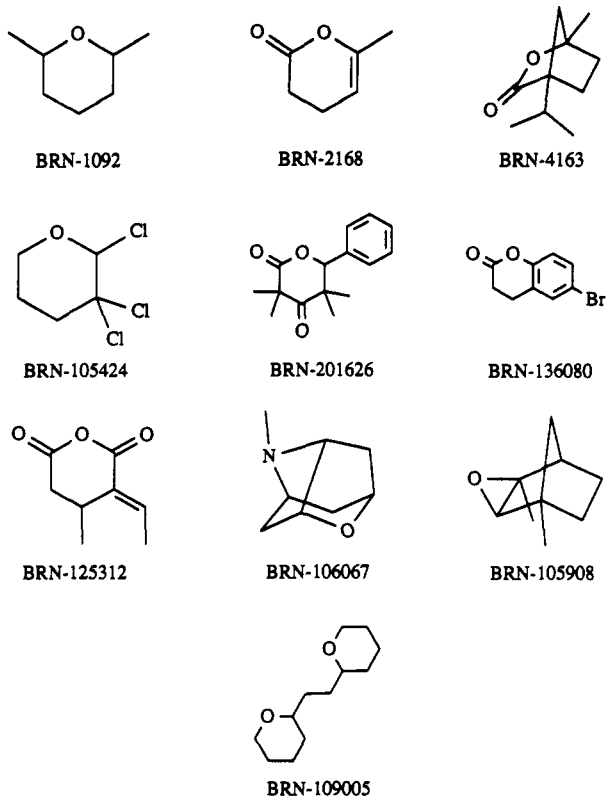
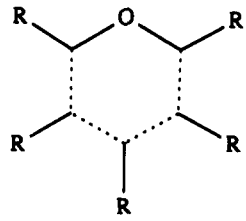


Figure 1. Structures of selected compounds taken from the pyran dataset, identified by Beilstein Registry Number (BRN).

METHODOLOGY

The procedures used in this study have been described previously.[1] All computations were performed on a Sun 4/110 workstation using the ADAPT software system running under the UNIX operating system.

Pyran Dataset. The structural information for 231 pyrans was received from the Beilstein Institute in the form of connection tables. The general structure for the compounds included in the pyran dataset (A) is given below, where the



dashed bonds indicate variable bonding. The dataset included a wide variety of compounds, examples of which are shown in Figure 1. Prior to entering the structures into ADAPT data files, the connection tables were examined manually, and compounds which could not be studied due to software limitations were removed. In addition, very complex compounds or those containing unusual substituents were also set aside. Eighteen compounds were set aside for these reasons, leaving 213 compounds. The remaining connection tables were processed as described in previous work and stored in ADAPT format. Aromatic ring systems were then converted from alternating single–double bonds to the ADAPT aromatic bond type. Once the structures were entered and stored, reasonable low-energy conformations were obtained using molecular mechanics programs.

Pyrrole Dataset. The structural information for 395 pyrroles was received from the Beilstein Institute in the same

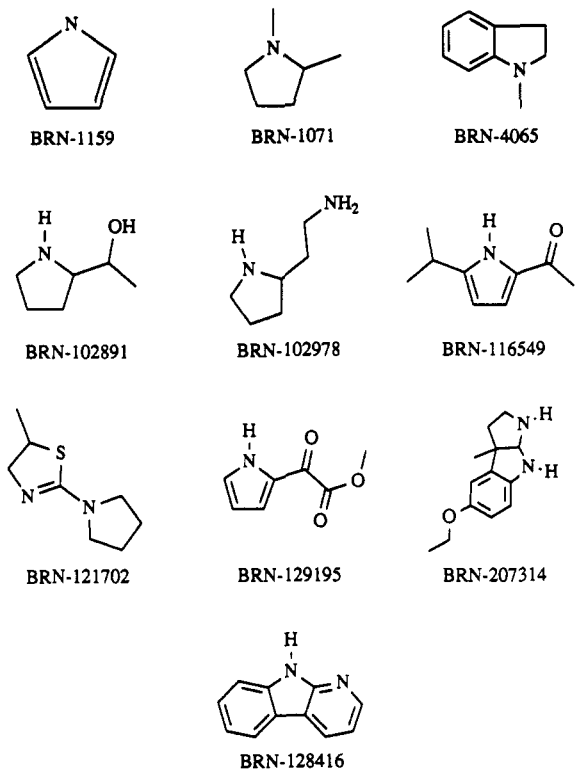
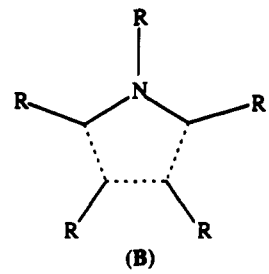


Figure 2. Structures of selected compounds taken from the pyrrole dataset, identified by Beilstein Registry Number (BRN).

format as was used for the pyran dataset. The general structure for the compounds included in the pyrrole dataset (B) is given below. The dataset was composed of a wide variety of



compounds containing the general ring structure noted. Examples selected to demonstrate the diversity of the dataset are shown in Figure 2. Processing of the pyrrole dataset was carried out in the same fashion used for the pyran dataset. A total of 377 compounds remained after this initial processing.

Boiling Point Data. The experimental normal boiling point data were received from the Beilstein Institute as an ASCII file containing boiling points (or ranges) and pressure data. The data were evaluated using a FORTRAN program which calculated and reported the mean boiling point for each compound in the list. Since it was of interest to model the normal boiling point, only those observations for which the reported pressure was within a range of 10 mmHg of standard atmospheric pressure (760 mmHg) were considered. Observations for which no pressure data were reported were assumed to have been determined at standard pressure. For compounds with multiple observations, the standard deviation of the experimental boiling point values was reported along with the mean boiling point. Data for compounds which yielded an unusually high variation in the experimental values were examined. If the problem could be resolved by examination of the experimental data, the compound was retained; otherwise, it was dropped from further consideration. Twenty-

eight of the 213 pyrans which remained after examination of the structural data were set aside in this step, leaving a total of 185 pyrans to be considered in subsequent steps. No members of the pyrrole dataset were removed at this step.

Descriptor Generation and Analysis. Once the energy-minimized structures were available, molecular structural descriptors were calculated. The descriptors chosen for calculation were those types that had been shown to be useful in past work. These capture structural information derived from the topologic, geometric, and electronic representations of the molecules. In addition, parameters which have been termed charged partial surface area or CPSA descriptors[4] and which were found to be particularly useful in past work were also calculated. In all, 126 different molecular structure descriptors were calculated. Among these were the descriptors that were included in the combined (furan/THF, thiophene) dataset model.[1] For the purposes of clarity, this first combined dataset model will be referred to as the *Combo-1* model.

Prior to their use in regression analysis, the descriptors were objectively examined. In this process, a descriptor is discarded if the majority (≥90%) of the values for the dataset are identical, since such a descriptor provides no means of discrimination between compounds. Also, it is often the case that a pair of descriptors are highly correlated (coefficient of simple correlation $r \geq 0.95$). Only one descriptor from each pair need be retained because they possess nearly the same information and such high correlations can cause problems during regression analysis. The descriptor retained is chosen on the basis of past utility and physical interpretability.

Design of the Hydrogen Bonding Descriptors. The set of CPSA descriptors has been shown to be quite useful in a variety of studies.[1,5] The combination of molecular surface area and atomic charge characteristics has advanced the study of physical properties of polar compounds. The structural features responsible for intermolecular hydrogen bonding were next studied with a similar approach.

The theory of hydrogen bonding was examined to determine which functional groups act together to form hydrogen bonds and how hydrogen bonding can have such a strong influence on the observed normal boiling points of organic compounds. The information used was drawn from Pimentel and McClellan[6] and Vinogradov and Linnell.[7] Each provided a broad coverage of the subject and included many examples and comprehensive bibliographies.

Hydrogen bonding is a strong interaction between two functional groups. The *acceptor* group provides an electron pair, and the *donor* group provides a proton. The proton is thought of as being *shared* between two atoms (usually heteroatoms). An equilibrium exists between the associated and nonassociated molecules in the bulk phase. The effect of intermolecular hydrogen bonding is an increase in the apparent molecular weight of the compound. Thus, one observes a higher than expected normal boiling point for compounds which participate in intermolecular hydrogen bonding. The exceptions to this are compounds which contain both an acceptor and a donor functional group. Given the right configuration, these two groups would form an *intramolecular* hydrogen bond. In such a case, the donor and acceptor groups are not available for intermolecular hydrogen bonding, and as a result, one would observe a normal boiling point similar to a compound that is unable to form intermolecular hydrogen bonds.

It was necessary to keep such factors in mind while attempting to develop useful descriptors to encode hydrogen-bonding characteristics. In this particular study, intramolecular interactions could not be considered because of the difficulty involved in determining the correct geometric orientation of the molecule. For that purpose, it would be necessary to perform molecular modeling on the intramolecular hydrogen-bonded species. However, none of the molecular mechanics programs available within ADAPT will allow for the minimization of the strain energy for such a compound. Thus, for this study, it was necessary to consider other approaches to detect intramolecular interactions if discrepancies were noted in modeling boiling points of compounds that could participate in such interactions.

Table I. Labels and Definitions for the 11 Hydrogen-Bond Specific Molecular Descriptors[a]

label	definition
SSAH	sum of the surface areas of hydrogens which can be donated
CHGD	maximum difference in charge between a hydrogen which can be donated and its covalently-bonded heteroatom
ACGD	average difference in charge between all pairs of H-bond donors
SSAA	sum of the surface areas of all H-bond acceptor groups
CNTH	simple count of all H-bond donor groups
CNTA	simple count of all H-bond acceptor groups
RHTA	ratio of the number of donor groups to the number of acceptor groups
RSAH	average surface area of hydrogens which can be donated
RSAA	average surface area of H-bond acceptor groups
RSHM	fraction of the total molecular surface area associated with hydrogens which can be donated
RSAM	fraction of the total molecular surface area associated with H-bond acceptor groups

[a] Molecular surface area is taken as the solvent-accessible surface area. Charges refer to the partial atomic charge on a given atom.

A set of 11 descriptors was devised for encoding the hydrogen-bonding features of a molecule. These 11 descriptors are listed in Table I. These represent a first attempt at producing these types of descriptors and were selected to examine a variety of ways of encoding the necessary information. Hydrogen-bond donor groups were considered as being any heteroatom (i.e., O, S, or N) that possessed a proton that can be donated. Other types of functional groups (e.g., terminal alkynes) were also included in the donor class. Acceptor groups included any functional group which possessed sufficient electron density to participate in a hydrogen bond. Although halogens and certain types of double (or aromatic) bonds can participate as acceptor groups in hydrogen bonding,[7-9] these were not included in this study in order to simplify this first attempt. The solvent-accessible surface areas and the partial atomic charges used to derive the new hydrogen-bond descriptors were calculated using the methods outlined for the CPSA descriptors.[4]

Regression Analysis. The methods of multiple linear regression analysis employed in this work differ from those used previously. Two approaches were used—interactive regression analysis (IRA) and leaps-and-bounds regression.[10] Interactive regression analysis is a generic approach that places all the control of the analysis in the hands of the user. This differs from other automated methods where the independent variables are chosen on the basis of an algorithm. In IRA, the user is free to select any combination of variables for the equation while the program provides the necessary statistics. In this fashion, it is possible to explore a wide variety of descriptor combinations.

The method of leaps-and-bounds regression is an automated method of regression analysis which approximates the method of all possible regressions. The current implementation of

this approach in ADAPT is the program LEAP. While LEAP performs a more complete analysis of the descriptor set with which it is presented, it is limited to the examination of 24 independent variables at a time, while a set of 49 descriptors can be examined using IRA. Thus, the user is limited to using IRA when larger sets of descriptors are available.

RESULTS AND DISCUSSION

Pyran Training Set Selection. Before a predictive model can be developed, a training set must be selected. The necessary characteristics of this training set include good quality experimental data for compounds which can be considered to be reasonably similar. The concept of similarity is subjective, and the previous study has shown that a final training set can include structure types which are quite diverse.

As a first experiment, the Combo-1 model was used for the prediction of the boiling points of the pyrans. The boiling points for the 185 pyrans were estimated and compared with the experimental data. Twenty-seven of the 185 compounds yielded absolute prediction error values in excess of an arbitrarily chosen cutoff of 40 °C. As has been noted in previous work, many of the compounds yielded large negative prediction errors. In past work, this type of error has been shown to indicate that the reported normal boiling point was actually the boiling point determined at reduced pressure. However, the structures and experimental boiling point data for all 27 compounds were examined in an attempt to discover a reason for the error. On the basis of the examination, 5 of the 27 compounds were retained for modeling, and the remaining 22 were dropped from further consideration leaving a training set of 163 observations.

Evaluation of the remaining compounds was begun using the 163-observation training set. A subset of 66 descriptors remained after descriptor analysis for this dataset. Regression analysis using the IRA method was then performed. A model containing nine descriptors was obtained that yielded calculated boiling points which were in reasonable agreement with the experimental data ($R^2 = 0.941$, $s = 16.7$ °C). However, the quality of this result was less than that obtained for previous work. Therefore, the process of data analysis described for both the furan/THF and thiophene datasets was begun. This analysis involved the refinement of the dataset using multiple linear regression and robust regression techniques,[11] which have been described in detail in the previous paper.[1] The result of this iterative refinement technique was a final training set for the pyrans which contained 146 observations.

Pyran Model Development. With the selection of the final training set membership complete, the process of developing the predictive model was begun. The steps of descriptor analysis and subsequent multiple linear regression analysis were applied as outlined in preceding sections. The regression method of choice for this study was IRA. The model obtained for the 146-observation dataset is shown in Table II. The model employs seven descriptors and yields a good correlation between the calculated and observed boiling points for all 146 observations. This correlation is shown graphically in Figure 3. Examination of the residuals shows the fit error to be evenly distributed over the range of predicted values and shows no apparent pattern. Table III gives the observed and fitted boiling points along with the fit error for the selected compounds shown in Figure 1.

The first step in model validation involved the calculation of jackknifed residuals.[12] It was noted that three compounds yielded larger than expected jackknifed residuals. However, it was found that removal of these observations from the training set made no difference in the fit of the equation or coefficients of the descriptors of the resulting model, so they were retained.

Table II. Boiling Point Prediction Model Based on the 146-Observation Pyran Class-Specific Dataset

$R^2 = 0.978$, $s = 10.2$ °C, $N = 146$

descriptor	regression coeff	SD coeff
no. of atoms	27.95	1.00
no. of single bonds	−7.28	0.55
valence corrected 3rd order path mol connectivity	22.69	2.23
6th order path-cluster mol connectivity	−6.65	0.52
FNSA-3	-1.20×10^3	67.15
WNSA-2	0.74	0.12
max positive charge	90.90	14.76
intercept	−97.64	

Table III. Example Results from Regression Analysis for Pyran and Pyrrole Datasets[a]

BRN No.	obsd boiling point, °C	fitted boiling point, °C	fit error, °C
	Pyran Dataset Results		
BRN-1092	114.7	114.7	0.0
BRN-2168	194.5	208.6	−14.1
BRN-4163	247.0	239.4	7.6
BRN-105424	212.0	210.0	2.0
BRN-105908	180.0	180.4	−0.4
BRN-106067	152.0	166.1	−14.1
BRN-109005	263.0	236.6	−0.6
BRN-125312	300.0	297.2	2.8
BRN-136080	321.0	306.4	14.6
BRN-201626	325.0	338.5	−13.5
	Pyrrole Dataset Results		
BRN-1071	94.9	99.3	−4.4
BRN-1159	130.3	119.3	11.0
BRN-4065	227.5	228.4	−0.9
BRN-102891	190.8	178.6	12.2
BRN-102978	170.0	147.4	22.6
BRN-116549	251.0	245.9	5.1
BRN-121702	247.5	253.1	−5.6
BRN-128416	363.6	332.7	30.9
BRN-129195	285.0	265.5	19.5
BRN-207314	309.0	309.1	−0.1

[a] Structures for these compounds are shown in Figures 1 and 2.

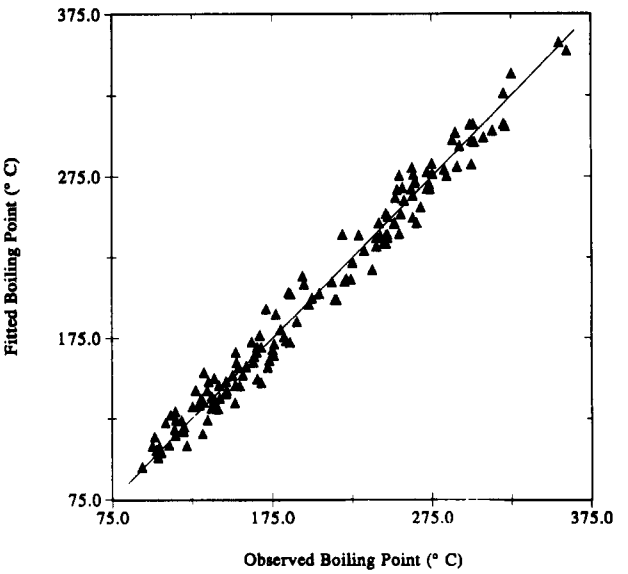


Figure 3. Scatter plot of the fitted and observed boiling points for the 146-observation pyran class-specific model.

The next test of the quality of the pyran-specific model involved setting aside 14 randomly chosen observations from the training set, followed by the recalculation of the model coefficients based on the remaining 132 observations. This process was repeated a total of five times. The results of the test indicated that the coefficients for the models based on the reduced datasets were stable within the standard deviations of the coefficients of the original model and that the coefficient of multiple correlation and standard deviation of regression changed little. Thus, it can be concluded that the model is stable at the 10% deletion level.

A final test of the robustness of the model was accomplished by splitting the original (146-observation) training set into two equal subsets using the DUPLEX method.[13] Each half was used to recalculate the coefficients of the descriptors from the original model while using the remaining half as a prediction set. The two subset models produced in this fashion showed a slight difference in the values of the coefficients compared to the original model, but the overall degree of fit and the standard deviation of regression appeared to remain essentially the same. The correlation between the predicted and observed boiling point values for the two prediction sets is high (r = 0.986 and 0.989) with the rms prediction error values being very similar to the standard deviation of regression of the original model (12.0 and 10.9 °C). On the basis of these results, the model appeared to be quite robust and applicable to the prediction of boiling points of similar compounds.

Development of the Reduced Descriptor Set. To this point, three individual datasets (i.e., furans/THFs, thiophenes, and pyrans) and one combination dataset had been examined and modeled. Since one of the main objectives of this work involves the study of the connection between molecular structure and boiling point, it was of interest to review the models developed thus far to determine the degree of similarity between models for the different datasets and also to investigate how they differ.

A review of the models developed thus far indicated that minimal overlap existed. The four models contain a total of 28 unique descriptors. Only one descriptor (number of single bonds) was common to all four models, and a second descriptor (FNSA-3, a CPSA descriptor) was common to three of the four models. These observations caused some concern because it was hoped that more similarity between models would have existed. The source of the dissimilarity could be attributed to how descriptor analysis was performed for each dataset. In each case, a large number of molecular descriptors (approximately 100–130) were calculated, and descriptor analysis was then performed. The results of descriptor analysis were highly dependent on the number of compounds in the dataset and the types of structures. Slight differences probably also occurred in descriptor analysis methodology. These factors lead to differences in the final descriptor pools which were then used in regression analysis. The work described in this section involved the selection of a reduced descriptor pool, from which new models of greater similarity were developed for each dataset studied to this point.

The 28 unique descriptors from the four existing models were used, and in order to make use of leaps-and-bounds regression, the set of 28 was reduced to 24. This was necessary because not all the descriptors were available in all four datasets, and some descriptors were very dataset specific. Thus, a set of 24 descriptors was identified and used with the appropriate final training set for each previous model. The regression method embodied in the program LEAP was then used to calculate the new models.

Table IV. Details of Furan/THF Class-Specific Model Based on 209-Observation Training Set and Reduced Descriptor Set

R^2 = 0.964, s = 11.8 °C, N = 209

descriptor	regression coeff	SD coeff
1st order mol connectivity	67.43	4.525
valence corrected 3rd order mol connectivity	11.72	2.787
no. of single bonds	−24.54	1.051
greatest positive partial atomic charge	149.4	15.15
LUMO energy (HMO method)	20.43	2.686
PPSA-1	0.2532	6.059×10^{-2}
DPSA-3	2.557	0.4004
FNSA-3	−500.2	125.7
WNSA-1	1.023	0.1177
intercept	−67.13	

Table V. Details of Thiophene Class-Specific Model Based on 134-Observation Training Set and Reduced Descriptor Set

R^2 = 0.972, s = 8.3 °C, N = 134

descriptor	regression coeff	SD coeff
no. of single bonds	−10.63	0.8777
av distance sum mol connectivity	27.40	3.873
molecular ID/no. of atoms	368.8	24.07
molecular weight	0.4033	3.670×10^{-2}
dipole moment	7.296	0.8821
radius of gyration	36.51	2.381
PPSA-1	0.3396	0.2826
FNSA-3	−747.4	83.92
intercept	−833.1	

The best model from each analysis with a coefficient of multiple correlation closest to 0.985 was chosen for further consideration. This was done in order to maintain the performance level of the new models in comparison to the original models while also keeping the number of descriptors involved in the models to a minimum. The models selected were then examined for the significance of the individual descriptors and for excess collinearity. As a result, three of the models were manipulated by using the program IRA in order to correct the collinearity problems. The result of these steps was a new model for each dataset that favorably compared to the original models in terms of fit and standard deviation of regression values, while having the additional feature of increased similarity between dataset models. These models are shown in Tables IV–VII.

The four new models contain 20 unique molecular descriptors. There is an increase in the similarity between the models with two descriptors (number of single bonds and FNSA-3) common to all four models, while a third descriptor (PPSA-1, another CPSA descriptor) is common to three of the four models. In addition, another 10 descriptors are used in at least two of the four models. Thus, it is apparent that the subset of 20 molecular descriptors is sufficient to allow the development of good quality models which show an increased similarity. The list of the 20 descriptors in the reduced descriptor set is given in Table VIII. The descriptors are listed along with the class or type of molecular representation which was used to derive them.

Several aspects of this set of descriptors are appealing. A large fraction of the descriptors are derived from the topological molecular representation and thus are insensitive to changes in the geometry of molecules. The CPSA descriptors, while

Table VI. Details of the Combo-1 Combined Class Model Based on 236-Observation Training Set and Reduced Descriptor Set

$R^2 = 0.964$, $s = 11.3$ °C, $N = 236$

descriptor	regression coeff	SD coeff
no. of single bonds	−16.60	1.055
no. of atoms	8.286	1.119
valence corrected 1st order mol connectivity	45.68	2.249
sum of atomic IDs for heteroatoms	3.956	0.6541
dipole moment	6.986	0.8158
radius of gyration	12.98	3.086
PPSA-1	0.1846	4.018×10^{-2}
PPSA-3	2.138	0.4852
FNSA-3	−1077.0	63.48
WNSA-2	0.9203	0.1211
RPCG	75.70	14.62
intercept	−133.1	

Table VII. Details of Pyran Class-Specific Model Based on 143-Observation Training Set and Reduced Descriptor Set

$R^2 = 0.976$, $s = 10.4$ °C, $N = 143$

descriptor	regression coeff	SD coeff
no. of single bonds	−5.610	0.7001
valence corrected 1st order mol connectivity	46.80	2.487
molecular ID/no. of atoms	233.3	34.50
av distance sum mol connectivity	37.34	6.126
FNSA-3	−1367.0	51.51
WPSA-3	10.92	0.9074
RPCG	85.12	14.64
intercept	−609.0	

Table VIII. List of 20 Molecular Descriptors of Which the Reduced Descriptor Set Is Composed[a]

descriptor	descriptor classification
no. of atoms	topologic
no. of single bonds	topologic
molecular weight	topologic
1st order mol connectivity	topologic
valence-corrected 1st order mol connectivity	topologic
valence-corrected 3rd order path mol connectivity	topologic
av distance sum mol connectivity	topologic
molecular ID/no. of atoms	topologic
sum of atomic IDs for heteroatoms	topologic
dipole moment	electronic/geometric
greatest positive partial atomic charge	electronic
LUMO energy	electronic
radius of gyration	geometric
PPSA-1	CPSA
PPSA-3	CPSA
DPSA-3	CPSA
FNSA-3	CPSA
WPSA-3	CPSA
WNSA-2	CPSA
RPCG	CPSA (electronic)

[a] Each descriptor is classified as to the type of structural information it encodes.

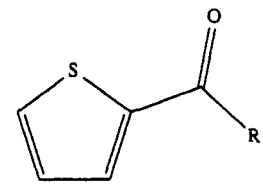
including surface area information, have shown limited sensitivity to differences in the geometry of molecules.[4] This means that the models developed are not highly dependent on the quality of the molecular modeling. In addition, several of the descriptors included in the reduced descriptor set are appealing from the point of view of physical interpretation. Descriptors such as molecular weight and the number of atoms are known to be correlated with the boiling points of sets of relatively similar compounds. Other descriptors such as dipole moment and the CPSA descriptors encode information necessary to account for the polar interactions between molecules. Thus, the reduced descriptor set not only allows for the development of good quality models, but the physical interpretation of the descriptors involved also agrees with our understanding of how different structural features affect observed boiling points.

The advantage of having selected the reduced descriptor set is that good models can be developed using substantially less time for descriptor calculation, analysis, and regression analysis. The results of regression analysis should be improved because methods such as leaps-and-bounds regression can be used in a single step to obtain good quality models quickly. Additionally, insight into the relationship between molecular structure and physical property can be obtained by examining the commonalities and differences between the new models.

Combined Dataset Modeling. Previously, different compound class datasets were combined for the purpose of developing more global boiling point prediction models. In this section, a valid boiling point prediction model is developed based on a dataset which includes observations from each of the three class-specific datasets examined to date (furans/THFs, thiophenes, and pyrans). Another goal of this work is to determine if the newly formed reduced molecular descriptor set would allow the rapid development of a valid model for this combined dataset. The model will be referred to as the *Combo-2* model.

A dataset containing 485 compounds was obtained by combining the final training sets for the Combo-1 model and pyran class-specific model. Due to software limitations within ADAPT, a maximum of 300 compounds is allowed in a single training set, so a random subset of 300 compounds was chosen from the original 485 observations. The remaining 185 compounds were set aside as an external prediction set.

Initial work indicated that eight compounds in the training set exhibited characteristics noted in previous work concerning the reduced pressure determination of the boiling point. These compounds were from the original furan/THF portion of the dataset which had not been as carefully screened as subsequent datasets, so the compounds were removed from the training set. Also, problems were encountered concerning the following thiophene moiety.



Only 26 of the original 485 compounds contained this moiety, and only nine of these 26 compounds were included in the initial training set. This resulted in poor performance of initial models for the prediction of the boiling point of the remaining 17 similar compounds in the prediction set. Since the goal of this work was to develop the most global model possible, the training set was altered to include more of these compounds. This allowed for greater influence of this moiety on descriptor selection. However, it also had the potential to make the model less stable. Because of the goal of this study, it was considered better to increase the representation of the thiophene moiety at the risk of potentially decreasing the stability of the resulting model. Thus, with all the adjustments

Table IX. Details of Final Combo-2 Boiling Point Prediction Model

$R^2 = 0.962$, $s = 11.8$ °C, $N = 299$

descriptor	regression coeff	SD coeff
no. of atoms	9.544	1.041
no. of single bonds	-11.47	0.6286
1st order mol connectivity	40.69	2.451
molecular ID/no. of atoms	89.55	21.22
sum of the atomic IDs for heteroatoms	1.970	0.5811
radius of gyration	15.66	3.054
dipole moment	5.863	0.7684
total positive charge	100.6	15.09
PPSA-3	1.327	0.3489
FNSA-3	-983.0	57.68
WNSA-2	0.9195	0.1086
intercept	-274.7	

noted, the populations of the training set and prediction set were 299 and 178, respectively, for a total of 477 observations.

The pairwise correlations for the 20 descriptors from the reduced descriptor set were checked, and there were no pairwise correlations above the usual critical value of 0.950. Leaps-and-bounds regression analysis was then done. The best model containing 11 molecular descriptors was chosen for further examination because of its size (number of descriptors) and degree of fit. The details of the model are given in Table IX, and the scatter plot of the fitted and observed boiling points is shown in Figure 4. The quality of the new Combo-2 model is similar to those obtained for other datasets. However, the development of the model was accomplished much more rapidly than before due to the use of the reduced descriptor pool and leaps-and-bounds regression.

The first steps of validation show the model to yield relatively low variance inflation factors (VIFs).[14] The mean VIF for the new model is 6.6 (high VIF = 14.3). These values are similar to the VIFs of models for other datasets. Comparisons of normal and jackknifed residuals show no large differences. These results suggest that the model is reasonably stable.

The ultimate test of any predictive model is its performance in external prediction. The Combo-2 model was applied to the prediction of the boiling points for the 178-observation prediction set which had been set aside. The scatter plot of the predicted and observed boiling points is shown in Figure 5. The total root mean squared error for the dataset was 13.5 °C, and the simple correlation coefficient for the predicted and observed values was 0.977.

An examination of the prediction error values suggested that the boiling points for two compounds were not as well predicted as those for bulk of the prediction set. The first of these is identical to a compound that was removed from the training set as an outlier. Normally, duplicates are removed in the early steps of a study, but this compound apparently existed in both the furan/THF dataset and also in the pyran dataset. Because of this, it would not have been detected as a duplicate. The second compound was the only furan/THF analogue containing a cyclopropyl ring. It is possible that the influence of the cyclopropyl ring is unique and, therefore, such a compound would not be similar enough to the training set to say that the Combo-2 model was appropriate for the prediction of this compound. Since both these compounds were questionable, they were set aside and the prediction statistics were recomputed. The correlation coefficient for the predicted and observed boiling points for the remaining 176 compounds was 0.981, and the total rms error was 12.2 °C. These results are in good agreement with the corresponding values for the training set. These results indicate that the

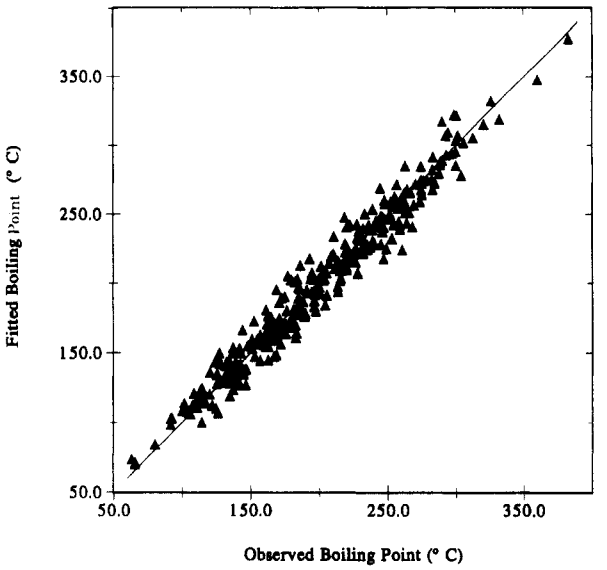


Figure 4. Scatter plot of the fitted and observed boiling points for the 299-observation Combo-2 combination dataset model.

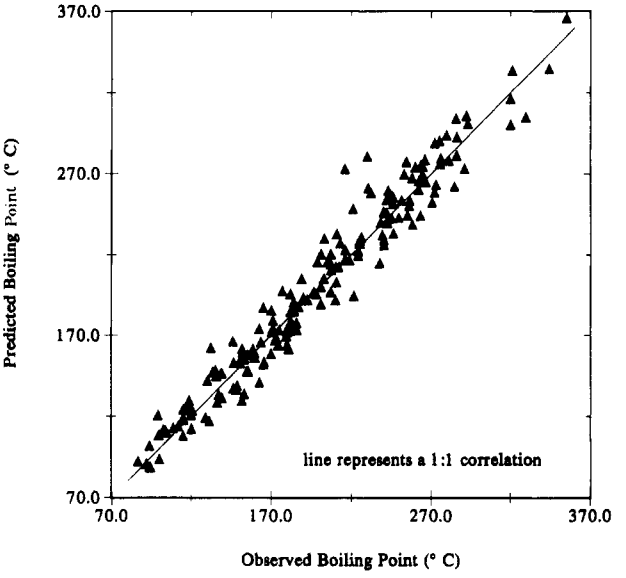


Figure 5. Scatter plot of the predicted and observed boiling points for the 178-observation Combo-2 external prediction dataset.

model is indeed applicable to predictions of similar compounds and represents a robust and valid predictive equation.

Finally, the Combo-2 model was used for the prediction of the boiling points for a set of furan/THF compounds which had not been previously included in modeling. The objective of this experiment was to determine if the Combo-2 model would perform better in prediction than did the Combo-1 model for the same dataset by virtue of the inclusion of the information from the pyran dataset. The results of the predictions of the boiling points of the 318 furan/THF compounds based on the Combo-2 model are shown in Figure 6. These results indicate that no significant changes have occurred for those compounds in the dataset which were determined at reduced pressure (those with large negative prediction errors). This further reinforces the idea of using these predictive models to detect error in a new dataset. In addition, the degree of positive prediction error has been reduced with the Combo-2 model, although the reduction is not as dramatic as was observed for the comparison of the predictions for the furan/THF class-specific model and the Combo-1 model.[1] This suggests that the addition of the pyrans to the combined data has improved the predictive value

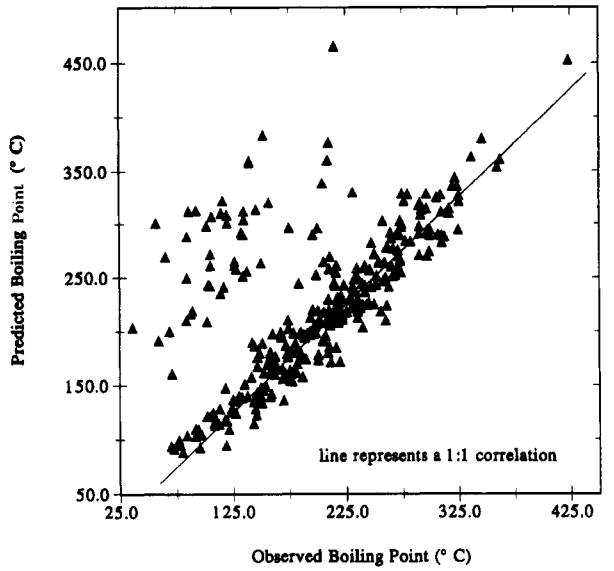


Figure 6. Scatter plot of the predicted and observed boiling points for the 318-observation furan/THF external dataset calculated by using the Combo-2 model.

of the model, but that a point is being reached where additional information will not provide a large increase in the quality of the model.

Pyrrole Dataset Study. The development of a boiling point prediction model for the pyrrole dataset was done similarly to that of the pyran dataset. The first step was to obtain a representative set of structures (the training set) upon which the model would be based. The second step of model development involved the selection and validation of the final model. Usually, several models are obtained for the final training set, and selection of the final model is based on several criteria. Among these are the size and accuracy of the model and the physical interpretation of the descriptors contained in the final model. Model validation also plays a role in model selection. Often, two models which produce very similar results will differ on the basis of their statistical quality. Thus, the final step is to assure that the model yields accurate results and is statistically sound.

Pyrrole Training Set Selection. The initial dataset received from the Beilstein Institute contained 395 compounds. This dataset was reduced to 377 observations by the removal of structures which could not be processed due to software limitations. The boiling points for these 377 pyrrole compounds were estimated by using the Combo-2 model. The boiling points for many of the pyrroles were well-predicted, indicating that the Combo-2 model incorporates information which is not specific to sulfur and oxygen heterocycles alone. However, many observations yielded large negative residuals. This type of result was expected since it is now well-established that there is a certain amount of error in the Beilstein database. Several observations also yielded large positive estimation error values. This too is expected because there are features of the structures involved which have not been encoded into the model. Using the iterative procedure involving descriptor analysis, regression analysis, and outlier detection described for past datasets, a final training set of pyrrole compounds was obtained which contained 278 observations.

Pyrrole Model Development. A model for the pyrrole compounds was then developed. The new set of hydrogen-bonding descriptors described above was considered in this portion of the work. A set of 25 descriptors remained from the combined set of 31 descriptors consisting of both the hydrogen-bond specific set and the reduced descriptor set after

Table X. Details of Pyrrole Class-Specific Boiling Point Prediction Model

$R^2 = 0.962$, $s = 12.3$ °C, $N = 278$

descriptor	regression coeff	SD coeff
molecular ID/no. of atoms	501.9	17.48
av distance sum mol connectivity	47.77	3.885
radius of gyration	44.01	1.641
dipole moment	13.44	1.189
DPSA-3	2.669	0.1419
RPCG	51.38	11.38
RSAA[a]	-0.4405	0.1078
intercept	-1171.0	

[a] Hydrogen-bonding specific descriptor.

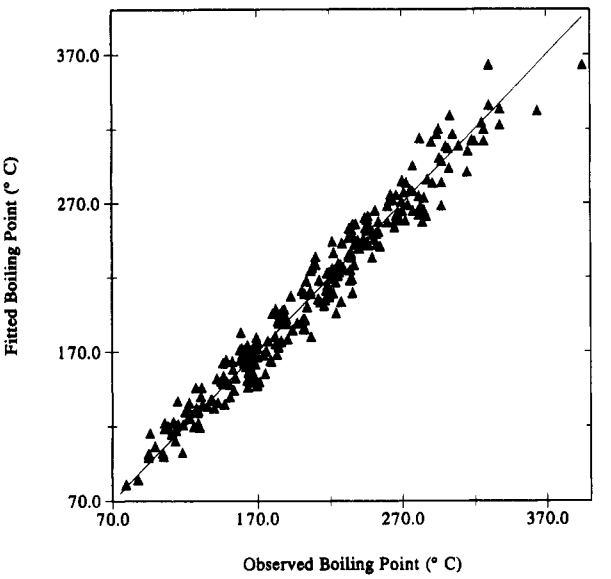


Figure 7. Scatter plot of the fitted and observed boiling points for the 278-observation pyrrole class-specific training set.

the usual descriptor analysis procedures had been applied. All 25 descriptors were then regressed along with the experimental boiling point data using leaps-and-bounds regression. The result was a 7-variable model which yielded a good fit (R^2 = 0.962) and standard deviation of regression (s = 12.3 °C). The details of the model are given in Table X. The scatter plot of the fitted vs observed boiling points for the 278 compounds is shown in Figure 7. The resulting boiling point estimates for the compounds shown in Figure 2 are given in Table III.

Examination of the error distribution for these fitted boiling points suggests that the model may not be as effective for compounds with boiling points at the extreme high end of the temperature scale. This may be the result of thermal instability of compounds with normal boiling points above 300 °C.

The types of descriptors involved in the model indicate a decrease of the importance of topological information and an increase of the influence of the geometric and CPSA descriptor types. This suggests that the polarity and shape of the pyrroles influence the observed boiling point more for this dataset than for the other class-specific datasets studied previously. Only one of the descriptors (RSAA) in the model is from the set of hydrogen-bonding specific parameters. This descriptor provides information concerning the average solvent-accessible surface area of hydrogen bonding acceptor groups in the molecule. It is encouraging that one of the descriptors derived for the purpose of encoding information concerning structural features that are responsible for a specific type of molecular interaction was found to be significant in the model. This

suggests that this form of descriptor encodes information different from that available from the CPSA descriptors. Since it has been found to be important for this particular dataset, it also suggests that this descriptor is encoding the desired information concerning hydrogen bonding. These results also seem to suggest that the original CPSA descriptors may already encode information concerning the hydrogen-bonding characteristics of the molecules. However, more work needs to be done in order to determine if there are better combinations of surface and charge information which can be used for this purpose.

The model was then validated. The variance inflation factors (VIFs) were calculated for each of the descriptors in the model. The mean VIF was found to be 2.7 with a high of 4.3 and a low of 1.8. These values are much lower than those obtained for models of past datasets, indicating that there may be less overlap of information between the descriptors in this model than has been observed for models for the other datasets. The examination of the jackknifed residuals for the dataset showed none of the compounds yielded jackknifed residuals which were much larger than the original fit residuals. This suggested that none of the compounds were unduly influencing the regression.

As a next step in the validation, 28 compounds (approximately 10%) were selected by using the DUPLEX method. These 28 compounds were set aside, and the coefficients of the original model were then recomputed on the basis of the 250-observation dataset. Comparisons of the original model and the reduced set model showed that the new coefficients were all within one standard deviation of the original coefficients. The fit and standard deviation of regression for the deletion set model were also in good agreement with that of the original model (R^2 = 0.964; s = 11.8 °C). Thus the model appeared stable at the 10% deletion level.

In order to examine the model more vigorously, the original dataset was divided in half, using the DUPLEX method. Each half was used alternately as both a training set and a prediction set. The resulting subset models still appeared to be in good agreement with the original model, but in a few instances the new coefficients differed from the original coefficients by slightly more than one standard deviation. The uncertainty of one coefficient for each subset model was thought to be large enough to be suspect, but both passed the partial-F test[15] at a confidence level of 95%. These results, taken together with the results noted above, suggest that the model is both stable and robust, and it should be quite effective for the prediction of normal boiling points of similar compounds.

Combined Dataset Model Development (Combo-3). The ultimate goal of this study and related past studies is to provide the means to estimate boiling points for very diverse sets of compounds within the Beilstein database. The ideal solution to this problem would be a single model which would perform equally well for any compound. Realistically, it is expected that a set of equations will be necessary in order to provide accurate estimated boiling points. However, it is not known just how diverse a dataset can be before it becomes too difficult or impractical to develop good models.

Past work involved the development of a model for combined sets of furans/THFs and thiophenes (Combo-1). Later, a diverse set of pyran compounds were added to this first combination set to produce a new combination model (Combo-2). Both these models proved to be stable and robust, and they yielded good predictive results. The next logical step involved combining the compounds from the final pyrrole dataset with the dataset used to obtain the Combo-2 model. The goals of this experiment were to determine if nitrogen heterocyclic compounds could be combined with the oxygen and sulfur heterocycles and to determine if such a model would be an improvement over the Combo-2 model. This new model will be referred to as the *Combo-3* model.

Table XI. Details of Model Developed for 752 Observations in Combined Furan, THF, Thiophene, Pyran, and Pyrrole Dataset (Combo-3)

R^2 = 0.954, s = 13.1 °C, N = 752

descriptor	regression coeff	SD coeff
no. of single bonds	−7.746	0.3653
valence corrected 1st order mol connectivity	27.07	2.680
valence corrected 3rd order mol connectivity	6.067	1.582
molecular ID/no. of atoms in mol	292.3	16.13
av distance sum mol connectivity	34.10	2.769
radius of gyration	30.33	2.376
max positive partial atomic charge	96.57	8.776
dipole moment	7.674	0.5342
PPSA-3	3.494	0.2332
FNSA-3	−816.0	38.40
WNSA-2	0.2816	0.0607
intercept	−763.8	

In order to construct the model, the training and prediction datasets used for the Combo-2 model were combined with the training set from the pyrrole study to give a final dataset of 753 compounds. The 20 descriptors of the reduced descriptor set and the 11 hydrogen-bonding descriptors were calculated for these compounds.

The multiple regression analysis routines available within the ADAPT system cannot be used to develop models for such a large dataset. Therefore, the Minitab statistical software package[16] was used. Minitab was used on both the Penn State University mainframe computer system and a MS-DOS compatible personal computer. The multiple linear regression analysis used was the method of best-subsets regression (BREG) as it is implemented in Minitab. The mainframe version of Minitab can examine a maximum of 20 descriptors in the BREG routine, while the PC version is limited to a set of 15 descriptors. For this reason, initial model development was accomplished by using the mainframe version, and subsequent model validation was done using the PC version.

Standard descriptor analysis methods were employed to examine the 31 available descriptors. In addition, some descriptors were set aside on a subjective basis in order to obtain a set of 20 descriptors which were thought to be most important. These 20 descriptors were submitted to regression analysis using BREG. The program reported the best five models of each size (1–20 descriptors). The standard deviation of regression declined as a function of the number of variables in the model. Based on these results, the 11-variable model was chosen for further examination.

Initial analysis of the 11-variable model for the 753 compounds showed that one of the 753 observations exhibited a great deal of influence on the regression by causing the uncertainty of many of the coefficients to be high. This single observation was removed, and the coefficients for the original 11-variable model were recomputed on the basis of the remaining 752 observations. The details of this final Combo-3 model are given in Table XI. A plot of the fitted vs observed boiling points for the 752 compounds is shown in Figure 8.

The stability of the Combo-3 model was then examined as described previously. The variance inflation factors were calculated, and the mean VIF for the model was 7.2 with a

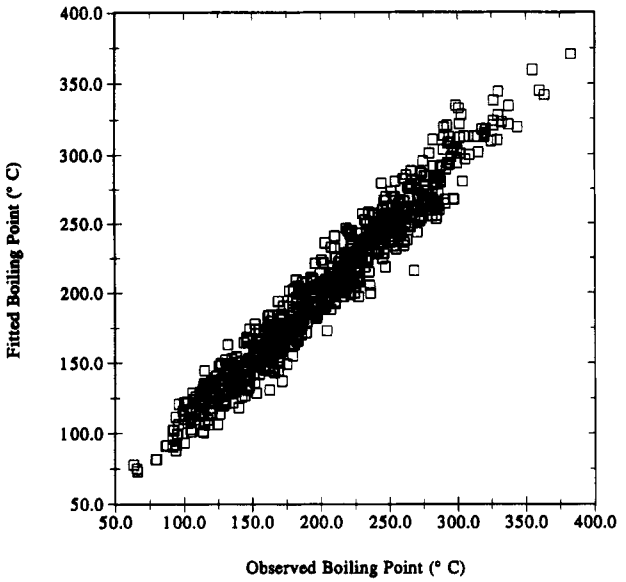


Figure 8. Scatter plot of the fitted and observed boiling points for the 752-observation Combo-3 dataset.

range of 29.3–2.3. These results are very similar to the results obtained for past studies. Although there is a certain amount of collinearity among the descriptors in the model, this has not appeared to affect the utility of past models.

Examination of the model through the use of subsetting indicated that the model is sensitive to the existence of the nitrogen heterocycles in the training set. For any given subset, the fit for the training set remained good, but the predicted boiling points for nitrogen heterocycles (from the pyrrole dataset) were typically poorer than the predicted boiling points for sulfur and oxygen heterocycles. This suggested that several problems exist. The structural features of the nitrogen heterocycles appear to differ sufficiently from those of the sulfur and oxygen heterocycles that the descriptors comprising the model cannot adequately encode enough information concerning them. Since the standard deviation of regression was observed not to decrease very much as additional descriptors are added to the model, simply using more descriptors would not solve the problem. Also, it becomes very difficult to examine other descriptors outside of the 31 examined here because of the size of the training set and the limitations of the statistical software. Thus, while it is possible to build a statistically reasonable model on this 752-member training set, the model may not perform optimally in prediction in its present form. Other problems associated with the management of a training set of this magnitude make model development itself a challenge.

CONCLUSIONS

The results of the studies described in the preceding sections should be viewed in light of the goals initially set forth. It has been demonstrated that it is possible to develop sound and reasonably accurate equations for the prediction of normal boiling points for a very broad set of heterocyclic organic compounds. The average error of the estimated boiling points is approximately 5%. These results are of about the same quality as many of the group contribution methods which can be used for the same purpose. The difference between this approach and that used in the group contribution methods is that the necessary structural information can be easily and quickly calculated directly from the structures of the compounds of interest. Since the structural information is already available in the Beilstein database, there is no need for manual structure entry, so this method is directly applicable to that database.

The quality of the predictions for new compounds, those which were part of truly external prediction sets, was demonstrated in several cases. Thus, the approach of developing models on large and very diverse sets of compounds has been shown to be reasonable. Future research will have to consider which datasets should be combined and how many of the combined class models are necessary to make an efficient working system. Also important is the question of how to decide objectively which model is appropriate for a given new observation. Predictions obtained from the class-specific models will be more accurate than those obtained for the more global combined class models.

The value of using these regression equations to detect error in an existing database has also been demonstrated. The most common error detected in the data received from the Beilstein Institute involved boiling points which had been determined at reduced pressure, while that pressure went unrecorded or was recorded incorrectly. This type of error was found to involve as many as 20% of the compounds in a given a subset of the database. The models developed can be used to detect this type of error fairly easily, allowing the recorded data to be reviewed and corrected if necessary. This would increase the quality of the database overall and would also help to prevent the same type of error from being incorporated in the future. Since many of the physical property values available in the Beilstein database are used as keys for searching the database, detecting and removing this type (and other types) of error will also improve the effectiveness of searches. An added advantage of using the molecular structure descriptors for the prediction models is that the descriptors themselves can be used as effective search keys. Since each of the descriptors are derived directly from structure, there is little error involved with these values. Also, many of the descriptors have a low rate of degeneracy, which make them quite useful for the purpose of searching databases. Thus, the molecular structure descriptors used for the predictive models have the additional advantage of being potential search keys.

Results obtained from modeling the Combo-3 dataset have highlighted the limitations of the approach employed thus far. It is clear that it is not possible to produce good quality predictive equations using datasets which are continually made larger and more diverse. The difficulty observed has two possible causes. The inability to examine a large enough set of descriptors in regression makes it very difficult to obtain an optimal model. Alternatively, the diversity of the dataset may make it impossible to encode all the molecular structure information required to produce an optimal model with the descriptors available.

The difficulty observed in modeling the nitrogen-containing heterocyclic compounds along with the other datasets indicates that it may be better to consider modeling such compounds separately. Ultimately, a small set of good quality models which, when taken together, cover a wide variety of structure types would be most effective for estimating boiling points of compounds in large databases. Such an approach would require some reliable method of selecting the best model for a given new compound. Approaches based on molecular similarity or cluster analysis may be employed to facilitate model selection.

Finally, it is possible to draw some conclusions about the relationship between structure and property from this work by examining the types of descriptors involved in the reduced descriptor set. The size and shape of the molecules involved

are important. The molecular connectivity descriptors encode such information along with information concerning the degree of branching in the molecule. It is known that the degree of branching in a molecule will affect the observed boiling point. Normal boiling points for a set of isomers will decrease with increasing branching. The first-order molecular connectivity index also becomes smaller as the structure becomes more branched. Thus, there is a positive correlation observed, and the sign of the descriptor in the models is also positive.

As has been noted previously, the polar interactions between molecules in the bulk phase also have a strong influence on the observed boiling points. The CPSA descriptors have been found to be important for encoding this type of information. Examination of the reduced descriptor set shows the CPSA descriptors, as a class, to be important in the models developed, with two of the CPSA descriptors found to be common to at least three of the four models which were developed from the reduced descriptor set. The hydrogen-bonding descriptor, which is found in the pyrrole class-specific model, suggests that is possible to encode information concerning specific types of polar interactions. However, additional work in the area of the hydrogen-bonding descriptors is still necessary.

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.

REFERENCES AND NOTES

(1) 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.

(2) Stuper, A. J.; Jurs, P. C. ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 99.

(3) Brugger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of Descriptors from Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 105.

(4) 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.

(5) Stanton, D. T.; Jurs, P. C. Computer-Assisted Study of the Relationship between Molecular Structure and Surface Tension of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109.

(6) Pimentel, G. C.; McClellan, A. L. *The Hydrogen Bond*; Freeman: San Francisco, 1960.

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

(8) Shulgin, A. T.; Kerlinger, H. O. The Carbonyl Double Bond: A New Hydrogen Bond Receptor. *Chem. Commun.* **1966**, *No. 9*, 249–250.

(9) Cairns, T.; Eglinton, G. Hydrogen Bonding in Phenols. Part II. Alkyl Substituted Bis(hydroxyphenyl)alkanes (Dinuclear Novolaks). *J. Chem. Soc.* **1965**, 5906–5912.

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

(11) Rousseeuw, P. J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871.

(12) Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection*; John Wiley and Sons: New York, 1987; p 226.

(13) Snee, R. D. Validation of Regression Models: Methods and Examples. *Technometrics* **1977**, *4*, 415.

(14) Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; John Wiley and Sons: New York, 1980.

(15) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1985; p 281.

(16) Minitab Statistical Software Package, Release 7; 1989; Minitab Inc., 3081 Enterprise Drive, State College, PA 16801.