Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure

Matthew D. Wessel and Peter C. Jurs*

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

Received June 16, 1994[⊗]

Computer assisted methods are used to investigate the relationship between normal boiling point and molecular structure for a set of hydrocarbons. Multiple linear regression methods are used to develop a six-variable linear model with a low root mean square (rms) error. The six descriptors in the linear model are also used to develop a computational neural network model with a significantly lower rms error. The methodology used in this study is also compared to Joback's group contribution method to estimate physical properties. The methods used here are found to be superior to Joback's method. However, when one additional variable encoding the square root of the molecular weight is added to Joback's groups, an excellent model is developed.

INTRODUCTION

The physical properties of organic compounds are often used as a means of determining the safety protocols associated with their usage. The normal boiling point of a compound is certainly one of the properties utilized. However, the boiling point is not always available in tables or other reference materials. As a result, methods of estimation have become increasingly important and useful to the producers and consumers of many organic chemicals. The development of models to predict boiling points, as well as other physical properties, from the structural features of a molecule is of high value because of this need.

There have been several investigations into the relationship between normal boiling point (NBP) and molecular structure. The validation of data in the Beilstein Institute physical property database was an important motivation behind some of those studies. Several classes of organic compounds taken from the Beilstein database have been studied. These include furans, tetrahydrofurans, and thiophenes[1] as well as pyrans and pyrroles.[2] More recently, the Design Institute for Physical Property Data (DIPPR)[3] has become interested in quantitative structure−property relationship (QSPR) methods to estimate and validate boiling points for their database. A structurally diverse set of organic compounds was taken from the DIPPR Project 801 database,[4] and a structure-boiling point relationship was developed.[5] Improvements to current QSPR methods are presently being developed and studied. The use of computational neural networks to enhance the predictve ability of a boiling point model has been shown.[6] The previous study of compounds taken from the DIPPR database is but one of several studies that demonstrate the applicability and power of ADAPT, the QSPR software system developed over the years in our laboratory.[7]

In the ADAPT approach to QSPR studies, the molecular structures are encoded by a set of descriptors (features, variables, attributes) that collectively characterize the structures of the compounds. The descriptors encode the aspects of the structures that are then related to the property under investigation, in this case the normal boiling point. The information stored in the descriptors is used in conjunction with multiple linear regression analysis or computational neural networks to develop models that can predict the normal boiling points. The goal is to identify a set of descriptors that support an accurate model for the prediction of the property of interest.

The relationship between structure and boiling point not only allows for the development of predictive models, but the descriptors in the model can also lend insight into the structural determinants of boiling point. A better understanding of the boiling phenomenon is thus obtained when a QSPR study is successful. The calculated boiling point value can also be used to estimate other properties such as critical temperature,[5,8] molar volume,[9] enthalpy of vaporization,[10] and chromatographic retention indices.[11]

The data set of hydrocarbons was chosen because of inconsistencies uncovered in the prior studies of diverse compounds. The wide range of compounds was difficult to model.[5] Developing a model to predict boiling point values for a homogeneous set of compounds should be relatively straightforward. By starting with a homogeneous data set, a natural progression into a more diverse data set can be made once the preliminary models have been developed. A model database can also be created when using this approach, where each model is specific to a certain class of compounds.

Presented here is a QSPR study using the ADAPT methodology for a set of 356 hydrocarbon compounds from the DIPPR Project 801 database.[12] The linear models developed with regression analysis are highly accurate. A computational neural network model was also developed, and it provides a significant improvement over the linear model. Joback's method of group contribution[13] is also compared to the ADAPT results. Since Joback's group contribution approach is a widely accepted method of boiling point estimation, the ADAPT model was compared to Joback's method for the identical set of compounds. The results show that the ADAPT approach provides superior models for the prediction of normal boiling points. However, when one additional variable encoding the square root of the molecular weight is added to Joback's groups, a surprisingly high quality model is produced.

EXPERIMENTAL SECTION

All computations involving the ADAPT software system and other software were carried out on a DEC 3000 AXP

[⊗] Abstract published in *Advance ACS Abstracts*, November 15, 1994.

0095-2338/95/1635-0068$09.00/0

Model 500 workstation. DIPPR provided the names and normal boiling points for all compounds in the data set.

Data Set. All the hydrocarbon compounds in the DIPPR Project 801 database[12] were used to form a working set of 362 compounds. Due to limitations with some of the ADAPT descriptor routines, six allenes were removed from further consideration. Of the remaining 356 compounds, 40 had calculated values for normal boiling point (DIPPR methods of estimation) as opposed to experimental values. These 40 compounds were placed in a new working set with 16 other compounds chosen at random to form a 56-member external prediction set. These 56 compounds were never used in any step of model development during this study. The remaining 300 structures comprised the training set. If the compounds with predicted boiling point values had been used to develop a model, the integrity of the model would be in question. The training set was used to develop models, and the external prediction set was used to validate the best model developed. Table 1 provides a list of the 356 compounds and the experimental normal boiling point values used in this study.

Uncertainties. The uncertainties associated with the normal boiling points in the data set were also supplied by DIPPR. Over 95% of the compounds were coded with an error of less than 1%. The boiling point values spanned the range from 169.4 to 770.1 K, and the average computed error was approximately 4.4 K. This value was used as a lower error bound for the models developed in this study.

Molecular Modeling. The structures of the compounds were entered by sketching, and they were then modeled using the program EHNDO (Extended Hückel Neglect of Differential Overlap) developed in our laboratory.[14] Molecular modeling was performed because accurate three-dimensional representations of the molecules were necessary for the development of some descriptors known to be important attributes for encoding the structures.

Descriptor Generation and Analysis. After molecular modeling was completed, descriptor generation was initiated. Three basic descriptor types were generated: topological,[15-20] geometric,[21-24] and electronic.[25] Charged partial surface area (CPSA)[26] descriptors are combinations of electronic and geometric descriptors that encode the capability for polar interactions. A total of 81 descriptors were generated for each compound.

In an effort to reduce the number of descriptors, a series of screening procedures was used to examine the information content of the descriptors in the training set. First, any descriptor with greater than 80% identical values was removed. Pairwise correlations among descriptors were examined next. One of any two descriptors with a correlation coefficient $r > 0.94$ was removed. Simplicity of calculation and/or ease of interpretation was used as a basis for deciding which of the two descriptors to retain. A vector space descriptor analysis routine that uses a Gram–Schmidt orthogonalization[27] to rank the descriptors on the basis of mutual orthogonality was also helpful in narrowing the descriptor pool. Collectively, these methods were effective in reducing the descriptor pool to about 40 members. This reduced descriptor pool was then screened further using multiple linear regression analysis.

Regression Analysis. Regression equations are of the form

$$BP_j = b_0 + \sum_{i=1}^{n} (b_i X_{ij}) \quad (1)$$

where BP_j is the normal boiling point of the jth compound, b_0 is the y intercept, b_i is the coefficient of descriptor X_{ij} for compound j, and n is the number of descriptors in the regression model. Regression by leaps-and-bounds[28] was used to develop some preliminary models for the training set using the 40-member descriptor pool. Leaps-and-bounds regression analysis uses the R^2 criterion to search for combinations of descriptors, but it does not provide any information concerning the internal stability of the models it discovers. Since our leaps-and-bounds routine is limited to 24 descriptors, some possible combinations of descriptors were probably overlooked. A genetic algorithm[29,30] descriptor selection routine recently coded in our laboratory was used to examine a descriptor space containing more than 24 members. The results of this search suggested several high quality models. An interactive regression analysis technique was used to refine the models suggested by the genetic algorithm search routine. Interactive regression analysis allows the user to observe how models change as new descriptors are added and older ones removed. It also provides a greater array of statistical information concerning the internal stability of the models developed.

Neural Networks. Computational neural networks were used to improve the accuracy of the linear predictions. The theory behind the BFGS (Broyden–Fletcher–Goldfarb–Shanno) quasi-Newton optimization as applied to neural networks has been discussed in detail elswhere.[31,32] The BFGS training algorithm has proven to be more accurate and efficient in training networks than the widely used back-propagation training algorithms.[31] Since neural networks are essentially a nonlinear technique, they do not reinterpret the descriptors in the linear model. Rather, they decrease the rms error of the training and prediction sets while maintaining the same structure–property relationship uncovered in the linear regression portion of the study.

RESULTS AND DISCUSSION

Leaps-and-bounds regression suggested several high quality models. After reviewing the best models, a noticeable nonlinear behavior was apparent. Closer examination revealed that the molecular weight was responsible for this nonlinear behavior. To perhaps rectify the problem, the square root of the molecular weight was calculated and added to the descriptor pool. Leaps-and-bounds regression was repeated, and several new models were discovered, many of which contained the new descriptor. A six-variable model with a high R^2 and low standard deviation of regression was the best model found. Examination of the model showed that the nonlinear behavior had essentially vanished.

The genetic algorithm feature selection routine was also used to search the space of 40 descriptors. This proved to be effective, and a six-variable model of higher quality (lower rms error) than the best model found by leaps-and-bounds regression was discovered. It should be noted that the genetic algorithm routine, much like leaps-and-bounds regression, did not provide any statistical information

Table 1. List of Hydrocarbons Taken from DIPPR Database

no.	compound	exper. NBP (K)	calc. NBP (K)[f]	no.	compound	exper. NBP (K)	calc. NBP (K)[f]
1	propylene	225.4	222.4	76	*n*-nonane	424.0	421.7
2	propane	231.1	230.7	77	2,2,3,3-tetramethylpentane	413.4	408.5
3	1,3-butadiene	268.7	270.1	78	*n*-butylbenzene	456.5	462.5
4	1-butene[a]	266.9	263.0	79	*sec*-butylbenzene	446.5	459.5
5	*cis*-2-butene	276.9	276.1	80	*tert*-butylbenzene	442.3	443.0
6	*trans*-2-butene	274.0	272.8	81	*p*-cymene	450.3	452.7
7	isobutene	266.2	257.1	82	*m*-diethylbenzene	454.3	458.7
8	*n*-butane	272.6	271.1	83	*o*-diethylbenzene	456.6	463.1
9	isobutane	261.4	262.2	84	*p*-diethylbenzene[b]	456.9	458.6
10	cyclopentadiene	314.6	322.4	85	isobutylbenzene	445.9	454.5
11	cyclopentene[b]	317.4	315.3	86	*n*-butylcyclohexane	454.1	450.0
12	isoprene	307.2	304.1	87	1-decene	443.8	441.8
13	cyclopentane[a]	322.4	317.6	88	*n*-decane	447.3	445.9
14	1-pentene	303.1	301.0	89	*m*-diisopropylbenzene	476.3	485.7
15	isopentane	301.0	303.5	90	*p*-diisopropylbenzene	483.6	485.5
16	neopentane	282.6	277.4	91	bicyclohexyl	512.2	506.8
17	*n*-pentane	309.2	307.4	92	1-dodecene	486.5	485.9
18	benzene	353.2	366.8	93	*n*-dodecane	489.5	489.7
19	cyclohexene	356.1	348.9	94	diphenylmethane[b]	537.4	545.4
20	cyclohexane	353.9	351.7	95	1,1-diphenylethane	545.8	554.4
21	2,3-dimethyl-1-butene	328.8	330.1	96	1,2-diphenylethane	553.6	563.8
22	2,3-dimethyl-2-butene[b]	346.3	341.4	97	1-tetradecene	524.2	524.7
23	2-ethyl-1-butene	337.8	334.2	98	*n*-tetradecane[a]	526.7	528.7
24	1-hexene	336.6	334.9	99	1-hexadecene	558.0	558.9
25	*cis*-2-hexene	342.0	344.8	100	*n*-hexadecane	560.0	462.5
26	*trans*-2-hexene	341.0	340.6	101	1-octadecene	588.0	587.2
27	methylcyclopentane	345.0	340.8	102	*n*-octadecane	589.9	590.7
28	2-methyl-1-pentene	335.2	333.0	103	*n*-nonadecane	603.0	603.3
29	4-methyl-1-pentene	327.0	328.7	104	2-methyl-1-butene	304.3	299.5
30	2,2-dimethylbutane	322.9	322.1	105	2-methyl-2-butene[b]	311.7	306.0
31	2,3-dimethylbutane	331.1	336.4	106	3-methyl-1-butene	293.2	294.5
32	*n*-hexane[b]	341.9	339.8	107	1,3-cyclohexadiene	353.5	353.5
33	2-methylpentane	333.4	335.3	108	methylcyclopentadiene	345.9	352.2
34	3-methylpentane[a]	336.4	339.4	109	2,3-dimethyl-1,3-butadiene	341.9	339.6
35	toluene	383.8	392.2	110	1,5-hexadiene	332.6	337.4
36	ethylcyclopentane	376.6	373.3	111	3,3-dimethyl-1-butene[b]	314.4	313.8
37	1-heptene	366.8	365.2	112	2-methyl-2-pentene	340.4	338.7
38	methylcyclohexane	374.1	373.0	113	3-methyl-1-pentene	327.3	332.0
39	2,3-dimethylpentane	362.9	367.6	114	ethane	184.6	188.2
40	*n*-heptane	371.6	369.4	115	*n*-undecane[b]	469.1	468.3
41	2-methylhexane	363.2	364.1	116	*n*-tridecane	508.6	509.7
42	3-methylhexane	365.0	367.8	117	*n*-pentadecane	543.8	546.4
43	2,2,3-trimethylbutane[b]	354.0	357.3	118	*n*-heptadecane	575.3	577.3
44	styrene	418.3	415.2	119	*n*-eicosane	616.9	613.6
45	ethylbenzene	409.3	416.7	120	*n*-heneicosane[a,c]	629.6	
46	*m*-xylene	412.3	414.8	121	*n*-docosane[a,c]	641.8	
47	*o*-xylene	417.6	419.1	122	*n*-tricosane[a,c]	653.3	
48	*p*-xylene	411.5	414.7	123	*n*-tetracosane[a,c]	664.5	
49	*cis*-1,2-dimethylcyclohexane	402.9	401.1	124	*n*-pentacosane[a,c]	675.0	
50	*trans*-1,2-dimethylcyclohexane	396.6	399.9	125	*n*-hexacosane[a,c]	685.3	
51	*cis*-1,3-dimethylcyclohexane	393.2	395.9	126	*n*-octacosane[a,c]	704.8	
52	*trans*-1,3-dimethylcyclohexane	397.6	397.4	127	*n*-triacontane[a,c]	722.8	
53	*cis*-1,4-dimethylcyclohexane[b]	397.5	397.3	128	*n*-dotriacontane[a,c]	738.8	
54	*trans*-1,4-dimethylcyclohexane	392.5	395.9	129	*n*-hexatriacontane[a,c]	770.1	
55	ethylcyclohexane	404.9	402.6	130	2-methylheptane	390.8	389.8
56	1-octene	394.4	392.5	131	3-methylheptane	392.1	394.0
57	*n*-propylcyclopentane	404.1	399.9	132	4-methylheptane	390.9	394.2
58	2,4,4-trimethyl-1-pentene	374.6	371.9	133	2-methyloctane	416.4	413.2
59	2,4,4-trimethyl-2-pentene	378.1	378.2	134	3-methyloctane	417.4	418.0
60	2,3-dimethylhexane	388.8	392.9	135	4-methyloctane[a]	415.6	417.8
61	2-methyl-3-ethylpentane	388.8	397.1	136	2-methylnonane[b]	440.1	435.7
62	*n*-octane	398.8	396.6	137	3-methylnonane	440.9	440.8
63	2,2,3-trimethylpentane[b]	383.0	384.2	138	4-methylnonane	438.8	440.5
64	2,2,4-trimethylpentane	372.4	378.8	139	5-methylnonane	438.3	440.6
65	2,3,3-trimethylpentane[a]	387.9	388.6	140	2,2-dimethylpentane	352.3	351.1
66	α-methylstyrene	438.6	435.8	141	2,4-dimethylpentane	353.6	362.1
67	cumene	425.6	434.9	142	3,3-dimethylpentane	359.2	359.5
68	*m*-ethyltoluene	434.5	436.1	142	2,2-dimethylhexane	380.0	376.2
69	*o*-ethyltoluene	438.3	439.0	144	2,4-dimethylhexane	382.6	390.1
70	*p*-ethyltoluene	435.2	436.0	145	2,5-dimethylhexane	382.3	386.9
71	*n*-propylbenzene	432.4	439.3	146	3,3-dimethylhexane[b]	385.1	385.0
72	1,2,3-trimethylbenzene	449.3	440.7	147	3,4-dimethylhexane	390.9	396.5
73	1,2,4-trimethylbenzene	442.5	437.9	148	2,2-dimethylheptane	405.8	398.2
74	*n*-propylcyclohexane[b]	429.9	426.9	149	2,6-dimethylheptane	408.4	409.1
75	3,3-diethylpentane	419.3	421.5	150	2,2-dimethyloctane	430.0	418.1

Table 1 (Continued)

no.	compound	exper. NBP (K)	calc. NBP (K)[f]	no.	compound	exper. NBP (K)	calc. NBP (K)[f]
151	2,3-dimethyloctane	437.5	437.0	226	cyclooctene	416.1	412.4
152	2,4-dimethyloctane	429.0	433.4	227	1-phenylindene[a,c]	610.0	
153	2,5-dimethyloctane	431.6	433.7	228	1,2,3-trimethylindene[a,c]	509.0	
154	2,6-dimethyloctane	433.5	433.7	229	β-pinene	439.2	431.5
155	2,7-dimethyloctane	433.0	430.2	230	α-pinene	429.3	437.9
156	3-ethylpentane[b]	366.6	373.4	231	*cis*-1,3-pentadiene	317.2	307.4
157	3-ethylhexane	391.7	400.0	232	*trans*-1,3-pentadiene[b]	315.2	306.7
158	3-methyl-3-ethylpentane	391.4	392.0	233	1,4-pentadiene	299.1	303.8
159	2,2,3,3-tetramethylbutane[a]	379.4	380.8	234	camphene	433.6	434.6
160	2,3,4-trimethylpentane	386.6	393.5	235	1,4-cyclohexadiene	360.1	351.7
161	2,2-dimethyl-3-ethylpentane	407.0	409.2	236	adamantane[a]	461.0	
162	2,4-dimethyl-3-ethylpentane	409.9	418.9	237	*trans*-1,4-hexadiene	338.1	336.9
163	3-ethylheptane	416.3	422.2	238	*cis,trans*-2,4-hexadiene	356.6	347.5
164	2,2,3,4-tetramethylpentane	406.2	409.6	239	*trans,trans*-2,4-hexadiene	355.0	345.9
165	2,2,4,4-tetramethylpentane[d]	395.4		240	1,5-cyclooctadiene[a]	423.3	413.5
166	2,3,3,4-tetramethylpentane	414.7	414.5	241	4-vinyl-1-cyclohexene[a,c]	401.0	
167	2,2,5-trimethylpentane	397.2	398.4	242	2,5-dimethyl-1,5-hexadiene[a]	387.4	389.4
168	2,4,4-trimethylhexane[b]	403.8	406.7	243	2,5-dimethyl-2,4-hexadiene	407.6	396.0
169	squalane[a,c]	720.0		244	dicyclopentadiene	443.0	466.6
170	cyclopropane	240.4	245.1	245	acetylene	189.4	190.5
171	cyclobutane	285.7	282.1	246	methylacetylene[a]	249.9	235.0
172	cycloheptane	391.9	385.6	247	vinylacetylene[b]	278.2	282.4
173	cyclooctane	423.8	416.2	248	dimethylacetylene	300.1	290.1
174	1,1-dimethylcyclopentane	361.0	360.6	249	ethylacetylene	281.2	274.7
175	*cis*-1,2-dimethylcyclopentane	372.7	372.1	250	2-methyl-1-buten-3-yne	305.4	315.7
176	*trans*-1,2-dimethylcyclopentane	365.0	370.4	251	1-pentene-3-yne	332.4	319.2
177	*cis*-1,3-dimethylcyclopentane	363.9	365.7	252	1-pentene-4-yne	315.6	314.9
178	*trans*-1,3-dimethylcyclopentane	364.9	365.4	253	3-methyl-1-butyne	302.1	306.1
179	isopropylcyclopentane[b]	399.6	398.3	254	1-pentyne	313.3	312.7
180	1-methyl-1-ethylcyclopentane	394.7	391.9	255	2-pentyne	329.3	317.6
181	*n*-butylcyclopentane	429.8	424.5	256	1-hexyne	344.5	346.0
182	1,1-dimethylcyclohexane	392.7	388.9	257	2-hexyne	357.7	347.7
183	isopropylcyclohexane	427.9	424.5	258	3-hexyne	354.3	345.4
184	*n*-decylcyclohexane	570.8	563.2	259	1-heptyne	372.9	375.8
185	*cis*-decalin	469.0	466.6	260	1-octyne	399.3	402.7
186	*trans*-decalin	460.5	465.4	261	*n*-pentylbenzene	478.6	484.1
187	ethylene	169.4	181.2	262	*n*-hexylbenzene[b]	499.3	504.4
188	1-nonene	420.0	417.9	263	*n*-heptylbenzene	519.2	523.8
189	1-undecene[b]	465.8	464.4	264	*n*-octylbenzene[a]	537.5	451.7
190	1-tridecene	505.9	505.7	265	*n*-nonylbenzene	555.2	558.7
191	1-pentadecene	541.6	542.4	266	*n*-decylbenzene[a,c]	571.0	
192	1-heptadecene	573.5	573.9	267	*n*-undecylbenzene	586.4	
193	1-nonadecene	602.2	599.4	268	*n*-dodecylbenzene[a,c]	600.8	
194	1-eicosene	615.5	610.8	269	*n*-tridecylbenzene[a,c]	614.4	
195	*cis*-2-pentene	310.1	312.2	270	*n*-tetradecylbenzene[a,c]	627.1	
196	*trans*-2-pentene	309.5	308.1	271	*n*-pentadecylbenzene[a,c]	639.1	
297	*cis*-3-hexene	339.6	343.2	272	*n*-hexadecylbenzene[a,c]	651.1	
198	*trans*-3-hexene	340.2	339.8	273	*n*-heptadecylbenzene[a,c]	662.1	
199	*cis*-2-heptene[b]	371.6	374.0	274	*n*-octadecylbenzene[a,c]	673.1	
200	*trans*-2-heptene	371.1	370.2	275	mesitylene	437.9	433.6
201	*cis*-3-heptene	368.9	372.5	276	*m*-cymene	448.2	452.7
202	*trans*-3-heptene	368.8	368.8	277	*o*-cymene	451.3	458.3
203	*cis*-2-octene	398.8	401.0	278	2-ethyl-*m*-xylene	463.2	462.3
204	*cis*-3-octene	396.0	399.7	279	2-ethyl-*p*-xylene	460.0	457.6
205	*cis*-4-octene	395.7	399.5	280	3-ethyl-*o*-xylene	467.1	459.3
206	*trans*-2-octene	398.1	397.4	281	4-ethyl-*m*-xylene	461.6	457.6
207	*trans*-3-octene	396.4	395.8	282	4-ethyl-*o*-xylene	462.9	457.1
208	*trans*-r-octene	395.4	397.7	283	5-ethyl-*m*-xylene	456.9	452.9
209	3-methyl-*cis*-2-pentene[b]	340.8	344.2	284	1-methyl-2-*n*-propylbenzene	457.9	461.0
210	3-methyl-*trans*-2-pentene	343.6	342.2	285	1-methyl-3-*n*-propylbenzene	454.9	457.2
211	4-methyl-*cis*-2-pentene	329.5	338.7	286	1-methyl-4-*n*-propylbenzene	456.4	457.2
212	4-methyl-*trans*-2-pentene	331.8	334.4	287	1,2,3,4-tetramethylbenzene[b]	478.2	461.1
213	2-methyl-1-hexene	365.0	362.6	288	1,2,3,5-tetramethylbenzene	471.1	457.8
214	3-methyl-1-hexene	357.0	361.8	289	1,2,4,5-tetramethylbenzene	470.0	456.1
215	4-methyl-1-hexene	359.9	363.9	290	*p-tert*-butylethylbenzene	485.2	472.9
216	2-methyl-1-heptene	392.4	388.0	291	ethynylbenzene	416.0	426.1
217	2-ethyl-1-pentene	367.1	366.6	292	*m*-methylstyrene	444.8	436.3
218	3-ethyl-1-pentene	357.3	362.8	293	*o*-methylstyrene	443.0	436.4
219	2,3,3-trimethyl-1-butene[b]	351.0	350.5	294	*p*-methylstyrene	445.9	435.5
220	2,3-dimethyl-1-hexene	383.6	390.1	295	*cis*-1-propenylbenzene[a]	452.0	443.4
221	2-ethyl-1-hexene[a]	393.1	390.6	296	*trans*-1-propenylbenzene	451.4	445.1
222	1-methylcyclopentene	348.6	348.6	297	*m*-divinylbenzene	472.6	464.2
223	3-methylcyclopentene	338.0	345.1	298	2-phenyl-1-butene[b]	455.1	458.5
224	4-methylcyclopentene	338.8	343.8	299	*cis*-2-phenyl-2-butene	467.8	463.1
225	cycloheptene	387.5	382.2	300	*trans*-2-phenyl-2-butene	447.1	464.8

Table 1 (Continued)

no.	compound	exper. NBP (K)	calc. NBP (K)[f]	no.	compound	exper. NBP (K)	calc. NBP (K)[f]
301	*p*-isopropenylstyrene[a,c]	515.0		329	chrysene	714.1	713.3
302	cyclohexylbenzene	513.3	517.2	330	biphenyl	528.1	531.2
303	*p-tert*-butylstyrene[a,c]	500.0		331	diphenylacetylene[a,c]	573.0	
304	4-isobutylstyrene[a,c]	524.0		332	*cis*-stilbene[a,c]	554.0	
305	naphthalene	491.1	492.7	333	*trans*-stilbene	579.6	577.8
306	1,2,3,4-tetrahydronaphthalene	480.8	477.8	334	*m*-terphenyl[a,c]	650.0	
307	1-methylnaphthalene	517.8	509.1	335	*o*-terphenyl[d]	610.6	
308	2-methylnaphthalene	514.3	506.3	336	*p*-terphenyl	649.1	651.6
309	2,6-dimethylnaphthalene	535.1	519.7	337	2,4-diphenyl-4-methyl-1-pentene[a,c]	614.0	
310	2,7-dimethylnaphthalene	536.1	519.6	338	2,3-dimethyl-2,3-diphenylbutane[a,c]	589.0	
311	1-ethylnaphthalene[b]	531.5	525.4	339	triphenylmethane[d]	632.1	
312	2-ethylnaphthalene	531.0	523.4	340	triphenylethylene[a,c]	669.0	
313	1-*n*-propylnaphthalene	545.9	544.0	341	1,1,2-triphenylethane[a,c]	622.0	
314	1-*n*-butylnaphthalene	562.5	462.3	342	tetraphenylmethane[a,c]	743.0	
315	2,6-diethylnaphthalene[a,c]	576.0		343	tetraphenylethylene[a,c]	760.0	
316	1-*n*-pentylnaphthalene	579.1	578.5	344	1,1,2,2-tetraphenylethane[a,e]	633.1	
317	1-phenylnaphthalene[a]	607.1	623.5	345	*d*-limonene	449.6	441.6
318	1-*n*-hexylnaphthalene	595.1	593.7	346	α-phellandrene	448.1	447.1
319	1-*n*-hexyl-1,2,3,4-tetrahydronaphthalene[a,c]	578.1		347	β-phellandrene	447.1	443.6
320	1-*n*-nonylnaphthalene[a,c]	639.0		348	α-terpinene	450.3	443.0
321	1-*n*-decylnaphthalene[a,c]	652.0		349	γ-terpinene[b]	456.1	441.7
322	acenaphthalene[d]	543.1		350	terpinolene	458.1	448.0
323	acenaphthene	550.5	548.1	351	2-norobrnene	368.6	379.0
324	fluorene	570.4	571.4	352	indene	455.8	455.7
325	anthracene	615.2	609.0	353	indane[a]	451.1	452.3
326	phenanthrene	613.5	611.8	354	5-vinyl-2-norobornene	413.6	422.0
327	fluoranthene[b]	656.0	661.2	355	1-methylindene	471.6	474.0
328	pyrene	668.0	669.2	356	2-methylindene	458.0	472.9

[a] Compound in external prediction set. [b] Compound in cross-validation set. [c] Compound had predicted boiling point (by DIPPR). [d] Compound flagged as an outlier in regression model. [e] Compound an outlier in external prediction set (see text). [f] Boiling point calculated using the final neural network model.

concerning the internal stability of the model. Both routines searched for models having low rms error values. Interactive regression analysis was then used to assess the statistical integrity of the model suggested by the genetic algorithm routine. The quality of the best model found was high, with excellent statistical properties and an rms error of 7.1 K for the 300-member training set.

In order to assure further stability, the training set was checked for outliers. Six standard statistical tests were utilized. They were the residual, standardized residual, studentized residual, leverage, DFFITS statistic, and Cooks Distance.[33] If a calculated boiling point value was flagged by four of the six tests, it was considered an outlier and labeled for removal. Four of the 300 compounds were subsequently flagged as outliers. The four compounds were 2,2,4,4-tetramethylpentane, acenaphthalene, *o*-terphenyl, and triphenylmethane. There was no distinct relation among the four outliers, and the four compounds all were known to within 1% of the boiling point value used. The model coefficients were recalculated and the outliers temporarily removed. The resulting model had an rms error of 6.3 K and a multiple correlation coefficient R of 0.997. The overall F statistic for the model was 9215.1 and each variable had a P value of essentially zero. Since it was apparent that the outliers had a significant effect on the coefficients, they were removed permanently. The revised model developed with the 296-member training set is shown in Table 2. A plot of calculated vs observed boiling points for this model is displayed in Figure 1.

Once the final model was determined, three more validation procedures were performed. First, the residual values were plotted against the calculated boiling points. There was no observable pattern in the residual plot. A correlation matrix was constructed to examine the 15 pairwise correlations among the six descriptors in the final model. The range of r values (0.0−0.67) was small, and the number of pairwise correlation coefficients greater than 0.5 was four. Finally, the 16-member external prediction set was used to test the predictive ability of the model. The model performed poorly at first, giving an rms error of 36.9 K. The large rms error was due to the compound 1,2,2,2-tetraphenylene. There were no other compounds in the training set that represented the structural moieties in the outlier. This is because all compounds with structural features similar to 1,1,2,2-tetraphenylethane have predicted boiling point values in the DIPPR database. As a result, these compounds were originally placed in the 56-member external prediction set, and therefore the model poorly estimates the boiling point of 1,1,2,2-tetraphenylethane. When this compound was removed from the 16-member external prediction set, the

Table 2. Final Model for the Prediction of Boiling Points of Hydrocarbons in DIPPR Database

coeff	sd of coeff	label	descriptor definition
237.4	24.86	QNEG	charge on most negative atom[a]
−0.2480	0.01663	DPSA	partial positive minus partial negative surface area[b]
114.2	20.70	FNSA	fractional negative surface area[b]
0.8120	0.06976	ALLP2	total paths/total no. of atoms[c]
−19.29	0.9542	MOLC 7	path cluster 3 molecular connectivity[d]
50.35	0.9302	SQMW	square root of molecular weight
−30.16	5.528		Y intercept
	$R = 0.997$	rms = 6.3 K	$N = 296$ cmpds. $F = 9215.1$

[a] see ref 25. [b] See ref 26. [c] See ref 15. [d] See ref 18.

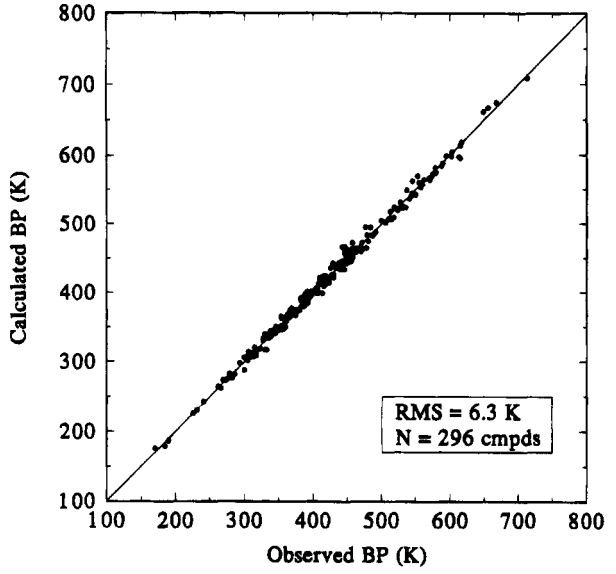

Figure 1. Plot of calculated vs observed normal boiling points for the 296-member training set using the regression model.

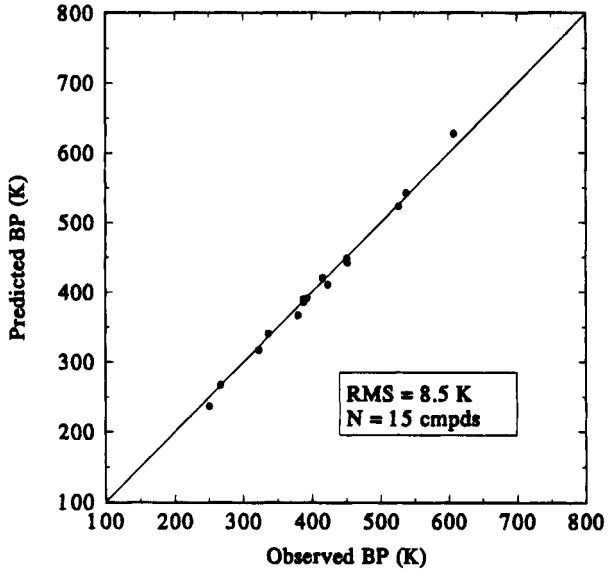

Figure 2. Validation of the linear model using the 15-member external prediction.

rms error was reduced to 8.5 K. A plot of the new 15-member external prediction set is shown in Figure 2. The revised 55-member prediction set (excluding the outlier) containing 40 values predicted with methods used by DIPPR was also used to test the model. The 55 compounds in this external prediction set (with experimental and predicted boiling point values) gave an rms error value of 14.4 K. The most useful and meaningful validation of the final model is, of course, provided by the 15-member prediction set. The final six variable model has the advantage of predicting the normal boiling points for many classes of hydrocarbons accurately and with a small number of variables. Also, since the variables are whole molecule descriptors, rather than substructures, as in group contribution methods, the relationship between boiling point and molecular structure is revealed on a more direct level.

Examination of the model in Table 2 provides some insight into the structural features that may influence the boiling point of a particular compound. The charge on the most negative atom, labeled QNEG,[25] is acting as an indicator variable. For instance, compounds with terminating alkene and alkyne functions are grouped together with relatively large QNEG values. Since the coefficient is positive, larger values tend to give a decrease in the boiling point, and this is observed for several compounds. However, unsaturation will also give rise to small dipole moments for hydrocarbons, and this may help increase the boiling point, so there is an offset factor to consider.

The two charged partial surface area descriptors, DPSA and FNSA,[26] are somewhat linearly related to the boiling point. DPSA, defined as the partial positive surface area minus the partial negative surface area is positive for most hydrocarbons. This is because the negative atoms of a hydrocarbon are the carbon atoms themselves. The positive hydrogen atoms are the main solvent accessible surface area of the compounds, explaining the positive values observed for DPSA. In the case of straight chain hydrocarbons, an increase in surface area will tend to increase the boiling point. However, this is not necessarily a linear relationship. As the number of carbon atoms increases, the boiling point will increase, but with smaller increments. The negative coefficient of DPSA may be encoding this trend. The positive coefficient of the fractional negative surface area, FNSA, may also account for the same trend.

ALLP 2, the total sum of paths (of length 1−44) in a structure divided by the total number of carbon atoms is observed to decrease in value as the branching increases for compounds with the same number of carbon atoms. At the same time, as the number of carbon atoms increases, the total number of paths will increase. This will cause the value of ALLP 2 to rise. Therefore, an increase in branching of hydrocarbons with the same molecular weight causes a decrease in boiling point, but adding another carbon causes an increase in boiling point.[17] ALLP 2 is effectively encoding this information.

As the amount of branching increases in hydrocarbons, the accessible surface area will decrease, and this will drive the boiling point lower. The path cluster-3 molecular connectivity index, MOLC 7, looks specifically for branching in compounds. The path cluster-3 is shown below:

If a compound does not contain this structural feature, there is no branching. Compounds that have this feature are obviously branched, and as branching increases, the value of MOLC 7 rises. Since MOLC 7 has a negative coefficient, the expected trend of decreasing boiling point with an increase in branching is observed.

Boiling point will generally increase as the molecular weight increases. Since this increase is not linear, as was discussed earlier, taking the square root of the molecular weight provides a much more linear correlation to the boiling point and is obviously effective in the model.

To assess the overall quality of this descriptor based approach to property estimation, a comparison between the ADAPT results and Joback's method of group contribution[13] was performed. In previous studies, it has been shown that ADAPT techniques were superior to the group contribution method.[6] The underlying problem with a group contribution approach is that it consistently overpredicts on the low and high end of the data range, thus giving large negative

Table 3. Group Increments for Joback and Comparative Methods

group		Joback	method 2	method 3	method 4
constant		198.00	197.13	188.62	−54.11
$-CH_3$		23.58	20.70	25.38	−13.01
$-CH_2-$	nonring	22.88	23.26	23.18	−5.643
$-CH_2-$	ring	27.15	26.94	28.26	−3.668
>CH−	nonring	21.74	27.90	23.18	−3.122
>CH−	ring	21.78	24.00	22.65	−6.579
>C<	nonring	18.25	27.70	17.81	−1.537
>C<	ring	21.32	17.99	15.36	−9.491
=CH−	nonring	24.96	28.70	28.45	−0.5981
=CH−	ring	26.73	25.17	26.81	−3.192
$=CH_2$		18.18	13.03	17.85	−13.28
=C<	nonring	24.14	33.65	28.97	5.160
=C<	ring	31.01	35.48	27.19	3.395
≡CH		9.20	8.31	12.73	−7.856
≡C−		27.48	36.82	36.48	9.654
⋯CH⋯				26.46	−3.313
⋯C⋯				32.10	7.340
⋯C⋯				36.71	14.15
square root MW					47.58

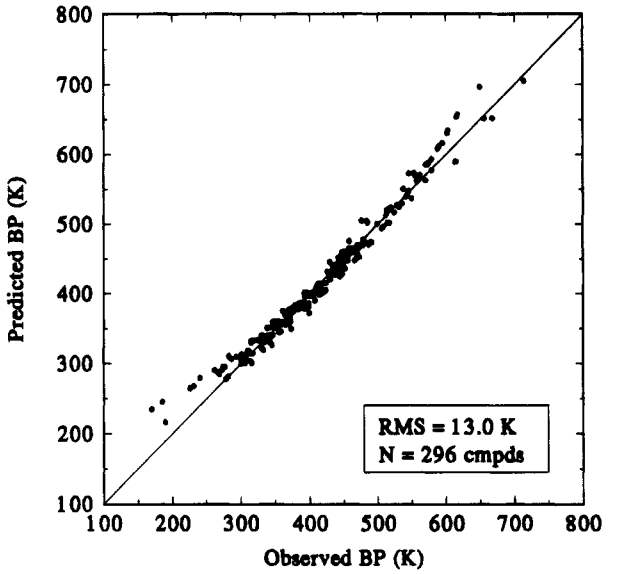

Figure 3. External prediction of the 296-member training set using the Joback coefficients and group counts.

residuals. A possible reason is that group additivities are not always linearly related to boiling points. In any event, it was necessary to compare the results here with the Joback's group contribution method.

The training set of 296 compounds was used to test three different approaches of group contribution. The 14 groups defined by Joback were assigned as counts for each compound.[13] Table 3 lists the value of the coefficients for each group used in the Joback approach. These coefficients were derived by Joback from a structurally diverse set of 438 organic compounds.[13] Using Joback's coefficients, the boiling point for each of the 296 compounds in the training set was predicted. The rms error was 13.0 K as compared to the rms error of 6.3 K for the ADAPT generated model using the same 296 compounds. A plot of the results for this group contribution method is shown in Figure 3. There is a noticeable curvature in the plot, giving rise to poor predictions outside of the intermediate range between about 300−550 K. Within this range, however, it appears that the boiling points are predicted relatively well.

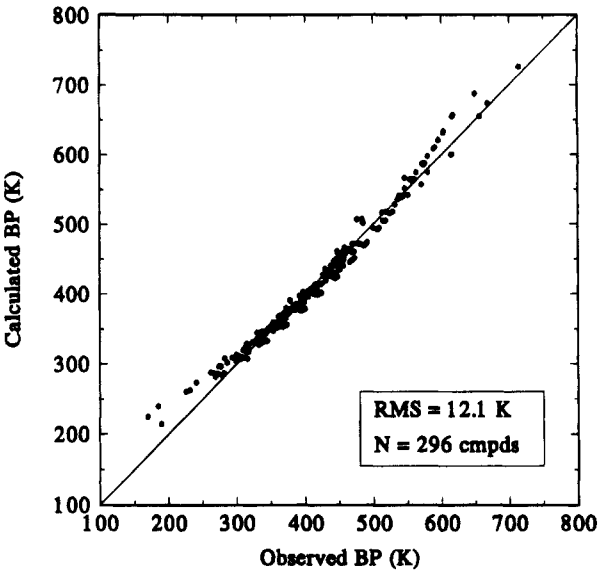

Figure 4. Plot of calculated vs observed normal boiling points for the 296-member training set using regression generated coefficients for the enhanced group counts.

Using each Joback group as a descriptor, a set of new coefficients was generated using multiple linear regression. The new coefficients, listed under method 2 in Table 3, are comparable to the Joback coefficients. A plot of the fitted vs observed boiling points for the 296-member training set shows curvature, but the rms error has decreased slightly to 12.3 K. The useful range of prediction is essentially the same as with the original coefficients, but the overall error is lower.

A third approach was taken that allowed a more detailed representation of aromatics. In the Joback approach, −CH= and =C< ring groups accounted for all aromatic and nonaromatic ring groups. ADAPT has the ability to recognize aromatic bonds. As a result, three new groups were included in the list of groups as shown in Table 3, and the groups mentioned above were restricted to nonaromatic ring systems. A new model was developed that included the aromatic groups as well as the standard Joback groups. The coefficients are shown in Table 3, listed under method 3. The rms error for this group additivity approach was 12.1 K, a slight improvement over the regression of Joback's group contributions. Figure 4 shows the results for this new enhance method. The rms error for any of these three methods is still considerably larger than the rms error associated with the ADAPT model based on calculated structural descriptors. From this comparison, it is clear that the ADAPT approach of descriptor generation, selection, and multiple linear regression is superior to Joback's method of group contribution.

The noticeable curvature observed in the plots of the results of the three group contribution approaches just described (Figures 3 and 4) does present a problem for simple boiling point estimation. Adding a term to account for some of the nonlinearity may improve the results for group contribution methods. The square root of the molecular weight was added to the method 3 approach (Joback's group counts plus three aromatic groups) as an eighteenth variable. These 18 variables were regressed for the 296-member training set, and a new set of coefficients was generated. The coefficients are listed in Table 3 under method 4. This approach produced excellent results. The training set rms error was

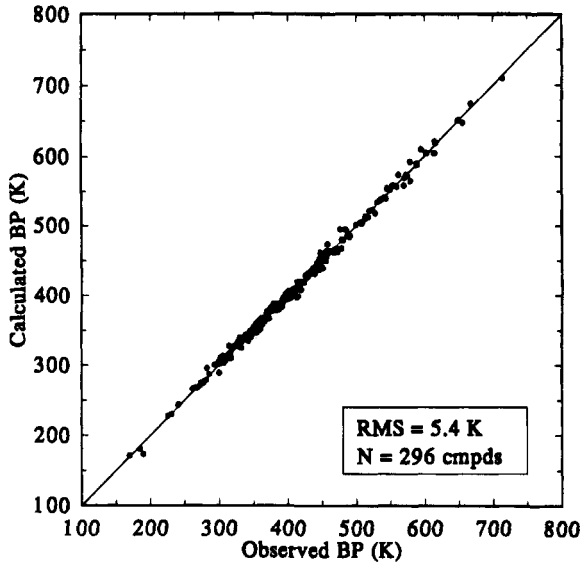

Figure 5. Plot of calculated vs observed normal boiling points for the 296-member training set using the enhanced group contribution method.

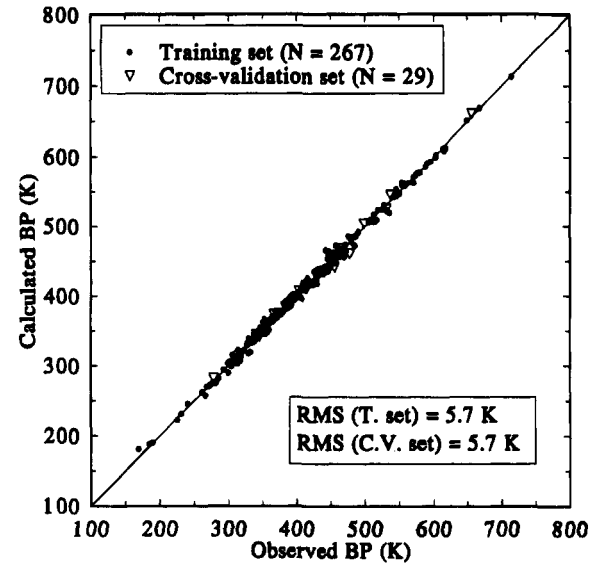

Figure 6. Plot of calculated vs observed normal boiling points for the 267-member training set and the 29-member cross-validation set using neural networks.

5.4 K, and the 15-member external prediction set rms error was 8.7 K. A plot of calculated boiling point vs observed boiling point for the training set is shown in Figure 5. The curvature associated with the first three group contribution methods is essentially absent in this new method. The advantage of this enhancement of Joback's method is that one needs only the coefficients and the molecular structure to estimate the boiling point. A computer is not necessary.

Computational neural networks were used to increase the predictive ability of the linear model developed using standard ADAPT methodology. The six descriptors associated with the linear regression model were fed directly to a fully-connected, feed-forward neural network with a three layer, 6:5:1 architecture. The BFGS quasi-Newton training algorithm was used to train the network, as mentioned previously.

The original training set of 296 compounds used to develop the regression model was split into a new training set of 267 compounds and a cross-validation set of 29 compounds. The original 15-member external prediction set was used to validate the final network model.

A neural network architecture of 6:5:1 was chosen after experimenting with different networks. The 6:5:1 network gave the best results of the many designs investigated. As a general rule, the number of adjustable parameters should be lower than half the number of observations in the training set. For this 6:5:1 setup, there were 41 adjustable parameters for 267 observations, a ratio of 1 to 6.5.

Once a network architecture was chosen, an automated version of the BFGS neural network routine was used to select an optimal network model. The 29-member cross-validation set, which is used to monitor the progress of the network as it trains, was also used to select the point during training that corresponded to the lowest cross-validation set rms error. It is thought that the network's ability to predict the boiling point of an external set of compounds is maximized at this point.[34] The network model chosen performed well, giving an rms error of 5.7 K for the training and cross-validation sets. Figure 6 shows the results for the training and cross-validation sets, respectively. The network model was validated with the 15-member external prediction

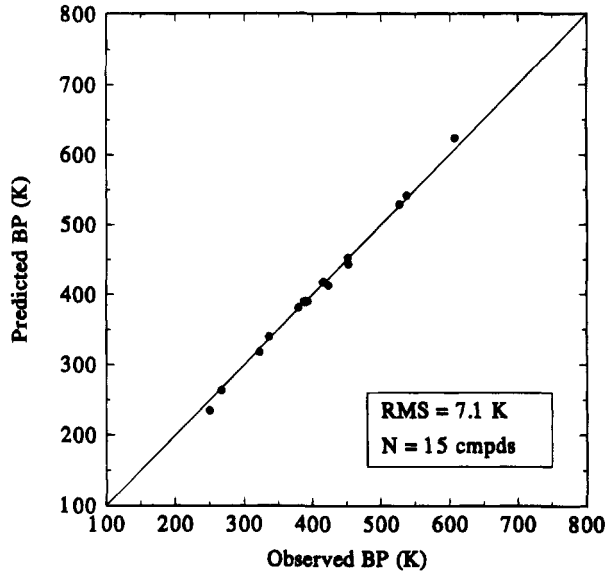

Figure 7. Validation of the neural network model using the 15-member external prediction set.

set. The rms error of prediction was 7.1 K, substantially better than the 8.5 K error from the linear regression model. The results for the prediction set are shown graphically in Figure 7.

Neural networks can improve the accuracy of the linear model primarily because they take advantage of nonlinear relationships between the descriptors and boiling point. The increase in adjustable parameters also helps to improve the accuracy, much like adding another descriptor to a linear model will improve the model. Nonlinear behavior can be seen in the plots of the Joback group contribution approach (Figures 3 and 4). Stein and Brown[35] use a nonlinear correction term for the prediction of boiling points for over 4000 compounds using a group contribution approach. It is clear that nonlinear relationships between structure and boiling point exist and neural networks can use nonlinearity to advantage for the prediction of normal boiling points of hydrocarbons.

CONCLUSIONS

The ADAPT methodology has been used to develop a multiple linear regression model that accurately predicts the normal boiling points of a set of hydrocarbons. This method was also shown to be superior to a group contribution approach, primarily because the descriptors in the linear model account for differences in structure and that simple group counts tend to neglect. However, when the variable encoding the square root of molecular weight was added to the group contribution approach, an excellent model, with the lowest rms error of any developed, was found. It should be noted that this new model had 18 descriptors as opposed to six descriptors in the ADAPT model. This increase in adjustable parameters helps the enhanced group contribution to achieve such excellent errors. This method is extremely convenient for simple estimation of boiling points because no computational tools are necessary.

Neural networks were also used to increase the accuracy of the linear model developed and were effective in doing so. The nonlinear behavior and number of adjustable parameters of the neural network model are the main reasons an increase in quality is observed. When developing models that link molecular structure to physical and chemical properties, the methodology described has been shown in this and previous studies to be extremely effective. Future work will involve the development of models to predict physical properties for organic compounds containing heteroatoms.

ACKNOWLEDGMENT

This research was supported by DIPPR Project 931:Data Prediction Methods, a project with 15 industrial sponsors.

REFERENCES AND NOTES

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

(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**, *33*, 306.

(3) Selover, T. B. DIPPR: Past-Present-Future. *AIChE Symp. Ser.* **1990**, *86*, 90.

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

(6) 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.

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

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

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

(10) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; American Chemical Society: Washington, DC, 1990.

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

(12) Daubert, T. E.; Danner, R. P.; Sibul, H. M.; Stebbins, C. C. *DIPPR Data Compilation of Pure Compound Properties*; Project 801 Sponsor Release, Design Institute for Physical Property Data, AIChE, New York, NY, July 1993.

(13) Joback, K. G. *A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques*. M. S. Dissertation, The Massachusetts Institute of Technology, Cambridge, MA, 1984.

(14) Dixon, S. L.; Jurs, P. C. Fast Geometry Optimization Using a Modified Extended Hückel Method: Results for Molecules Containing H, C, N, O, and F. *J. Comput. Chem.* **1994**, *15*, 733.

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

(16) Randić, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. *Search* for all Self-Avoiding Paths for Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5.

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

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

(19) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley and Sons, Inc.: New York, 1986.

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

(21) Pearlman, R. S In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker, Inc.: New York, 1980.

(22) Vogel, A. I. *Textbook of Practical Organic Chemistry*; Chaucer, 1977, p 1034.

(23) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950.

(24) Miller, K. J.; Savchik, J. A. A New Emperical Method to Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206.

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

(26) 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.

(27) Bradley, G. L. *A Primer of Linear Algebra*; Prentice-Hall, Inc.: New Jersey, 1975.

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

(29) Lucasius, C. B.; Kateman, G. Understanding and Using Genetic Algorithms Part 1. Concepts, Properties and Context. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 1.

(30) Hibbert, D. B. Genetic Algorithms in Chemistry. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 277.

(31) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure-Property Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxical. Chem.* **1994**, *13*, 841.

(32) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480.

(33) Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression* Diagnostics; Wiley: New York, 1980.

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

(35) Stein, S. E.; Brown, R. L. Estimation of Normal Boiling Points from Group Contributions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 581.

CI940068W