Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure

Leanne M. Egolf, Matthew D. Wessel, and Peter C. Jurs*

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

Received January 28, 1994*

Numeric representations of molecular structure are used to predict the normal boiling points and critical temperatures for compounds drawn from the Design Institute for Physical Property Data (DIPPR) database. Multiple linear regression analysis and computational neural networks (i.e., using back-propagation and quasi-Newton training) are employed to develop models which can accurately predict the boiling points of 298 organic compounds. This approach is assessed by comparing its results against results obtained using the Joback group contribution approach. Finally, the same methodology is used to develop two separate critical temperature models, one based on the methods of corresponding states and the second based on structurally derived parameters alone.

INTRODUCTION

Over the years much research has been devoted to improving the prediction of physical and biological properties of organic compounds. The development of methodology for generating structure–property relationships has facilitated this area of research. One such approach is implemented in a software system called ADAPT.[1] The strength of this methodology lies in the use of what are termed *descriptors*. These descriptors, which are derived strictly from molecular structure, encode topological, geometrical, and electronic information. Ultimately, by accessing this information, researchers can both focus on key structural features and use the numeric representations of those features to design accurate predictive models.

Previous studies have demonstrated how this approach can be used to screen and model the normal boiling points of heterocyclic organic compounds drawn from the Beilstein database.[2,3] In this study we focus on an alternate database, the Design Institute for Physical Property Data (DIPPR) database.[4] This data compilation consists of compounds which have been deemed industrially important from the chemical engineer's perspective.

The primary goal of this study is to use the ADAPT techniques to develop a boiling point predictive equation which encompasses as much structural diversity as possible. This equation will be used not only to evaluate the quality of new experimental data but also to fill in data gaps which currently exist within the DIPPR boiling point tables. In addition, the utility of applying computational neural network techniques to upgrade boiling point estimations will be further assessed through this work. Neural network systems based on the back-propagation learning algorithm[5] as well as the BFGS (Broyden–Fletcher–Goldfarb–Shanno) quasi-Newton optimization method[6] will be examined.

Investigations will also be initiated to explore where our methodology stands in relation to other predictive tools. In response to a recent challenge issued at an assembly of industrial engineers, experiments have been designed to weigh the performance of our methods against the more widely accepted group contribution methods. Group contribution, or additivity, methods[7] assume that the structural pieces of a molecule act independently of one another. Consequently, the characteristics which a molecule exhibits (i.e., physical properties, biological activities) are equal to the sum of the constants, or contributions, assigned to each individual atom, bond, and/or group fragment.[7]

Normal boiling point is a property which has been thoroughly tested and specifically parametrized in many of the group additivity designs. Thus, it serves as an appropriate test property for our head-to-head comparison. Joback's group contribution approach[8] will be used as the reference system since it is considered by many to be the most well-established, accurate, and widely used of the group contribution methods.

Finally, relationships between boiling points and other thermophysical properties are considered. Boiling points have historically received a great deal of attention because they are closely related to a number of other physicochemical properties and can be used for their estimation. Enthalpies of vaporization,[9] liquid molar volumes,[10] and critical temperatures[11] are three thermodynamic properties which have been closely linked with boiling point processes. Relying on the methods of corresponding states,[12] these properties are frequently estimated using equations in which accurate boiling point values function as the principal input variable. The final objective of this study is to use our descriptors to determine if they can enhance this accepted mode of prediction. Critical temperatures will first be modeled using the experimental boiling point as the primary variable, supplemented by a limited number of calculated structural descriptors. This model, once developed, will be compared to a second critical temperature model derived strictly from structural information alone.

METHODOLOGY

Linear regression computations were performed on a Sun 4/110 workstation using the ADAPT software package.[1] The neural network computations were carried out on a DEC 3000 AXP Model 500 workstation.

The Data Set. A group of 298 diverse organic compounds comprise the data set. Although 38% of the compounds are saturated or unsaturated hydrocarbons (i.e., straight-chain, branched, and cyclic), the remainder contain halogen, alcohol, cyano, amino, ester, ether, carbonyl, or carboxylic acid functionalities. Figure 1 illustrates several representative structures.

* Abstract published in *Advance ACS Abstracts*, May 15, 1994.

0095-2338/94/1634-0947$04.50/0

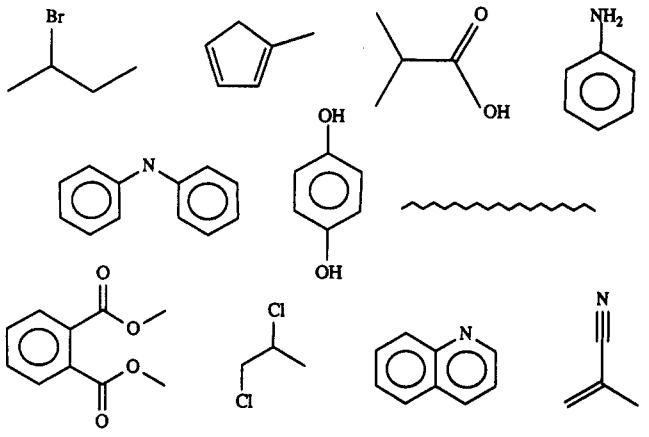

Figure 1. Representative structures from the DIPPR boiling point data set.

The experimental boiling points for each of the 298 compounds were extracted from the DIPPR database.[4] Of those, 165 compounds also have a reported experimental critical temperature. The boiling points range from 225 to 648 K, while the critical temperatures span from 365 to 782 K.

The experimental uncertainties which accompany these data have been established by members of the DIPPR database project staff.[4] Considering only those compounds in our data sets, calculations reveal that a 2.1% error is associated with the boiling points and a 1.3% error is associated with the critical temperatures. These uncertainties, while thought to be conservative, indicate that the target errors for our boiling point and critical temperature models are 9 and 8 K, respectively.

Molecular Modeling. The molecular structures were entered by sketching and stored in the form of connection tables. Because the three-dimensional features of a molecule are thought to have a direct impact on boiling point tendencies, molecular modeling routines were used to place the compounds in energy-minimized conformations.[13,14] It is these final structural representations which were repeatedly accessed during the descriptor generation phase of this work.

Descriptor Generation. After the molecular structures were stored along with their corresponding boiling points and critical temperatures, descriptor generation began. The majority of descriptors calculated for this study fell into three main categories: topological, geometrical, and electronic. Descriptors which unite information from two or more of the general categories were also utilized. Examples of such combination descriptors include the CPSA descriptors[15] and the hydrogen bonding descriptors, both discussed later in the text.

An effective model-building strategy used throughout our studies has been to focus on descriptors which have a sound, theoretical basis with respect to the property of interest. To this end, features which reflect branching, molecular flexibility, dipole–dipole interactions, and hydrogen bonding were encoded for the current work. The importance of these structural features was originally explored during the investigations of boiling points of pyridine compounds[3] and is briefly reviewed here. First, boiling point processes are significantly influenced by dispersion forces, the strength of which are determined by a molecule's shape.[16] Thus, two isomers can exhibit very different physical property values. The electrons in an elongated molecule such as normal pentane are much more easily polarized than are those in a more compact, symmetrical isomer such as neopentane. This tendency helps to explain *n*-pentane's higher normal boiling point and critical temperature. In comparing isomers again, it is conceivable that a molecule such as cyclopentanol would enter the vapor phase much more readily than 3-pentanol since it is locked in a rigid conformation and is less free to move and adapt to externally applied forces (i.e., thermal energy). Molecular flexibility descriptors were calculated to explore this concept.

In previous investigations of boiling points, residual trends encountered during the early stages of model development showed the need to design descriptors to characterize two additional intermolecular forces. The first of these, the dipole–dipole interaction, enables a liquid to resist a transition to the gaseous phase. In a polar substance, a molecule adjusts to its immediate environment by orienting the positive end of its dipole with the negative end of an adjacent molecule. This partial ordering of molecules causes a pure substance to persist as a liquid at temperatures higher than otherwise expected.[17] The realignments which facilitate dipole–dipole attrractions also encourage the development of an even stronger intermolecular interaction, hydrogen bonding. Liquids which contain a highly electronegative atom such as N or O as well as an H atom bonded to a highly electronegative atom exhibit substantially elevated boiling points.[17] This phenomenon is attributed, at least in part, to the hydrogen bonds which can from between two neighboring molecules or among small clusters of molecules. The resulting dimers or polymer-like structures stabilize the substances in the liquid state to a marked degree.

Descriptor Screening and Regression Analysis. Once a pool of descriptors is generated, subsets of potentially useful descriptors are next identified through a series of screening procedures. The first step is to eliminate all descriptors which contain ≥90% identical values since these descriptors provide little discrimination among the compounds under investigation. When only informationally rich descriptors remain, all possible pairs of descriptors are then examined in order to reduce information overlap. One descriptor is discarded from each pair exhibiting a pairwise correlation of ≥0.95. Physical interpretability and ease of calculation often direct this feature selection. Since some of the model-building routines can screen only a limited number of descriptors simultaneously, robust subsets of descriptors were finally assembled using both chemical intuition and vector space descriptor analysis,[18] a descriptor-ranking algorithm based on the Gram–Schmidt orthogonalization procedure.

Once a select group of descriptors is pinpointed, regression analysis can begin. Equations linking the structural features to normal boiling points (or critical temperatures) are developed which follow the general form

$$PP_j = b_0 + \sum_{i=1}^{n} b_i X_{ij}$$

where PP_j is the physical property of compound j, b_0 is the y-intercept of the regression model, b_i is the coefficient of the ith descriptor x_i, and n is the number of descriptors in the final model.

Leaps-and-bounds regression[19] and interactive regression analysis (IRA) are the linear regression techniques applied in these studies. Leaps-and-bounds is a highly automated routine which self-sufficiently identifies models based on the R^2 criterion. The IRA algorithm, by contrast, is designed to be used interactively. This routine allows the researcher to watch how a model reacts as new descriptors are added and the old ones removed. Capitalizing on this capability, IRA

Table 1. Boiling Point Model Developed from the ADAPT Descriptors Using Multiple Linear Regression Analysis[a]

descriptor	coeff		SD
fractional negatively charged SA – 3	−514.5	±	62.9
surface-weighted negatively charged SA – 2	1.212	±	0.112
surface-weighted negatively charged SA – 3	−9.353	±	0.871
number of oxygens	−10.04	±	1.44
Weiner number	−0.1102	±	0.0064
molecular ID	15.90	±	0.26
$[\lvert(q_{acc})(SA_{acc})\rvert + \lvert(q_{don})(SA_{don})\rvert]/SA_{tot}$	545.7	±	26.4
av charge on carbonyl or cyano carbons	−70.51	±	10.72
intercept	158.9		

[a] $R = 0.988$; $s = 11.85$ K; $N = 268$ compounds.

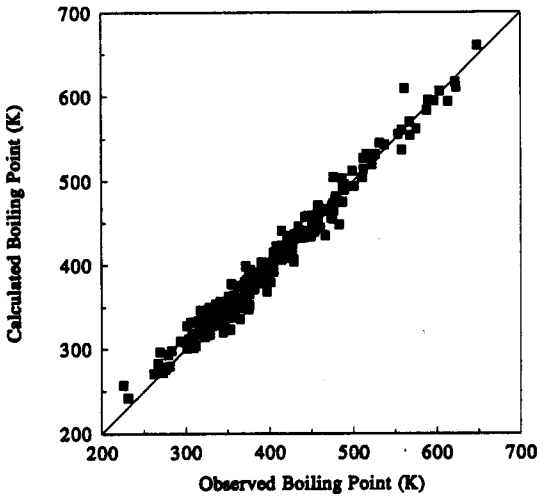

Figure 2. Calculated vs observed normal boiling points for 268 diverse organic compounds using the regression model.

can be used to refine descriptor subsets suggested by less internally critical programs like leaps-and-bounds.

Model development on the normal boiling points began by randomly splitting the compounds into a training set (268 compounds) and a prediction set (30 compounds). These subdivisions were subsequently used in both the linear regression and neural network work.

DISCUSSION

Linear Regression Analysis. The research conducted in past boiling point investigations greatly accelerated the model development process for this study. By capitalizing on an optimized pool of theory-based descriptors, an accurate predictive model was generated in a fairly straightforward and efficient manner. The regression model is summarized in Table 1, and a plot of calculated vs observed boiling points is shown in Figure 2. Recall that the experimental error associated with the boiling point data was 2.1% at the mean of the boiling point range. Thus, our model's correlation R of 0.988 and standard deviation of regression s of 11.85 K (2.7% error at the mean) are excellent, especially given the fact that a mere eight variables were used to obtain these results. Furthermore, all 268 data points could be retained in the development of the final model. This suggests two things: (1) the data were well-screened prior to their inclusion in the DIPPR database and (2) the ADAPT techniques can recognize the underlying structural trends in a diverse group of compounds and effectively reflect those trends through high-quality predictive equations.

Validation procedures were used to verify the strength of the proposed model. Internal validation tests began with a visual inspection of the calculated vs experimental plot, Figure 2, and the residual plot. Figure 2 illustrates the very tight fit of the data, with no discernable skew on either side of the ideal one-to-one correlation line. The residual plot reflected a random scatter of the calculation error, which is also ideal. Results of the outlier detection schemes were examined next. Although a few compounds were cited for exhibiting characteristics which slightly exceeded statistical norms for the data set, the model was not significantly improved upon the temporary removal of that data. Thus, the original model was deemed sound for the data set as a whole and was subsequently reinstated.

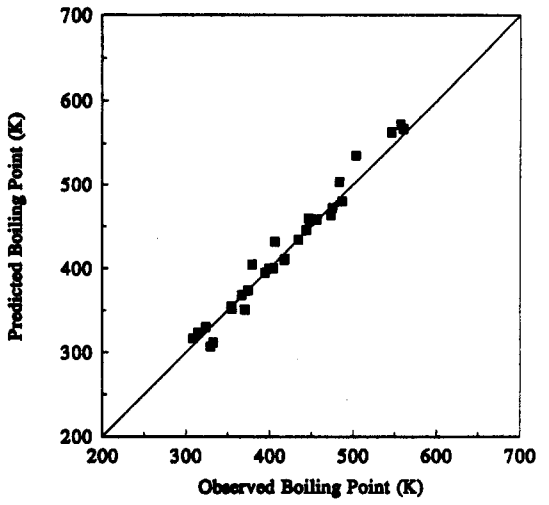

Figure 3. External prediction results obtained using the regression boiling point model.

External validation was used to assess the universal applicability of the established relationship. Figure 3 illustrates the model's performance on the external prediction set. The prediction set's R of 0.987 and RMS error of 13.23 K are virtually identical to the results obtained on the training set. This degree of reproducibility is exceptional and supports the model's eventual use as a tool to provide accurate estimations in the absence of experimental data.

The individual descriptors which define the DIPPR boiling point equation offer unique structural information. The first three descriptors (see Table 1) are known as CPSA descriptors.[15] These descriptors, which combine the solvent-accessible surface area of each atom and the partial charge of each atom in a variety of ways, were designed to focus on features which contribute to polar interactions. The CPSA descriptors selected for this model access negative charge, negatively charged surface area, and total molecular surface area information and present this information in three distinct combinations. The persistence of these specific structural features suggests that the accessibility of the heteroatoms plays a prominent role in determining a compound's propensity to enter the vapor phase.

The next three descriptors are all topological. The number of oxygens is simply a count of the number of oxygens in any given compound. The Weiner number[20] and molecular ID,[21] however, are less intuitive. These descriptors, which are derived using a graph theoretical perspective of molecular structure, denote branching indices which enable the researcher to distinguish among structural isomers. The Weiner number represents a compound as the sum of the entries in the lower triangle of the topological distance matrix. The molecular ID, on the other hand, is based on the summation of all unique paths in the molecule.

As introduced in the Descriptor Generation section of this paper, the final two descriptors were proposed in direct response

Table 2. Summary of Neural Network Performance Using Three Different Network Architectures and Back-Propagation Training

	2 hidden neurons		3 hidden neurons		4 hidden neurons	
trial	CV error	TR error	CV error	TR error	CV error	TR error
1	11.46	10.85	9.90	10.07	10.60	9.96
2	12.88	10.62	12.56	9.67	12.58	9.51
3	11.55	10.90	9.50	10.01	9.02	9.89
mean	11.96 K	10.79 K	10.65 K	9.92 K	10.73 K	9.79 K

to needs established in the pyridine boiling point work. There, the development of superior predictive models was facilitated by effectively encoding both hydrogen bonding and dipole–dipole interactions. Turning to the current study, a number of hydrogen bonding descriptors were found to offer significant independent information, as noted during the descriptor analysis procedures. Of the remaining pool of hydrogen bonding descriptors, though, it was curious to note that the one which proved most useful was the same descriptor selected in the pyridine boiling point study, its first application. A feature specific to dipole–dipole interactions likewise entered our structure–boiling point relationship. By introducing the average charge on the carbonyl and cyano carbons as a descriptor, the strength of the induced dipole and its overall effect on the boiling point process can be realized.

Neural Networks. Once eight collectively useful descriptors were identified through linear regression analyses, investigations were initiated to determine if the prediction accuracy could be bettered through the use of neural networks trained with back-propagation and quasi-Newton methods. The theory of these techniques has been discussed in depth in previous publications.[5,6] Consequently, only those details which are specific to the current work will be presented here.

The experimental design was as follows. The 268 compounds which comprised the regression training set were first randomly subdivided into a training set (241 compounds) which is used to develop the structure–property relationship and a cross-validation set (27 compounds) which is used to pinpoint the optimal set of weights for a given network system. Through monitoring the results of three different training set–cross-validation set pairs, our goal was to quantitatively assess the impact of using two, three, or four hidden layer neurons. The sole pieces of information submitted to each network were the eight structural descriptors. By refining the ways in which this information is processed, progressively more accurate boiling point estimations can be attained.

The performances of networks trained using back-propagation are summarized in Table 2. Before judging one system as being superior overall, it is important to recall that the success of any model is a function of both its accuracy and the number of parameters required to produce that accuracy. In Table 2 it is apparent that approximately 1 K of accuracy is gained when three hidden layer neurons are employed as opposed to two. However, the results obtained using three hidden neurons are virtually identical to those acquired when using four, despite the 10 additional adjustable parameters in the latter system. The best network is therefore the 8:3:1 network. In comparing the neural network results to those obtained through linear regression techniques, only a 1 K improvement is observed. This amount was determined to be unsubstantial, and therefore, this avenue of study was not pursued further.

The BFGS training method was then tested using a somewhat different approach. The BFGS technique uses a quasi-Newton optimization algorithm to train the neural network rather than the simple gradient descent optimization technique employed by back-propagation training.[6] The training period is shorter and the accuracy of the estimations is improved for the BFGS training as compared to back-propagation training.[22]

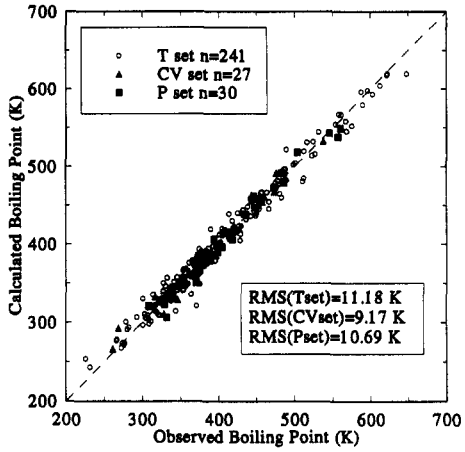

Figure 4. Calculated vs observed plot for the training, CV, and prediction sets using BFGS neural network methods.

The same 8:3:1 network architecture was used for the BFGS-trained neural networks. A new cross-validation set of 27 compounds was randomly selected from the original regression training set. An automated version of the BFGS method was then used to develop a highly accurate neural network model. The automated version conducted 500 individual network trainings (each having different random initial weights and biases), and the lowest cross-validation set error was reported for each session. Once the lowest error was found, the network model associated with that session was used to estimate the boiling points of the 30 compounds in the prediction set. The results of these efforts are shown in Figure 4. The RMS errors for the training set and cross-validation set are comparable to those obtained with regression and the back-propagation-trained neural network. The RMS error for the prediction set of 30 compounds not used for the training is substantially better than the regression value. No substantial outliers are seen, and the data cover the range from 225 to 648 K quite evenly.

Group Contribution Methods. Tests to explore the effectiveness of the linear regression methods versus the Joback group contribution methods were initiated next. In principle, our techniques should yield superior predictions since our approach views the atoms in a molecule as interactive pieces of a larger unit. First, our descriptors can encode information which is identical to that represented via the group contribution approaches, the atom, bond, and group counts. However, rather than being restricted to strictly topological information which can be inferred from the contributions selected, our techniques can focus on the geometrical and electronic aspects of the molecule as well. These latter features are intrinsic to many of the intermolecular forces (e.g., dipole–dipole interactions, hydrogen bonding) which, in turn, are fundamental to the boiling point process. Finally, where chemical theory suggests the study of structural features not parametrized through existing software, ADAPT's capabilities can be expanded to quantitate information on any potentially useful structural nuance.

The success of the Joback method was assessed through a parallel investigation of the boiling point data set. The boiling point group contributions derived by Joback were obtained by regressing the normal boiling points of a group of 438 diverse organic compounds against the numbers of each structural

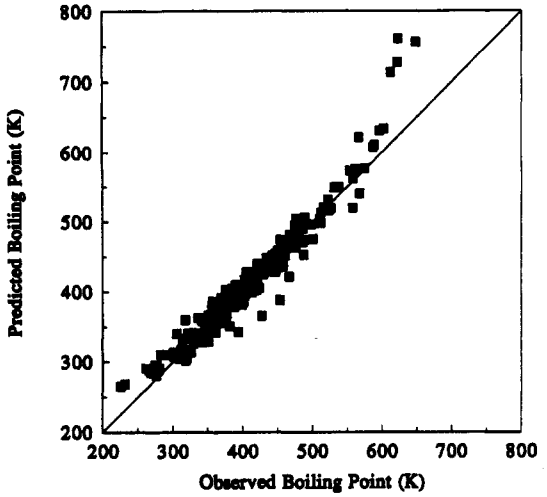

Figure 5. External prediction of the 268 training set compounds using the Joback group contribution approach.

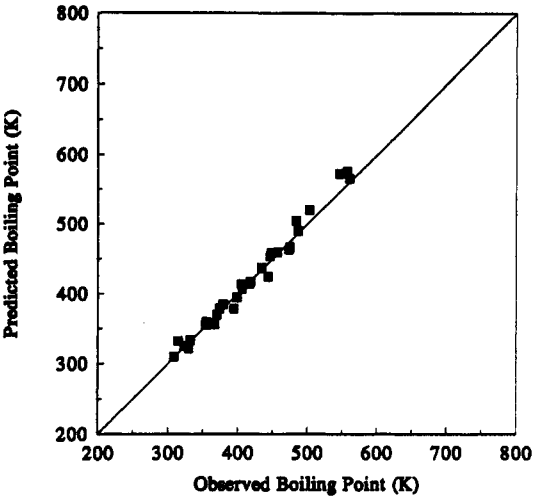

Figure 6. External prediction of the 30 prediction set compounds using the Joback group contribution approach.

group present.[12] The coefficients encompass information from hydrocarbons as well as compounds which contain one or more of the following functionalities: halogens, alcohols, ethers, esters, carbonyls, carboxylic acids, nitrates, nitriles, amines, imines, sulfides, thiophenes, thiols, isocyanates, phthalates, acrylates, anilines, furans, pyrroles, pyridines, and piperidines. To obtain the normal boiling point estimation for a given compound, the group increments are entered into the following equation.

$$T_b\,(\mathrm{K}) = 198 + \sum(\text{group increments})$$

Although boiling points were readily obtained for the majority of compounds, the technique proved to be somewhat ambiguous for one chemical class, the formates. After investigating several interpretations and matching our formate predictions with those presented in Joback's thesis (i.e., three of five were presented), we identified the scheme that appears to be the one Joback intended for use.[8]

Once the formates were predicted using the Joback format, the overall performance of this group contribution approach was evaluated. First, the Joback parameters were applied to the prediction of the training set compounds. The correlation between estimated and observed boiling points is 0.969 with an rMS error of 20.48 K, results which are clearly inferior to those acquired through our approach. In fact, the predictive accuracies for these compounds would be sacrificed by more than 8 K if efforts were restricted to this mode of estimation. An analysis of the correlation plot, Figure 5, reveals possible boundaries associated with the Joback approach. Although the majority of the data appears to be faily well predicted in the 275–600 K region, prediction skews are evident at each end of the correlation. Unfortunately, there was no single structural feature which could be pinpointed as the sole source of this limitation. Large diesters, esters, and carboxylic acids seem to be predicted with the least accuracy at the high end of the range, while propane and propylene are the compounds predicted with the least accuracy at the low end. If a single questionable moiety had been isolated, the problem most likely could have been rectified. Instead, the results may simply be a function of the approach itself, both in the types of structural features which have been parametrized and the contributions which have been permanently assigned to each group.

One way to investigate the latter possibility is to generate group contribution values (i.e., coefficients) for the specific data set at hand. Working with the 268-compound training set alone, we simply enter the 29 groups which are represented, provide a count of each relevant group, and regress the group count information against the target boiling points. The resulting (new) coefficients, which noticeably differ from Joback's, yield improved predictions. The final training set statistics have an R of 0.978 and a standard error of 16.65 K. Thus, the results can be upgraded by recalculating the coefficients to match the specific contributions of the compounds being accessed. An examination of the correlation plot, however, reveals that while the compounds near the center of the boiling point range have been pulled into the correlation, the undesirable skews at the end of the range have not been eliminated. As hypothesized earlier, the persistence of this problem suggests that the groups which Joback has parametrized cannot adequately account for the variance in the data at the extreme ends of the boiling point range.

Next, the Joback method (with Joback's coefficients) was used to predict the boiling points for the 30 prediction set compounds in an identical manner. Figure 6 shows the predicted vs observed boiling point values. Here, we see a much more linear distribution of the prediction results. The associated statistics, as one might surmise, are also much better, with an R of 0.991 and an RMS error of 10.80 K. Although these results seem altogether superior to those acquired on the training set compounds, one must remember that the procedures used to obtain these results were no different. These results, while initially misleading, do seem to confirm our suspicions that the Joback method performs best in the 275–600 K region identified earlier. All of the prediction set compounds have boiling points within this range. Finally, considering the overall results, we would hesitate to accept a Joback prediction outside the 275–600 K range. Obviously, the method then becomes very restrictive, especially when working with larger molecules or those which contain heteroatoms, both attributes which may force the normal boiling point above the 600 K limit.

Returning to our comparison of the ADAPT methods versus the Joback methods, only eight structural parameters were utilized for prediction purposes with ADAPT, whereas 29 were required with the Joback system. The need to access so few parameters implies that the ADAPT model is very stable since all eight descriptors probably play a vital role in almost every compound's prediction. Finally, unlike the Joback parameters, which are strictly counts of the number of each structural moiety present in the molecule, the individual ADAPT descriptors can lend insight into the chemistry

Table 3. Prediction Results from Two Boiling Point Models

	compound	exptl[a]	Model 1[b]	Joback[c]		compound	exptl[a]	Model 1[b]	Joback[c]
1.	acrylonitrile	350.50	359.07	366.80	78.	2-methyl-2-butanol	375.15	374.27	402.75
2.	acrylic aldehyde	325.84	329.57	313.38	79.	3-methyl-1-butanol	404.35	412.27	405.54
3.	ethylenecarboxylic acid	414.15	440.88	410.23	80.	3-methyl-2-butanol	384.65	381.74	405.10
4.	3-chloropropene	318.11	327.76	302.15	81.	1-pentanol	410.95	412.89	405.98
5.	propylene	225.43	257.27	264.72	82.	2-pentanol	392.15	384.64	405.54
6.	1,2-dichloropropene	361.25	376.13	342.46	83.	3-pentanol	388.45	378.48	405.54
7.	allyl alcohol	370.23	380.60	356.90	84.	neopentyl glycol	483.00	448.42	494.93
8.	methyl vinyl ether	278.65	293.50	287.14	85.	1,5-pentanediol	512.15	514.34	498.16
9	*n*-propionaldehyde	321.15	317.69	316.70	86.	*o*-dichlorobenzene	453.57	444.48	443.20
10.	ethyl formate	327.46	349.45	339.12	87.	bromobenzene	429.24	404.40	429.52
11.	methyl acetate	330.09	342.29	326.26	88.	chlorobenzene	404.87	407.01	400.79
12.	propionic acid	414.32	423.46	413.55	89.	benzene	353.24	362.86	358.38
13.	1-bromopropane	344.15	319.98	334.20	90.	phenol	454.99	454.12	439.00
14.	isopropyl chloride	308.85	310.64	305.03	91.	*p*-hydroquinone	558.15	536.98	519.62
15.	*n*-propyl chloride	319.67	318.72	305.47	92.	aniline	457.60	445.04	435.89
16.	allylamine	326.45	347.80	337.25	93.	2-methylpyridine	402.55	399.34	417.06
17.	propane	231.11	242.11	268.04	94.	adiponitrile	568.15	554.02	540.84
18.	methyl ethyl ether	280.50	279.56	290.46	95.	cyclohexene	356.12	348.15	360.06
19.	*n*-propanol	370.35	364.69	360.22	96.	cyclohexanone	428.90	407.06	428.72
20.	methylal	315.00	320.73	312.88	97.	mesityl oxide	402.95	394.86	394.59
21.	1,2-propylene glycol	460.75	444.07	451.96	98.	cyclohexane	353.87	345.48	360.90
22.	isopropylamine	305.55	331.84	340.13	99.	2,3-dimethyl-1-butene	328.76	333.70	332.80
23.	*n*-propylamine	321.65	337.22	340.57	100.	2,3-dimethyl-2-butene	346.35	328.75	340.60
24.	trimethylamine	276.02	276.12	280.48	101.	2-ethyl-1-butene	337.82	335.18	333.24
25.	1,3-butadiene	268.74	296.44	284.28	102.	1-hexene	336.63	340.64	333.36
26.	divinyl ether	301.45	327.90	306.70	103.	*cis*-2-hexene	342.03	333.97	340.84
27.	methacrolein	341.15	341.37	336.14	104.	*trans*-2-hexene	341.02	333.48	340.84
28.	*tert*-crotonic acid	458.15	453.68	440.59	105.	methylcyclopentane	344.96	343.63	351.96
29.	methyl acrylate	353.35	344.78	345.82	106.	2-methyl-1-pentene	335.25	336.78	333.24
30.	vinyl acetate	345.65	355.01	345.82	107.	4-methyl-1-pentene	327.01	336.72	332.92
31.	*n*-butyronitrile	390.75	379.17	393.00	108.	butyl vinyl ether	366.97	373.80	355.78
32.	1-butene	266.90	283.11	287.60	109.	cyclohexanol	434.00	446.15	448.41
33.	*cis*-2-butene	276.87	278.13	295.08	110.	2-hexanone	400.85	380.24	390.55
34.	*trans*-2-butene	274.03	278.01	295.08	111.	methyl isobutyl ketone	389.65	403.28	410.20
35.	isobutene	266.25	281.56	287.48	112.	ethyl *n*-butyrate	394.65	399.11	394.90
36.	*n*-butyraldehyde	347.95	346.18	339.58	113.	2-ethyl butyric acid	466.95	464.02	481.75
37.	isobutyraldehyde	337.25	332.69	339.14	114.	*n*-hexanoic acid	478.85	481.83	482.19
38.	methyl ethyl ketone	352.79	323.25	344.79	115.	isobutyl acetate	389.80	395.79	394.46
39.	tetrahydrofuran	338.00	332.67	337.82	116.	cyclohexylamine	407.65	422.06	428.76
40.	*n*-butyric acid	436.42	440.88	436.43	117.	2,2-dimethylbutane	322.88	327.42	333.45
41.	ethyl acetate	350.21	344.47	349.14	118.	2,3-dimethylbutane	331.13	328.30	335.80
42.	isobutyric acid	427.85	428.62	435.99	119.	*n*-hexane	341.88	331.42	336.68
43.	methyl propionate	352.60	337.43	349.14	120.	2-methylpentane	333.41	329.93	336.24
44.	*n*-propyl formate	353.97	377.50	362.00	121.	3-methylpentane	336.42	330.07	336.24
45.	1-bromobutane	374.75	347.90	357.08	122.	diisopropyl ether	341.45	351.42	358.22
46.	1-chlorobutane	351.58	346.57	328.35	123.	di-*n*-propyl ether	362.79	360.56	359.10
47.	*n*-butane	272.65	271.94	290.92	124.	acetal	376.75	379.34	381.08
48.	isobutane	261.43	270.68	290.48	125.	diisopropylamine	357.05	364.23	385.97
49.	*n*-butanol	390.81	387.65	383.10	126.	di-*n*-propylamine	382.00	372.98	386.85
50.	*sec*-butyl alcohol	372.70	359.15	382.66	127.	benzaldehyde	451.90	443.62	434.90
51.	*tert*-butyl alcohol	355.57	362.72	379.87	128.	benzoic acid	522.40	519.52	531.75
52.	diethyl ether	308.58	301.56	313.34	129.	toluene	383.78	386.01	386.24
53.	isobutanol	380.81	371.15	382.66	130.	benzyl alcohol	477.85	478.25	478.42
54.	1,3-butanediol	480.15	475.30	474.84	131.	*m*-cresol	475.43	473.18	466.86
55.	*n*-butylamine	350.55	362.58	363.45	132.	*o*-cresol	464.15	466.76	466.86
56.	diethylamine	328.60	317.19	341.09	133.	*m*-toluidine	476.55	464.78	463.75
57.	isobutylamine	340.88	357.14	363.01	134.	*o*-toluidine	473.55	456.31	463.75
58.	pyridine	388.41	382.89	389.20	135.	*n*-butyl acrylate	421.00	412.39	414.46
59.	cyclopentene	317.38	318.66	332.91	136.	isobutyl acrylate	405.15	407.87	414.02
60.	isoprene	307.21	318.32	307.04	137.	ethylcyclopentane	376.62	373.29	374.84
61.	acetylacetone	413.55	416.56	421.54	138.	2,3-dimethylpentane	362.93	357.53	358.68
62.	allyl acetate	377.15	395.03	368.70	139.	*n*-heptane	371.58	359.95	359.56
63.	ethyl acrylate	372.65	362.14	368.70	140.	2-methylhexane	363.20	358.76	359.12
64.	cyclopentane	322.40	314.47	333.75	141.	3-methylhexane	365.00	359.04	359.12
65.	1-pentene	303.11	311.75	310.48	142.	styrene	418.31	418.49	405.80
66.	diethyl ketone	375.14	350.39	367.67	143.	acetophenone	475.15	455.32	462.99
67.	2-pentanone	375.46	352.45	367.67	144.	ethylbenzene	409.35	411.28	409.12
68.	valeraldehyde	376.15	374.18	362.46	145.	*m*-xylene	412.27	409.01	414.10
69.	ethyl propionate	372.25	360.46	372.02	146.	*p*-xylene	411.51	408.78	414.10
70.	isobutyl formate	371.22	399.00	384.44	147.	2,6-xylenol	474.22	470.22	494.72
71.	*n*-propyl acetate	374.65	372.31	372.02	148.	*N*,*N*-dimethylaniline	466.69	435.55	421.56
72.	valeric acid	458.65	461.27	459.31	149.	*cis*-1,2-dimethylcyclohexane	402.94	403.19	397.32
73.	1-chloropentane	381.54	374.37	351.23	150.	*trans*-1,2-dimethylcyclohexane	396.58	401.81	397.32
74.	isopentane	300.99	300.44	313.36	151.	*cis*-1,3-dimethylcyclohexane	393.24	401.83	397.32
75.	neopentane	282.65	297.92	310.57	152.	*trans*-1,3-dimethylcyclohexane	397.61	403.10	397.32
76.	*n*-pentane	309.22	301.85	313.80	153.	*cis*-1,4-dimethylcyclohexane	397.47	402.72	397.32
77.	2-methyl-1-butanol	401.85	396.33	405.54	154.	*trans*-1,4-dimethylcyclohexane	392.51	401.70	397.32

Table 3 (Continued)

	compound	exptl[a]	Model 1[b]	Joback[c]		compound	exptl[a]	Model 1[b]	Joback[c]
155.	ethylcyclohexane	404.95	403.33	401.99	227.	*tert*-butylamine	317.55	346.18	360.22
156.	*n*-propylcyclopentane	404.11	401.85	397.72	228.	valeronitrile	414.45	406.20	415.88
157.	2,4,4-trimethyl-1-pentene	374.59	386.28	375.77	229.	2-methyl-1-butene	304.30	308.42	310.36
158.	2,4,4-trimethyl-2-pentene	378.06	381.84	383.25	230.	2-methyl-2-butene	311.71	303.76	317.84
159.	isobutyl isobutyrate	420.65	432.67	439.78	231.	3-methyl-1-butene	293.21	308.83	310.04
160.	2,3-dimethylhexane	388.76	385.67	381.56	232.	1,5-dichloropentane	453.15	444.53	388.66
161.	2-methyl-3-ethylpentane	388.80	386.19	381.56	233.	methyl isopropyl ketone	367.55	349.29	367.23
162.	*n*-octane	398.83	388.09	382.44	234.	isopropyl acetate	361.65	361.19	371.58
163.	2,2,3-trimethylpentane	383.00	383.23	378.77	235.	2-methylbutyric acid	450.15	448.61	458.87
164.	2,2,4-trimethylpentane	372.39	382.94	378.77	236.	methyl *n*-butyrate	375.90	365.62	372.02
165.	2,3,3-trimethylpentane	387.92	383.29	378.77	237.	piperidine	379.55	377.34	386.57
166.	di-*n*-butyl ether	413.44	415.96	404.86	238.	2,2-dimethyl-1-propanol	386.25	379.61	402.75
167.	2-ethyl-1-hexanol	457.75	471.14	474.18	239.	ethyl propyl ether	337.01	329.86	336.22
168.	quinoline	510.75	504.05	504.68	240.	methyl *sec*-butyl ether	332.15	331.54	335.78
169.	α-methylstyrene	438.65	438.95	428.56	241.	methyl *tert*-butyl ether	328.35	326.59	332.99
170.	ethyl benzoate	486.55	502.82	490.22	242.	methyl isobutyl ether	331.70	327.90	335.78
171.	cumene	425.56	434.85	431.56	243.	*n*-pentylamine	377.65	389.08	386.33
172.	*o*-ethyltoluene	438.33	435.39	436.98	244.	3-methylpyridine	417.29	408.67	417.06
173.	*p*-ethyltoluene	435.16	433.88	436.98	245.	1,3-cyclohexadiene	353.49	351.27	359.22
174.	mesitylene	437.89	432.26	441.96	246.	methylcyclopentadiene	345.93	348.46	359.93
175.	*n*-propylbenzene	432.39	436.27	432.00	247.	2,3-dimethyl-3-butadiene	341.93	340.93	329.80
176.	1,2,3-trimethylbenzene	449.27	433.74	441.96	248.	1,5-hexadiene	332.61	348.36	330.04
177.	1,2,4-trimethylbenzene	442.53	432.91	441.96	249.	hexanenitrile	436.75	436.03	438.76
178.	isophorone	488.35	491.46	497.07	250.	3,3-dimethyl-1-butene	314.40	333.86	330.13
179.	*n*-propylcyclohexane	429.90	431.52	424.87	251.	2-methyl-2-pentene	340.45	331.22	340.72
180.	3,3-diethylpentane	419.34	413.68	402.09	252.	3-methyl-1-pentene	327.33	337.12	332.92
181.	*n*-nonane	423.97	415.19	405.32	253.	ethyl isopropyl ketone	386.55	382.01	390.11
182.	2,2,3,3-tetramethylpentane	413.44	408.34	398.86	254.	1-hexanal	401.45	403.09	385.34
183.	dimethyl terephthalate	561.15	609.41	576.30	255.	3-hexanone	396.65	368.71	390.55
184.	1,2,3,4-tetrahydronaphthalene	480.77	478.87	475.54	256.	*sec*-butyl acetate	385.15	382.40	394.46
185.	*n*-butylbenzene	456.46	459.46	454.88	257.	*tert*-butyl acetate	369.15	374.78	391.67
186.	*tert*-butylbenzene	442.30	457.14	451.65	258.	ethyl isobutyrate	383.00	391.59	394.46
187.	*p*-cymene	450.28	457.10	459.42	259.	*n*-propyl propionate	395.65	387.79	394.90
188.	*m*-diethylbenzene	454.29	458.98	459.86	260.	*n*-butyl ethyl ether	365.35	359.22	359.10
189.	*o*-diethylbenzene	456.61	460.83	459.86	261.	2-ethyl-1-butanol	419.65	419.63	428.42
190.	isobutylbenzene	445.94	458.74	454.44	262.	1-hexanol	430.15	437.29	428.86
191.	*n*-butylcyclohexane	454.13	458.37	447.75	263.	2-hexanol	413.04	408.75	428.42
192.	*n*-decane	447.30	441.12	428.20	264.	2-methyl-1-pentanol	421.15	420.49	428.42
193.	2-ethylhexyl acrylate	489.15	496.47	505.54	265.	4-methyl-2-pentanol	404.85	392.42	427.98
194.	diphenyl ether	531.46	545.06	549.74	266.	methyl *tert*-pentyl ether	359.45	348.78	355.87
195.	diphenylamine	575.15	561.75	577.49	267.	1,6-hexanediol	516.15	531.95	521.04
196.	diethyl phthalate	567.15	569.68	622.06	268.	*n*-hexylamine	404.65	415.07	409.21
197.	*m*-diisopropylbenzene	476.33	504.58	504.74	269.	acetone[d]	329.44	307.16	321.91
198.	bicyclohexyl	512.19	527.36	513.06	270.	2-propanol[d]	355.41	351.83	359.78
199.	1-dodecene	486.50	490.85	470.64	271.	ethyl vinyl ether[d]	308.70	316.93	310.02
200.	*n*-dodecane	489.47	489.23	473.96	272.	cyclopentadiene[d]	314.65	323.00	332.07
201.	di-*n*-hexyl ether	498.85	511.97	496.38	273.	*n*-butyl formate[d]	379.25	404.58	384.88
202.	diphenylmethane	537.42	542.89	550.20	274.	isovaleric acid[d]	448.25	455.95	458.87
203.	benzyl benzoate	596.65	594.70	631.30	275.	*n*-butyl acetate[d]	399.15	399.71	394.90
204.	1,2-diphenylethane	553.65	555.69	573.08	276.	hexamethylene imine[d]	404.85	400.46	413.72
205.	1-tetradecene	524.25	529.88	516.40	277.	*p*-cresol[d]	475.13	472.02	466.86
206.	*n*-tetradecane	526.73	531.91	519.72	278.	*p*-toluidine[d]	473.40	463.33	463.75
207.	dibutyl phthalate	613.15	593.94	713.58	279.	1-heptene[d]	366.79	368.54	356.24
208.	1-hexadecene	558.02	560.72	562.16	280.	methylcyclohexane[d]	374.08	374.05	379.11
209.	dibutyl sebacate	622.15	616.82	727.68	281.	2,2,3-trimethylbutane[d]	354.03	354.60	355.89
210.	1-octadecene	587.97	583.68	607.92	282.	*o*-xylene[d]	417.58	410.09	414.10
211.	stearic acid	648.35	660.16	756.75	283.	1-octene[d]	394.44	394.87	379.12
212.	*n*-octadecane	589.86	595.25	611.24	284.	benzyl acetate[d]	486.65	479.62	490.22
213.	*n*-nonadecane	603.05	606.28	634.12	285.	*m*-ethyltoluene[d]	434.48	434.21	436.98
214.	*n*-butyl stearate	623.15	610.80	760.98	286.	dimethyl phthalate[d]	556.85	571.98	576.30
215.	methyl isopropenyl ketone	371.15	357.25	364.23	287.	*sec*-butylbenzene[d]	446.48	459.47	454.44
216.	1,1-dichloropropane	361.25	361.02	342.46	288.	*p*-diethylbenzene[d]	456.94	458.32	459.86
217.	1,3-dichloropropane	393.55	392.89	342.90	289.	1-decene[d]	443.75	445.50	424.88
218.	1,3-propylene glycol	487.55	475.05	452.40	290.	1-decanol[d]	503.35	535.11	520.38
219.	isobutyronitrile	376.76	377.60	392.56	291.	*p*-diisopropylbenzene[d]	483.65	503.30	504.74
220.	1,4-dichlorobutane	427.05	419.43	365.78	292.	1,1-diphenylethane[d]	545.78	562.60	572.64
221.	2-bromobutane	364.37	336.10	356.64	293.	*n*-hexadecane[d]	560.01	566.75	565.48
222.	*sec*-butyl chloride	341.25	334.56	327.91	294.	propionitrile[d]	370.50	351.02	370.12
223.	methyl isopropyl ether	323.75	303.00	312.90	295.	2-bromopropane[d]	332.56	311.86	333.76
224.	1,4-butanediol	501.15	493.86	475.28	296.	*tert*-butyl chloride[d]	323.75	330.37	325.12
225.	2,3-butanediol	453.85	439.65	474.40	297.	4-methylpyridine[d]	418.50	411.25	417.06
226.	*sec*-butylamine	336.15	353.71	363.01	298.	*n*-pentyl formate[d]	406.60	431.62	407.76

[a] DIPPR experimental boiling points (see ref 4). [b] Values obtained using the regression model (see Table 1). [c] Values obtained using the Joback approach. [d] Compound was a member of the external prediction set.

Table 4. Critical Temperature Models Developed through ADAPT

descriptor	coeff		SD
(a) Critical Temperature Model Developed by Capitalizing on Boiling Point Data[a]			
experimental normal boiling point	1.213	±	0.016
number of rings	34.77	±	1.61
valence-corrected sixth-order path mol conn	−54.45	±	5.40
intercept	97.02		
(b) Critical Temperature Model Derived from Structural Features Alone[b]			
relative negatively charged SA	2.080	±	0.170
number of oxygens	−13.64	±	2.02
number of aromatic bonds	7.370	±	0.561
atom-corrected κ-1 topological shape index	16.61	±	0.60
molecular ID/no. of atoms	508.7	±	15.0
path cluster-4 molecular connectivity	17.71	±	1.90
$[\lvert(q_{acc})(SA_{acc})\rvert + \lvert(q_{don})(SA_{don})\rvert]/SA_{tot}$	792.8	±	34.5
av charge on positively charged carbons	149.8	±	18.4
intercept	−514.0		

[a] $R = 0.994$; $s = 8.48$ K; $N = 147$ compounds. [b] $R = 0.989$; $s = 11.88$ K; $N = 147$ compounds.

involved in boiling point processes, as illustrated earlier. Boiling point estimations for the individual compounds are presented in Table 3. The values in the column titled Model 1 are the fitted boiling points for the 268 compounds of the training set and the predicted values for the 30 compounds of the prediction set. For comparison purposes, the values in the column titled Joback are the boiling points obtained using the Joback group contribution approach with Joback's coefficients.

Critical Temperature Investigations. The prediction of critical temperatures formed the next step in this investigation. Because the supply of experimental critical temperature data is limited, a new data set had to be defined for this part of the study. Of the 298 compounds which comprised the original data set, 165 had reported experimental critical temperatures. Those compounds which had experimental data were thus retained with their original training or prediction set status, and all others were simply dropped from further consideration. The final data breakdown for the critical temperature investigation was a 147-compound training set and an 18-compound prediction set.

Critical temperatures were initially modeled through the methods of corresponding states. Using the training set compounds' experimental boiling points as the primary variable in linear regression routines, a critical temperature model was developed. Only two additional pieces of structural information, the number of rings and the valence-corrected sixth-order path molecular connectivity,[23] were employed to refine the prediction accuracy. The final model is presented in Table 4a, and the fit plot is shown in Figure 7. Given that the target, or experimental, error is 8 K (i.e., 1.3%), our multiple correlation R of 0.994 and standard error s of 8.48 K are excellent. Note that none of the original 147 compounds had to be removed in order to generate these results. This again verifies the strength of the DIPPR data as well as our estimation techniques. The universality of the model is exemplified through the results, an R of 0.992 and an RMS error of 8.75 K, obtained on the external prediction set of 18 compounds. Finally, since the two supplemental descriptors are both topological in nature, the researcher is freed from having to access geometrical information available only through molecular modeling routines.

A logical question to pose at this time is, "Could estimated normal boiling points be successfully utilized in lieu of experimental values?" First, the experimental boiling points are correlated at 0.973 with the experimental critical temperatures. By contrast, the boiling point estimates from ADAPT correlated at 0.950 with the experimental critical temperatures. While the estimated values show a strong relationship, any critical temperature estimations would naturally be compromised to some extent. To limit any problem, we recommend estimated boiling points only be used where accurate experimental values are unavailable.

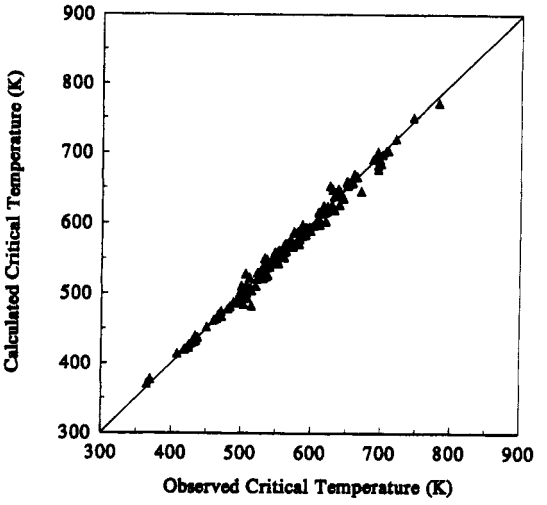

Figure 7. Correlation obtained with the critical temperature model which uses the experimental boiling point as a descriptor.

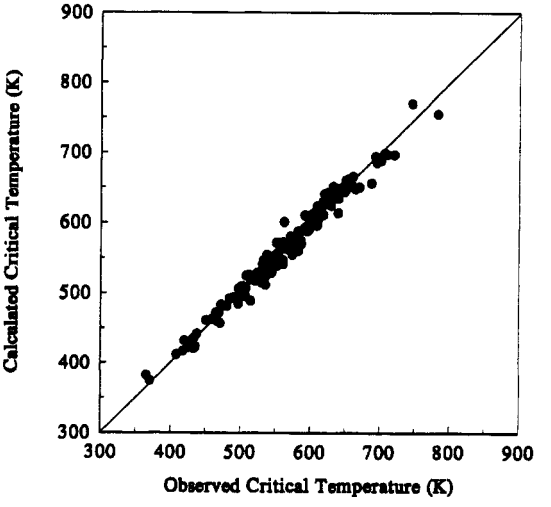

Figure 8. Correlation obtained with the critical temperature model which accesses structural information alone.

In some situations, it would be advantageous to be able to predict a compound's critical temperature solely from its molecular structure. One example is when a synthetic chemist proposes a new compound which he hopes will exhibit certain desirable characteristics. In the past, such a compound would have had to be synthesized and be physically in hand in order to ascertain any property information. Recent advances in property estimation techniques, though, enable current researchers to bypass the costly and highly inefficient synthesis stage and extract equivalent information through computational methods alone.

Although many approaches such as Joback's require boiling point values to be entered as inputs in their critical temperature estimation equations,[9] ADAPT is not hindered in this respect. As with the boiling point work, a high-quality model was developed from eight descriptors derived directly from the structure itself. The correlation, graphically depicted in Figure 8, is very tight. The model is shown in Table 4b. In comparing our boiling point and critical temperature models, it was

Table 5. Prediction Results from Two ADAPT Critical Temperature Models

	compound	exptl[a]	T_c (bp)[b]	T_c (struct)[c]		compound	exptl[a]	T_c (bp)[b]	T_c (struct)[c]
1.	3-chloropropane	514.50	482.87	488.93	78.	toluene	591.79	593.81	610.88
2.	propylene	364.76	370.45	382.36	79.	*m*-cresol	705.85	702.92	699.57
3.	allyl alcohol	545.05	546.09	529.21	80.	*o*-cresol	697.55	690.41	693.03
4.	ethyl formate	508.40	494.21	524.88	81.	*m*-toluidine	709.15	703.54	697.41
5.	methyl acetate	506.80	497.40	506.49	82.	*o*-toluidine	694.15	701.42	691.68
6.	propionic acid	604.00	599.57	613.84	83.	ethylcyclopentane	569.52	577.50	569.77
7.	*n*-propyl chloride	503.15	484.76	509.83	84.	2,3-dimethylpentane	537.35	537.24	535.70
8.	propane	369.82	377.34	374.03	85.	*n*-heptane	540.26	538.10	537.16
9.	methyl ethyl ether	437.80	437.25	441.99	86.	2-methylhexane	530.37	537.56	530.56
10.	*n*-propylamine	536.71	546.24	534.96	87.	3-methylhexane	535.25	539.75	533.68
11.	methylal	480.60	479.10	481.77	88.	ethylbenzene	617.17	621.57	629.79
12.	isopropylamine	471.85	467.64	483.87	89.	*m*-xylene	617.05	623.19	624.84
13.	*n*-propylamine	496.95	487.17	506.73	90.	*p*-xylene	616.26	624.89	625.23
14.	trimethylamine	433.25	431.82	420.41	91.	2,6-xylenol	701.05	697.97	688.76
15.	1,3-butadiene	425.37	422.99	421.83	92.	*N,N*-dimethylaniline	687.15	690.89	656.36
16.	*n*-butyronitrile	582.25	570.98	588.14	93.	*cis*-1,2-dimethylcyclohexane	606.15	602.39	598.02
17.	1-butene	419.59	420.76	431.70	94.	*trans*-1,2-dimethylcyclohexane	596.15	594.68	597.30
18.	*cis*-2-butene	435.58	432.85	423.16	95.	*cis*-1,3-dimethylcyclohexane	591.15	584.21	590.41
19.	*trans*-2-butene	428.63	429.40	422.84	96.	*trans*-1,3-dimethylcyclohexane	598.00	589.51	590.91
20.	isobutene	417.90	419.97	417.19	97.	*trans*-1,4-dimethylcyclohexane	590.15	589.74	591.78
21.	methyl ethyl ketone	535.50	524.94	512.05	98.	ethylcyclohexane	609.15	604.01	596.37
22.	tetrahydrofuran	540.15	541.77	551.21	99.	isobutyl isobutyrate	602.00	596.77	608.23
23.	*n*-butyric acid	628.00	626.38	634.07	100.	2,3-dimethylhexane	563.40	568.57	564.01
24.	ethyl acetate	523.30	521.81	524.90	101.	2-methyl-3-ethylpentane	567.00	568.62	562.97
25.	isobutyric acid	609.15	615.98	623.50	102.	*n*-octane	568.83	567.17	563.96
26.	methyl propionate	530.60	524.71	514.76	103.	2,2,3-trimethylpentane	563.50	561.58	569.08
27.	*n*-propyl formate	538.00	526.37	554.51	104.	2,2,4-trimethylpentane	543.96	548.71	542.91
28.	*n*-butane	425.18	427.73	430.06	105.	2,3,3-trimethylpentane	573.50	567.55	580.19
29.	isobutane	408.14	414.12	411.74	106.	2-ethyl-1-hexanol	640.25	640.88	634.80
30.	*n*-butanol	562.93	571.05	563.55	107.	quinoline	782.15	772.04	755.25
31.	*sec*-butyl alcohol	536.01	549.09	523.56	108.	cumene	631.15	638.98	645.18
32.	*tert*-butyl alcohol	506.20	528.31	506.63	109.	*o*-ethyltoluene	651.15	654.31	647.25
33.	diethyl ether	466.70	470.10	469.99	110.	*p*-ethyltoluene	640.15	644.34	644.79
34.	isobutanol	547.73	558.92	537.09	111.	mesitylene	637.36	648.27	639.38
35.	*n*-butylamine	531.90	522.22	540.99	112.	*n*-propylbenzene	638.38	645.53	647.46
36.	diethylamine	496.60	495.60	490.31	113.	1,2,3-trimethylbenzene	664.53	666.26	647.93
37.	pyridine	619.95	602.92	640.86	114.	1,2,4-trimethylbenzene	649.13	658.43	644.51
38.	cyclopentane	511.76	522.85	525.90	115.	*n*-propylcyclohexane	639.15	626.41	614.22
39.	1-pentene	464.78	464.68	472.30	116.	3,3-diethylpentane	610.05	605.66	602.95
40.	diethyl ketone	560.95	552.05	541.58	117.	*n*-nonane	595.65	592.85	588.58
41.	2-pentanone	561.08	552.43	542.79	118.	2,2,3,3-tetramethylpentane	610.85	598.50	624.19
42.	ethyl propionate	546.00	548.54	543.66	119.	1,2,3,4-tetrahydronaphthalene	720.15	720.49	697.26
43.	isobutyl formate	551.35	547.29	571.69	120.	*n*-butylbenzene	660.55	669.84	665.34
44.	*n*-propyl acetate	549.40	551.45	553.60	121.	*p*-cymene	653.15	656.66	661.51
45.	valeric acid	651.00	653.34	654.07	122.	isobutylbenzene	650.15	659.20	659.98
46.	isopentane	460.43	462.11	464.01	123.	*n*-decane	618.45	616.34	611.59
47.	neopentane	433.78	439.86	436.01	124.	*n*-nonadecane	658.20	657.86	654.44
48.	*n*-pentane	469.65	472.09	472.53	125.	*n*-tetradecane	692.40	693.43	694.48
49.	2-methyl-2-butanol	545.15	552.06	535.65	126.	*n*-octadecane	745.26	750.76	770.12
50.	3-methyl-1-butanol	579.45	587.48	581.31	127.	*sec*-butyl chloride	520.60	510.94	517.75
51.	1-pentanol	586.15	595.48	587.97	128.	methyl isopropyl ether	464.50	465.66	464.64
52.	bromobenzene	670.15	645.59	650.52	129.	*sec*-butylamine	514.30	504.75	520.32
53.	chlorobenzene	632.35	618.93	651.49	130.	*tert*-butylamine	483.90	482.19	492.46
54.	benzene	562.16	560.26	601.26	131.	2-methyl-1-butene	465.00	466.12	460.90
55.	phenol	694.25	682.11	692.89	132.	2-methyl-2-butene	471.00	475.11	457.09
56.	aniline	699.00	684.82	689.07	133.	3-methyl-1-butene	450.37	452.67	461.28
57.	2-methylpyridine	621.00	617.36	641.01	134.	methyl isopropyl ketone	553.00	542.84	539.49
58.	cyclohexane	560.40	563.75	547.21	135.	methyl *n*-butyrate	554.50	552.97	543.27
59.	cyclohexanone	629.15	648.10	624.65	136.	piperidine	594.05	592.17	595.39
60.	cyclohexane	553.54	561.02	557.21	137.	ethyl propyl ether	500.23	505.80	502.19
61.	1-hexene	504.03	505.34	506.55	138.	methyl *tert*-butyl ether	497.10	495.29	485.10
62.	methylcyclopentane	532.79	550.21	546.43	139.	3-methylpyridine	645.00	635.24	650.62
63.	cyclohexanol	625.15	653.24	642.59	140.	1,5-hexadiene	507.00	500.46	495.45
64.	2-hexanone	587.05	583.23	570.20	141.	hexanenitrile	622.05	623.73	639.70
65.	methyl isobutyl ketone	571.40	569.65	580.70	142.	3-hexanone	582.82	578.14	560.33
66.	ethyl *n*-butyrate	571.00	571.78	573.88	143.	ethyl isobutyrate	553.15	561.58	571.50
67.	isobutyl acetate	561.00	569.83	572.52	144.	*n*-propyl propionate	578.00	573.00	570.29
68.	2,2-dimethylbutane	488.78	488.66	494.07	145.	1-hexanol	611.35	614.47	608.62
69.	2,3-dimethylbutane	499.98	498.66	503.51	146.	2-hexanol	586.20	598.02	575.17
70.	*n*-hexane	507.43	511.70	507.35	147.	4-methyl-2-pentanol	574.40	588.08	554.88
71.	2-methylpentane	497.50	501.43	498.80	148.	acetone[d]	508.20	496.61	486.97
72.	3-methylpentane	504.43	505.08	503.41	149.	2-propanol[d]	508.31	528.11	504.38
73.	diisopropyl ether	500.05	511.18	509.44	150.	ethyl vinyl ether[d]	475.15	471.46	459.92
74.	di-*n*-propyl ether	530.60	531.51	532.23	151.	isovaleric acid[d]	634.00	640.73	649.66
75.	diisopropylamine	523.10	530.10	528.23	152.	*n*-butyl acetate[d]	579.15	575.64	579.22
76.	di-*n*-propylamine	555.80	553.56	550.85	153.	*p*-cresol[d]	704.65	703.73	699.85
77.	benzaldehyde	695.00	676.48	685.31	154.	*p*-toluidine[d]	693.15	701.23	697.64

Table 5 (Continued)

	compound	exptl[a]	T_c (bp)[b]	T_c (struct)[c]		compound	exptl[a]	T_c (bp)[b]	T_c (struct)[c]
155.	1-heptene[d]	537.29	536.36	536.39	161.	*p*-diethylbenzene[d]	657.96	665.28	664.06
156.	methylcyclohexane[d]	572.19	574.42	574.71	162.	1-decene[d]	617.05	615.79	610.42
157.	2,2,3-trimethylbutane[d]	531.17	526.44	543.27	163.	propionitrile[d]	564.40	546.42	555.85
158.	*o*-xylene[d]	630.37	632.25	629.26	164.	4-methylpyridine[d]	646.15	636.71	648.80
159.	1-octene[d]	566.60	565.97	562.88	165.	*n*-pentyl formate[d]	576.00	585.69	603.00
160.	*m*-ethyltoluene[d]	637.15	648.08	644.42					

[a] DIPPR experimental critical temperatures (see ref 4). [b] Values obtained using the experimental boiling point as the primary variable. [c] Values obtained using structurally derived parameters alone. [d] Compound was a member of the external prediction set.

interesting to note that the same hydrogen bonding descriptor and a unique dipole–dipole descriptor were found to offer useful critical temperature information. The count of the number of oxygens likewise proved important with each property. The remaining features contributed electronic, geometrical, branching, and group contribution information. Finally, the external applicability of the model is excellent. The correlation of 0.989 is identical to that obtained on the training set, while the accuracy is slightly better, 10.80 K versus 11.88 K. Critical temperature estimations for the individual compounds are presented in Table 5.

CONCLUSIONS

Several conclusions can be drawn from the structure–property investigations of the DIPPR database. First, in contrast to the Beilstein boiling point studies, where inaccurate data greatly impeded model-building progress, the DIPPR data proved to be consistently reliable. This resulted in the generation of comprehensive, high-quality predictive equations which relate molecular structural features to the normal boiling points of diverse sets of organic compounds. Secondly, the results conclusively demonstrate that the predictive capabilities of the ADAPT system exceed those of the more widely used Joback group contribution approach. Finally, neural networks utilizing the BFGS quasi-Newton training optimization algorithm proved to yield results which were superior to those acquired using the more widely employed back-propagation training algorithm. Further investigations of the DIPPR data will reveal whether the quality of the developed models (i.e., both regression and neural network) will be maintained, or compromised, as the complexity of the studied structures is steadily increased.

As a property which is thought to be closely associated with boiling points, critical temperatures may depend on similar intra- and intermolecular interactions. Two critical temperature models were developed to investigate this hypothesis. The experimental boiling point served as the principal parameter in the first predictive equation. The unparalleled strength of this variable (relative to a pool of calculated descriptors) supports the contention that there is an intricate link between boiling point and critical temperature processes. The second model, by contrast, was built exclusively from structure-based descriptors. Although this model yields slightly less accurate estimations, it has the advantage that its use does not hinge on the availability of high-quality boiling point data. The independent successes of these diverse modeling approaches illustrate the utility and flexibility of the ADAPT system. ADAPT can be used not only to estimate properties strictly from structural descriptors but also to explore whether one physical property is an important determinant of another. Through this work we have demonstrated that critical temperatures may be strongly interwined with boiling points. Future work will establish if the same holds true for other properties (e.g., liquid molar volumes and enthalpies of vaporization) which are hypothesized to be related to boiling point processes in a similar manner.

ACKNOWLEDGMENT

The funding for this work was provided by the Design Institute for Physical Property Data (DIPPR).

REFERENCES AND NOTES

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

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

(3) Egolf, L. M.; Jurs, P. C. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616.

(4) Design Institute for Physical Property Data (DIPPR). *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation*; Daubert, T. E., Danner, R. P., Eds.; Hemisphere Publishing: New York, 1989; Vols. 1–4.

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

(6) Fletcher, R. *Practical Methods of Optimization*; Wiley: New York, 1980; Vol. 1.

(7) Copeman, T. W.; Mathias, P. M.; Klotz, H. C. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: New York, 1988; p 351.

(8) Joback, K. G. M.S. dissertation, The Massachusettes Institute of Technology, Cambrige, MA, 1984.

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

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

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

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

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

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

(15) 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.

(16) Petrucci, R. H. *General Chemistry. Principles and Modern Applications*, 3rd ed.; MacMillan: New York, 1982; p 266.

(17) Streitwieser, A., Jr.; Heathcock, C. H. *Introduction to Organic Chemistry*, 2nd ed.; MacMillan: New York, 1981.

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

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

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

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

(22) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure-Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.*, in press.

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