Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds 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 March 17, 1995®

Models that accurately predict normal boiling points for organic compounds containing heteroatoms have been developed with regression and computational neural network methods. The structures of the compounds are represented by calculated structural descriptors. Two models are presented—one for a set of 277 compounds containing only O, S, and halogens, and a second for a set of 104 compounds all containing N. Root-mean-square errors of about 9 K result. The accuracy of prediction of these models is compared to a widely used group contribution method for boiling point estimation.

INTRODUCTION

Quantitative structure−property relationships (QSPRs) attempt to quantitatively link attributes of molecular structure to physico-chemical properties. These mathematical links, or models, have a variety of important uses. The prediction of certain properties is an important function that QSPRs fulfill. There has been a substantial amount of interest in the prediction of normal boiling points.[1−3] The normal boiling point is one of the simplest physical properties that can be used to identify an unknown compound, and boiling points can also be used to determine other physical properties, such as heats of vaporization[4] and critical temperatures.[5] Prediction methods have also been important and useful in managing and updating physical property databases.[6,7] These methods provide estimates of boiling points to fill in gaps in databases and can also be used for validation of the data.

The Design Institute for Physical Property Data (DIPPR)[8] has been interested in prediction methods for the purpose of compiling a high quality, accurate database of physical properties for several thousand compounds of industrial importance. In a previous paper a diverse set of 298 organic compounds was taken from the DIPPR Project 801 database[9] and used to develop a boiling point model.[7] The model developed provided clues as to the direction that needed to be taken to develop better models with the DIPPR data. It was apparent that the compounds needed to be divided into subsets. More recently, a set of 356 hydrocarbons was taken from the DIPPR database, and a high quality model for boiling points was developed.[10] This model reinforced the idea that subsetting was the best approach to take in order to ensure that the models developed would be of high quality.

Having built a model for predicting boiling points for hydrocarbons, the next step was to develop a model to estimate boiling points for heteroatom-containing compounds. This paper describes the work completed using the heteroatom-containing compounds taken from the DIPPR Project 801 database. The QSPR software system ADAPT,[11] developed over the years in our laboratory, was used to create the models. We have consistently compared the results obtained with our methodology to the Joback group contribution method,[12] a widely used method for estimating boiling points.

EXPERIMENTAL SECTION

A DEC Alpha 3000 AXP Model 500 workstation running DEC OSF/1 ver. 1.2 B field test software was used to perform all computations involving the ADAPT software system. The names and normal boiling points for all compounds used in this study were provided by DIPPR or taken from other published sources.

Data Set. All organic compounds containing evaluated experimental normal boiling points and one or more heteroatoms were extracted from the current DIPPR database. The range of molecular weights was 27−346 amu, and the number of carbons per compound spanned the range 1−15. The heteroatoms included N, O, S, F, Cl, Br, and I. This screening procedure produced a working data set of 327 compounds. Of these 327 compounds, 27 were chosen randomly and comprised an external prediction set, and the remaining 300 comprised the training set. The external prediction set was never used until after a model had been developed using the training set.

Uncertainties. The experimental error for each compound in the data set was provided by DIPPR. However, the error codes provided only an upper limit of error for each compound. Therefore low and high error bounds were estimated for the data set. The mean absolute error (mae) was estimated at 2.3 K for the low error bound. This corresponded to an rms error of 5.4 K. The high error bound was estimated at 7.6 K mae or an rms error of 11.4 K for the data set. Therefore, in order to prevent overfitting of the data, the rms errors of the neural network models should fall within the approximate range of 5.4 to 11.4 K.

Molecular Modeling. The 327 structures in the data set were modeled using the semiempirical molecular orbital routine MOPAC (vers. 6.0)[13] with the PM3 Hamiltonian.[14] Since there are several molecular interactions crucial to boiling point that are dependent on the geometry of a given compound, molecular modeling was deemed necessary.

Descriptor Generation and Analysis. Molecular descriptors fall into three general categories—topological,[15−20] geometric,[21−24] and electronic.[25] Topological descriptors are numerical representations of the two-dimensional carbon (or heavy atom) skeleton of a molecule. Geometric descriptors rely on the actual three-dimensional conformation of the molecule for the calculation of quantities such as volume, surface area, and moments of inertia. Electronic descriptors include partial atomic charges and numerical combinations of various partial charges in a molecule. The three categories of descriptors are by no means mutually exclusive; for

® Abstract published in *Advance ACS Abstracts,* September 1, 1995.

0095-2338/95/1635-0841$09.00/0

Table 1. List of Compounds Taken from DIPPR Project 801 Database[9] Used To Develop the N-Excluded Data Subset Model

no.	compound	exptl NBP (K)	calcd NBP (K)[d]	no.	compound	exptl NBP (K)	calcd NBP (K)[d]
1	bromotrichloromethane	378.0	385.9	74	methyl ethyl ether[b]	280.5	277.1
2	chlorotrifluoromethane	191.7	190.8	75	1-propanol	370.4	368.3
3	carbon tetrachloride	349.8	360.0	76	1,2-propylene glycol	460.8	457.3
4	carbon tetrafluoride	145.1	161.7	77	1,3-propylene glycol	487.6	485.5
5	tribromomethane	422.3	415.0	78	isopropyl mercaptan[b]	325.7	326.7
6	chlorodifluoromethane	232.3	220.6	79	*n*-propyl mercaptan	340.9	336.7
7	dichlorofluoromethane[b]	282.0	278.5	80	furan	304.5	308.4
8	chloroform[a]	334.3	341.8	81	thiophene	357.3	359.1
9	trifluoromethane	191.0	180.7	82	2,5-dihydrofuran	339.0	335.9
10	dibromomethane[b]	370.1	369.3	83	methacrylic acid	434.2	418.5
11	dichloromethane	312.9	311.9	84	1,2-dichlorobutane	397.1	394.3
12	difluoromethane	221.5	201.6	85	2,3-dichlorobutane	392.6	393.4
13	diiodomethane[c]	455.2		86	1-butanal[b]	348.0	355.7
14	formaldehyde	254.0	267.5	87	methyl ethyl ketone	352.8	342.4
15	formic acid[b]	373.7	402.5	88	2-methylpropanal	337.3	353.0
16	methyl bromide	276.7	290.8	89	tetrahydrofuran[b]	339.1	347.2
17	methyl chloride[a]	248.9	252.0	90	*n*-butyric acid[a]	436.4	439.6
18	methyl fluoride	194.8	204.6	91	ethyl acetate	350.2	346.8
19	methyl iodide	315.6	312.4	92	isobutyric acid	427.7	436.4
20	methanol	337.8	338.5	93	methyl propionate	352.6	338.7
21	methyl mercaptan	279.1	282.4	94	*n*-propyl formate	354.0	360.4
22	1,1-dichlorotetrafluoroethane	276.2	293.1	95	tetrahydrothiophene[a]	394.3	392.3
23	hexafluoroethane[b]	194.9	195.6	96	1-bromobutane	374.8	375.8
24	1,1-dichloro-2,2,2-trifluoroethane	301.0	301.7	97	2-bromobutane	364.4	369.3
25	trichloroethylene	360.1	345.5	98	*n*-butyl chloride[a]	351.6	343.8
26	trichloroacetaldehyde[c]	370.8		99	*sec*-butyl chloride	341.3	334.3
27	pentachloroethane	433.0	429.2	100	*tert*-butyl chloride	323.8	328.6
28	pentafluoroethane	225.1	219.2	101	isobutyl chloride	342.0	338.2
29	1,1,2,2-tetrabromoethane	516.7	520.0	102	1-butanol[a]	390.8	389.1
30	1,1,1,2-tetrachloroethane	403.7	391.9	103	2-butanol	372.7	361.9
31	1,1,2,2-tetrachloroethane	418.3	417.3	104	diethyl ether[b]	307.6	304.2
32	1,1-difluoroethylene	187.5	210.2	105	2-methyl-1-propanol	380.8	381.1
33	1,1,2,2-tetrafluoroethane[b]	250.1	253.2	106	2-methyl-2-propanol	355.6	367.9
34	vinyl bromide	288.9	299.7	107	methyl *n*-propyl ether[a]	312.2	304.9
35	acetyl chloride	323.9	318.7	108	1,3-butanediol	480.2	475.5
36	1,1,1-trichloroethane[a]	347.2	359.3	109	1,4-butanediol[b]	501.2	495.0
37	1,1,2-trichloroethane	387.0	373.1	110	2,3-butanediol	453.9	451.8
38	vinyl fluoride	200.9	226.8	111	2-methyl-1,3-propane	487.2	479.1
39	1,1,1-trifluoroethane	225.8	213.9	112	*n*-butyl mercaptan	371.6	367.1
40	1,1-dibromoethane	381.1	386.7	113	*sec*-butyl mercaptan[a]	358.1	342.0
41	1,2-dibromoethane	404.5	407.8	114	*tert*-butyl mercaptan	337.4	347.3
42	1,1-dichloroethane	330.4	330.3	115	isobutyl mercaptan	361.6	361.7
43	1,2-dichloroethane	356.6	349.0	116	methyl *n*-propyl sulfide	368.7	364.7
44	1,1-difluoroethane	247.4	236.2	117	2-methylthiophene	385.7	383.6
45	1,2-difluoroethane	283.6	268.5	118	methyl isopropyl ketone	367.5	366.2
46	acetic acid[a]	391.1	409.4	119	1-pentanal	376.1	383.0
47	methyl formate	304.9	307.3	120	2-pentanone	375.5	369.4
48	bromoethane	311.5	318.5	121	3-pentanone	375.1	363.2
49	ethyl chloride	285.4	283.8	122	*n*-butyl formate[b]	379.3	385.2
50	ethyl fluoride	235.4	242.7	123	*sec*-butyl formate	366.5	381.6
51	ethyl iodide	345.4	338.0	124	*tert*-butyl formate	356.0	379.7
52	dimethyl ether	248.3	250.8	125	ethyl propionate	372.3	410.3
53	ethanol	351.4	347.8	126	isobutyl formate	371.2	382.6
54	ethylene glycol	470.5	478.7	127	isopropyl acetate	361.6	366.1
55	dimethyl sulfide	310.5	311.2	128	methyl *n*-butyrate	375.9	362.0
56	ethyl mercaptan	308.2	305.6	129	isovaleric acid	448.3	456.7
57	dimethyl disulfide[a]	382.9	399.5	130	neopentanoic acid	437.0	446.1
58	1,2-ethanedithiol	419.2	418.2	131	*n*-pentanoic acid	458.9	458.6
59	2-chloropropene	295.8	297.2	132	*n*-propyl acetate	374.6	371.4
60	1,2,3-trichloropropane	430.0	425.6	133	1-chloropentane	381.5	374.3
61	1,2-dichloropropane[a]	369.5	370.3	134	2,2-dimethyl-1-propanol	386.3	391.0
62	acetone	329.4	321.0	135	ethyl isopropyl ether	326.1	329.1
63	1-propanal	321.1	320.7	136	ethyl propyl ether	337.0	332.7
64	1,2-propylene oxide	307.6	306.1	137	2-methyl-1-butanol[a]	401.9	401.9
65	ethyl formate	327.5	334.3	138	2-methyl-2-butanol	375.1	377.3
66	methyl acetate	330.1	321.6	139	3-methyl-1-butanol	404.4	410.1
67	propionic acid[b]	414.3	413.4	140	3-methyl-2-butanol	384.6	381.5
68	1-bromopropane	344.1	344.0	141	methyl *n*-butyl ether	343.4	333.0
69	isopropyl chloride	308.8	308.9	142	methyl *sec*-butyl ether	332.1	327.4
70	*n*-propyl chloride	319.7	313.4	143	methyl *tert*-butyl ether	328.4	324.7
71	isopropyl iodide[a]	362.6	364.3	144	1-pentanol[b]	410.9	410.0
72	*n*-propyl iodide	375.6	366.4	145	2-pentanol[a]	392.1	384.5
73	isopropanol	355.4	356.9	146	3-pentanol	388.4	375.8

Table 1. (continued)

no.	compound	exptl NBP (K)	calcd NBP (K)[d]	no.	compound	exptl NBP (K)	calcd NBP (K)[d]
147	1,5-pentanediol	512.2	507.6	213	2-heptanone	424.0	419.2
148	methyl *n*-butyl sulfide	396.6	390.1	214	4-heptanone[b]	417.2	414.2
149	methyl *tert*-butyl sulfide	372.0	381.1	215	1-methylcyclohexanol	441.2	460.4
150	*n*-pentyl mercaptan	399.8	396.7	216	5-methyl-2-hexanone	418.0	415.9
151	hexachlorobenzene	582.6	579.1	217	*n*-butyl propionate	419.8	410.2
152	hexafluorobenzene	353.4	353.9	218	ethyl isovalerate	407.5	407.8
153	1,2,4-trichloro benzene[a]	486.2	497.7	219	*n*-heptanoic acid	496.2	495.0
154	*m*-dichlorobenzene	446.2	456.3	220	*n*-hexyl formate	428.6	433.1
155	*o*-dichlorobenzene[b]	453.6	450.4	221	*n*-propyl *n*-butyrate	416.5	410.2
156	*p*-dichlorobenzene	447.2	457.3	222	1-heptanol	449.5	450.5
157	*m*-chlorophenol	487.0	489.4	223	2-heptanol	432.4	428.4
158	*o*-chlorophenol	447.5	479.9	224	5-methyl-1-hexanol	445.2	450.0
159	*p*-chlorophenol[b]	493.1	488.3	225	*n*-heptyl mercaptan	450.1	450.7
160	phenol	455.0	452.0	226	vanillin	558.0	569.5
161	1,2-benzenediol	518.7	530.5	227	*p*-ethylphenol	491.1	486.3
162	1,3-benzenediol	549.7	550.3	228	2,3-xylenol[b]	490.1	473.9
163	*p*-hydroquinone	558.2	548.0	229	2,4-xylenol	484.1	474.8
164	cyclohexanone	428.9	430.4	230	2,5-xylenol[a]	484.3	475.1
165	ethyl acetoacetate	454.0	428.0	231	2,6-xylenol	474.2	460.2
166	diethyl oxalate[a]	458.9	449.8	232	3,4-xylenol	500.2	483.7
167	cyclohexanol	434.0	459.5	233	3,5-xylenol	494.9	484.8
168	3,3-dimethyl-2-butanone	379.4	386.0	234	diethyl maleate[b]	498.2	493.5
169	ethyl isopropyl ketone	386.5	386.7	235	1-octanal	447.2	453.3
170	1-hexanal	401.5	408.8	236	2-octanone	445.8	441.2
171	2-hexanone	400.9	394.9	237	*n*-butyl *n*-butyrate[b]	438.2	433.3
172	3-hexanone	396.6	389.4	238	*n*-heptyl formate	451.3	455.7
173	methyl isobutyl ketone	389.6	391.2	239	*n*-hexyl acetate	444.7	439.6
174	3-methyl-2-pentanone[a]	390.6	389.9	240	isobutyl isobutyrate	420.6	431.3
175	*n*-butyl acetate	399.1	394.4	241	di-*sec*-butyl ether	394.2	405.3
176	*sec*-butyl acetate	385.1	396.9	242	2-ethyl-1-hexanol	457.8	442.8
177	*tert*-butyl acetate[a]	369.1	379.0	243	1-octanol[a]	468.3	469.1
178	ethyl *n*-butyrate	394.6	387.0	244	2-octanol	453.0	448.9
179	2-ethyl butyric acid	466.9	462.0	245	*n*-octyl mercaptan	472.2	474.0
180	ethyl isobutyrate	383.0	389.4	246	ethyl benzoate[b]	486.6	471.3
181	*n*-hexanoic acid	478.8	478.7	247	benzyl ethyl ether	458.1	463.9
182	isobutyl acetate	389.8	391.5	248	isophorone	488.4	486.5
183	*n*-pentyl formate	405.5	409.3	249	diisobutyl ketone	441.4	451.3
184	*n*-propyl propionate	395.6	387.6	250	2-nonanone	467.5	461.2
185	*n*-butyl ethyl ether	365.3	361.1	251	5-nonanone	461.6	457.8
186	*tert*-butyl ethyl ether	346.0	351.8	252	*n*-heptyl acetate	465.6	461.2
187	diisopropyl ether	341.5	354.3	253	*n*-nonanoic acid[a]	528.8	528.3
188	2-ethyl-1-butanol	419.7	421.8	254	*n*-octyl formate	472.0	477.4
189	1-hexanol	430.6	430.4	255	2,6-dimethyl-4-heptanol	451.0	457.7
190	2-hexanol	413.0	406.5	256	1-nonanol	486.3	487.0
191	2-methyl-1-pentanol[a]	421.2	422.8	257	2-nonanol	471.7	468.7
192	4-methyl-2-pentanol	404.9	419.8	258	*n*-nonyl mercaptan	493.0	495.7
193	methyl *n*-pentyl ether	372.0	362.3	259	dimethyl phthalate	556.8	550.6
194	1,6-hexanediol	516.2	520.3	260	anethole	508.5	481.4
195	hexylene glycol	470.6	466.8	261	*p*-*tert*-butylphenol[a]	512.9	519.4
196	di-*n*-propyl sulfide[b]	416.0	411.2	262	*n*-decanoic acid	543.2	543.5
197	ethyl *tert*-butyl sulfide[a]	393.6	403.5	263	isopentyl isovalerate	467.2	470.7
198	*n*-hexyl mercaptan	425.8	424.0	264	*n*-octyl acetate	484.5	481.1
199	di-*n*-propyl disulfide	469.0	469.2	265	1-decanol	504.1	503.6
200	benzotrichloride[b]	486.7	476.3	266	*n*-decyl mercaptan	512.3	514.5
201	benzyl dichloride[a]	487.0	467.3	267	2-ethylhexyl acrylate	489.2	489.5
202	benzoic acid	522.4	520.2	268	*n*-undecanoic acid	557.4	557.6
203	*p*-hydroxybenzaldehyde	583.2	570.0	269	methyl decanoate	505.0	495.2
204	salicylaldehyde	469.7	485.7	270	1-undecanol	518.2	519.1
205	benzyl chloride	452.6	447.3	271	diethyl phthalate[a]	567.2	567.1
206	benzyl alcohol[b]	477.9	471.5	272	*n*-decyl acetate	517.2	517.8
207	*m*-cresol	475.4	468.5	273	*n*-dodecanoic acid	571.9	569.9
208	*o*-cresol	464.2	453.0	274	di-*n*-hexyl ether	498.9	505.4
209	*p*-cresol	475.1	468.0	275	*n*-tridecanoic acid[b]	585.3	581.5
210	diethyl malonate	472.0	476.2	276	anthraquinone	653.1	652.8
211	diisopropyl ketone	397.6	407.0	277	*p*-*tert*-octylphenol	563.6	575.6
212	1-heptanal	426.0	432.2				

[a] Compound in external prediction set. [b] Compound in cross-validation set. [c] Compound an outlier in N-excluded data subset model. [d] Boiling point calculated using the final neural network model.

example, the dipole moment, considered to be an electronic descriptor, is dependent on geometry. A fourth class of descriptors combines solvent-accessible surface area and atomic charge information to produce charged-partial surface area (CPSA) descriptors.[26] The same information used to develop CPSA descriptors is also used to calculate descriptors that encode hydrogen bonding, such

as the surface area or charges on donatable hydrogen atoms or acceptor N, O, or F atoms.[6]

A total of 124 descriptors were generated for each compound in the data set. The descriptor pool was pruned using a series of objective methods. Any descriptor that contained more than 90% identical values was removed. Since the data set was structurally diverse, it was thought that some descriptors acting as indicator variables (i.e., a high percentage of identical values) might be useful later in the study, hence the high cutoff percentage. The pool of descriptors was then checked for pairwise correlations, and if two descriptors correlated with $r \geq 0.94$ one of them was removed. The descriptor which was correlated with a smaller number of descriptors or was easiest to explain or calculate was retained. These methods were effective in removing 47 descriptors, leaving a reduced descriptor pool with 77 members. To avoid chance correlations, it is recommended that the number of descriptors submitted to regression analysis be less than 60% of the number of observations in the training set.[27] Since there were 300 compounds in the training set, this condition was met, and regression analysis was carried out.

Regression Analysis. Multiple linear regression analysis was used to develop linear models that linked the boiling points to a small subset of descriptors. Since the dependent variable was used to develop the models, this process can be thought of as subjective feature selection. Regression equations have the form

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

where BP_j is the boiling point for compound j, b_0 is the intercept term, and b_i is the coefficient for descriptor X_{ij}. Two automated routines were used to perform subjective feature selection with regression analysis. The first routine used a genetic algorithm (GA) method[28,29] to choose high quality subsets of descriptors from the reduced pool. The GA routine used the root-mean-square (RMS) error to evaluate the quality of the models. Internal statistics, such as individual F values, were not utilized by the GA routine. A second routine employed simulated annealing (SA)[30,31] to choose subsets of descriptors. As in the GA routine, the SA routine evaluated the quality of the models with the RMS error and internal statistics were not used.

Since both methods discussed above utilized only RMS errors to evaluate quality, it was important to examine models suggested by these methods to determine their overall quality. An interactive regression analysis (IRA) routine allowed for this. A pool of descriptors that are of potentially high quality are entered into IRA. Standard regression analysis is used to develop models from the entered pool at the discretion of the user. The number of descriptors in a potential model is also controlled by the user. The routine also provides statistical guidelines, in the form of F-to-enter and T-values. As descriptors are added and removed, the changes in the statistics from model to model can be monitored. Therefore, a more statistically comprehensive evaluation of a given model and the pool of descriptors is achieved.

Neural Networks. Computational neural networks were effective in further minimizing the RMS errors associated with the linear models. Neural networks can provide nonlinear mappings of the descriptors to the boiling point values thus accounting for variability that a linear function can not encode. Their nonlinear nature, coupled with an increase in the number of adjustable parameters, allow neural networks to perform better than linear methods in our studies. A Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton minimization algorithm was employed to train the computational neural network.[32] The theories behind computational neural networks, and the BFGS algorithm have been discussed previously.[33]

RESULTS AND DISCUSSION

Both GA and SA routines were effective in developing an 11 descriptor model that fit the boiling points for the 298-compound training set with an RMS error of 13.5 K. There were two outliers in the 300-compound training set, malononitrile and cyanogen, the only dinitriles present in the data set. The RMS error was relatively high, and examination of the data showed that nitrogen-containing compounds were mostly responsible for this. When the boiling points of the 48 nitrogen-containing compounds were calculated by the model, the RMS error found was 21.3 K. The remaining 250 compounds had an RMS error of just 13.0 K. Thus, the nitrogen-containing compounds were seriously degrading the quality of the model. It is possible that the compounds containing nitrogen were not properly encoded with the 11 descriptors selected and thus were not modeled adequately. Subdividing the data set is an effective method that can be used to investigate this phenomenon. The 27-compound external prediction set had an RMS error of 20.2 K, which provided further evidence that there was a flaw in the model.

As a result of the previous observations, the approach was altered. Since the nitrogen-containing compounds were problematic, the original 327-compound data set was split into a set of 50 N-containing compounds and 277 non-nitrogen containing (N-excluded) compounds. Then models were developed for each subset. It should be noted that the 50-compound N-containing subset was augmented with an additional 54 compounds taken from the literature to bring the N-containing subset total to 104 compounds. This was done because it was necessary to have a data set of at least 100 compounds to insure that a statistically valid model could indeed be developed.

Part I. N-Excluded Data Subset. The 277 compounds in the N-excluded data subset were split randomly into a training set of 250 compounds and an external prediction set of 27 compounds. Table 1 lists the compounds comprising the N-excluded subset. Since descriptors were already calculated for these 277 compounds, feature selection was immediately performed as described previously. A reduced descriptor pool with 77 members was obtained after objective feature selection was completed.

The GA and SA descriptor selection routines each produced the same 10-member subset of descriptors. However, it was apparent that the model was not encoding the data set adequately. The residual values produced by this model were sorted in descending order. After examining the sorted residuals, it was clear that several trends existed that were not encoded. Structural features such as carbonyl groups and fluorine or sulfur atoms were inflating the RMS error, as structures with these features tended to exhibit large positive or negative residuals.

To rectify the problem, seven new descriptors were specifically calculated to encode the structural moieties. Two descriptors, the number of sulfur and fluorine atoms, which were originally removed because they contained greater than 90% identical values, were also added because they can act as indicator variables. The new reduced pool was then submitted to regression analysis.

After repeating the GA and SA feature selection routines, several high quality models were developed using the 250-compound training set. IRA was used to determine which

Table 2. Final Linear Regression Model for the Prediction of Normal Boiling Points for the N-Excluded Subset Taken from the DIPPR Project 801 Database[9]

coeff	sd of coeff	label	descriptor definition
0.3009	0.01105	PPSA	partial positive surface area[a]
−3.690	0.1656	PNSA	partial negative surface area[a]
−51.78	6.394	RPCG	relative positive charge[a]
9.515	0.3571	NRA	number of ring atoms
19.21	0.5395	SQMW	square root of molecular weight
554.7	19.06	SADH	surface area of donatable hydrogens[b]
−25.52	0.9835	NF	number of fluorines[c]
19.52	2.364	KETO	ketone indicator[c]
50.84	4.345	SULF	number of sulfide groups[c]
−135.0	13.55	S/NA	number of sulfurs/total number of atoms[c]
59.86	5.411		Y-intercept

$R = 0.991$; RMS error = 11.6 K; $N = 248$ compds; $F = 1253.7$

[a] See ref 25. [b] See text and ref 6. [c] Descriptor added to pool after preliminary screening (see text).

Table 3. Example Calculation of Normal Boiling Points for Compounds **81** and **138** Using the Linear Regression Model in Table 2

descriptor label	model coeff	value for thiophene (**81**)	value for 2-methyl-2-butanol (**138**)
Y-intercept	59.86	1.0	1.0
PPSA	0.3009	121.3	24.18
PNSA	−3.690	−6.565	−12.52
RPCG	−51.78	0.2517	0.3311
NRA	9.515	5.000	0.0
SQMW	19.21	9.165	9.381
NF	−25.52	0.0	0.0
SADH	554.7	0.0	0.05154
KETO	19.52	0.0	0.0
SULF	50.84	1.0	0.0
S/NA	−135.0	0.2000	0.0
calcd bp:	355.0, 370.5 K		
exptl bp:	357.3, 375.1 K		

model was of highest statistical quality. The best model developed was a 10-variable model with an RMS error of 12.3 K.

Several validation methods were used to examine the stability of the model. The model was first checked for outlier values. The six statistical tests used were the residual, standardized residual, studentized residual, leverage, DFFITS statistic, and Cooks distance.[34] If four or more of these six tests for a given compound exceeded the cutoff value for that test, the compound was considered an outlier. Only two compounds were flagged by the test, diiodomethane and trichloroacetaldehyde. When the model coefficients were recalculated, the RMS error was 11.6 K, a significant improvement. Therefore, the outliers were removed permanently, leaving 248 compounds in the training set. Table 2 shows the final 10-descriptor model and Table 3 provides an example calculation for the compounds thiophene and 2-methyl-2-butanol.

To further validate the model several other tests were performed. The residual values were plotted against the calculated boiling point values, and there was no observable pattern. The largest pairwise correlation coefficient was 0.592 for the 10 descriptors. The variance inflation factor (VIF), defined as $(1 - R^2)^{-1}$, was determined for each descriptor as well. A VIF larger than 10 is indicative of multicolinearity problems.[35] The highest VIF was 2.28 for the model. These tests demonstrate the high internal stability of the model. A plot of calculated vs observed boiling points, given in Figure 1, showed no observable pattern or deviation from normal behavior.

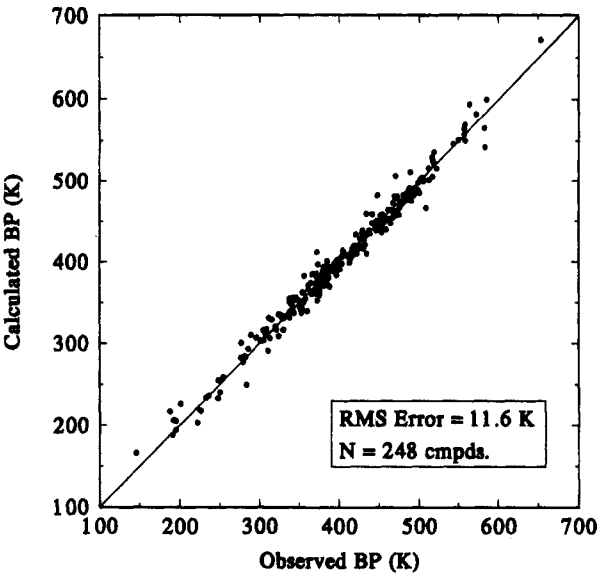

Figure 1. Plot of calculated vs observed boiling points for the training set using the N-excluded linear regression model.

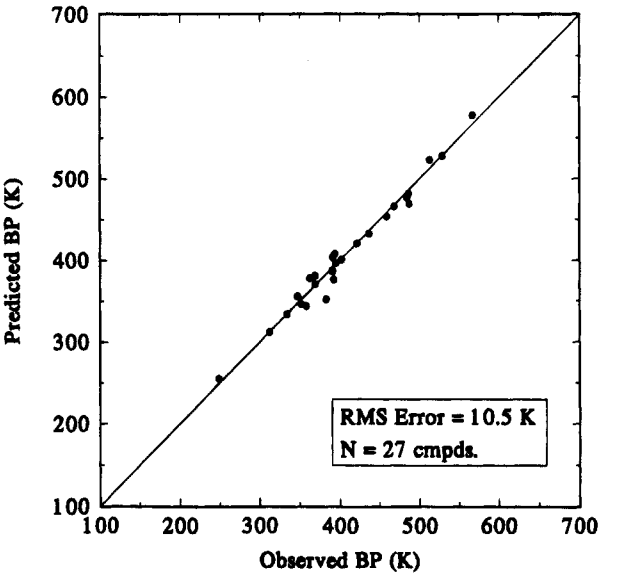

Figure 2. Plot of predicted vs observed boiling points for the external prediction set using the N-excluded linear regression model.

The final validation step for the model involved predicting the boiling points for the 27 compounds in the external prediction set. The RMS error was 10.5 K for the external prediction set, thus completing the final validation step. Figure 2 shows a plot of predicted vs observed boiling points for the prediction set.

The descriptors in the final model cannot be directly linked to boiling point in a causal sense. They are numerical representations of molecular structure that collectively will encode the boiling point with a high degree of accuracy. The descriptors themselves do not cause the boiling point value to change. However, it is reasonable to use the descriptors as an aid in deciphering what structural features are likely to be important. To this end, examination of the descriptors clearly shows that for a diverse group of compounds, indicators of specific substructure features are important. For instance, the number of fluorines has a large negative coefficient, thus suggesting that fluorine atoms will tend to

Table 4. List of Unique Structural Groups and Their Coefficients for the Three Group Contribution Approaches for the N-Excluded Data Subset

group		Joback	regression I	regression II
constant		198.18	232.85	103.17
$-CH_3$		23.58	11.12	3.06
$-CH_2-$	nonring	22.88	21.56	7.85
$-CH_2-$	ring	27.15	30.47	13.71
>CH−	nonring	21.74	25.00	5.09
>CH−	ring	21.78	17.79	−3.15
>C<	nonring	18.25	12.32	−13.43
>C<	ring	21.32	−34.36	−32.51
$=CH_2$		18.18	−15.88	−18.82
=CH−	nonring	24.96	24.42	14.29
=CH−	ring	26.73	22.23	11.02
=C<	nonring	24.14	26.21	4.85
=C<	ring	31.01	29.02	14.49
−F		−0.03	−13.24	−23.54
−Cl		38.13	31.73	6.66
−Br		66.86	57.11	−4.37
−I		93.84	80.00	−31.27
−OH	alcohol	92.88	79.25	74.15
−OH	phenol	76.34	84.69	73.10
−O−	nonring	22.42	20.80	11.73
−O−	ring	31.22	−4.19	6.75
>C=O	nonring	76.75	77.10	54.49
>C=O	ring	94.97	20.78	45.71
−HC=O		72.24	60.03	46.40
−COOH		169.09	141.88	112.28
−COO−	ester	81.10	74.49	37.96
−SH		63.56	64.34	44.63
−S−	nonring	68.78	68.33	41.31
−S−	ring	52.10	40.79	20.38
SQMW				21.05

lower the boiling point of a compound. In fact, the boiling point of fluoromethane is higher than the boiling point of tetrafluoromethane. This trend is not observed for the other halo-methane compounds. Most of the descriptors in the model are fairly straightforward. The three CPSA descriptors encode intermolecular interaction properties, such as dipole−dipole interactions and London dispersion forces. Collectively, the descriptors accurately estimate boiling points for organic compounds that do not contain nitrogen.

Once a final model was determined, it was compared to the Joback group contribution method. In group contribution methods, a compound is broken down into unique substructure groups, and the frequencies of those groups are regressed against the dependent variable of interest, in this case boiling point. Joback used a diverse set of 438 compounds to generate coefficients.[12] In that data set, there were 43 unique substructure groups. A total of 28 groups were unique to the 248 member training set used in this study. Joback's original coefficients for those 28 groups, shown in Table 4, were used to predict the boiling points for the training set. The RMS error was 26.7 K for the training set. This rather large value is not surprising, since the original coefficients were derived from a different set of compounds.

A new set of coefficients was calculated by regressing the group counts against the boiling points for the training set. These coefficients are listed in Table 4 under the regression I heading. The RMS error was 17.6 K, a significant improvement. In a previously published paper, we used the square root of the molecular weight in combination with Joback's groups and developed an excellent model for hydrocarbon boiling points.[10] Taking that approach here, new coefficients were calculated for the 28 groups plus the square root of molecular weight. Table 4 lists the coefficients

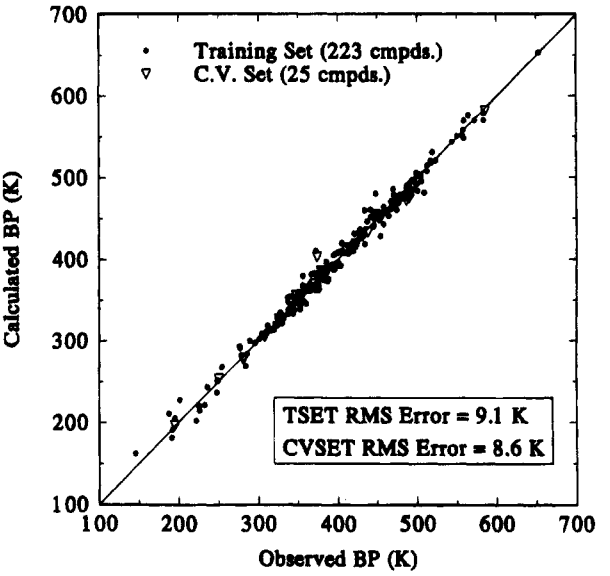

Figure 3. Plot of calculated vs observed boiling points for the N-excluded training and CV sets using the computational neural network model.

under the regression II heading. The RMS error was 14.7 K for the training set with this enhanced group contribution method. The three group contribution methods do not yield results that are comparable to the ADAPT methodology. Additionally, the best linear model from ADAPT has 11 adjustable coefficients, whereas the best group contribution model has 30 adjustable coefficients and performs worse.

The ten descriptors from the final linear model were used to develop a computational neural network model. Neural networks take advantage of nonlinearity to improve the link between the descriptors and boiling points. The 248-compound training set was split randomly into a new training set of 223 compounds and a cross-validation (CV) set of 25 compounds. The original 27-compound external prediction set was used to validate the best network model developed.

Several trial runs with different network architectures were performed to determine the best setup using the program QNET, developed in our laboratory.[32] A 10:6:1 (10 input, 6 hidden, and 1 output neuron) network architecture was chosen. This architecture minimized the number of adjustable parameters without sacrificing network performance. As a general rule, the ratio of observations (223) to adjustable network parameters (73) should be no less than 2.0, and the ratio for this network is much greater.[36]

Once the network architecture was established, an automated version of QNET was used to find an optimal network model. The automated version conducted 500 individual training sessions and recorded the lowest RMS errors for the training and CV set. Since the training algorithm is very sensitive to the starting set of weights and biases, the automated version is used to find quality starting points that lead to low RMS errors. Sessions that provided low RMS errors were examined with more scrutiny.

The session that provided the lowest and best behaved CV set RMS error was chosen as the best network model. This model had an RMS error of 9.1 K for the training set and 8.6 K for the CV set. A plot of calculated vs observed boiling points for the training and CV sets is shown in Figure 3. The external prediction set was used to validate the final network model. The prediction set RMS error was 9.0 K.

Table 5. List of Compounds in the N-Containing Data Subset[a]

no.	compound	exptl NBP (K)	calcd NBP (K)[f]	no.	compound	exptl NBP (K)	calcd NBP (K)[f]
1	hydrogen cyanide	298.8	315.3	53	1-methylpyrrole[d]	386.0	377.8
2	nitromethane	374.4	385.0	54	2-methylpyrrole[d]	420.7	420.0
3	methylamine	266.8	278.3	55	3-methylpyrrole[d]	416.0	422.8
4	acetonitrile[b]	354.8	341.1	56	pyridine[d]	388.4	378.8
5	ethyleneimine	329.0	311.6	57	2-methylpyridine[d]	402.5	406.7
6	nitroethane	387.2	391.9	58	4-methylpyridine[d]	418.5	409.4
7	dimethylamine[a]	280.0	286.6	59	piperidine[d]	379.4	385.3
8	ethylamine	289.7	293.7	60	1-naphthylamine[a,d]	573.8	574.8
9	cyanogen[c]	252.0		61	2-naphthylamine[b,d]	579.3	571.3
10	malononitrile[c]	491.5		62	2-propylamine[a,e]	304.9	314.4
11	acrylonitrile[b]	350.5	359.5	63	2-methyl aniline[e]	473.5	471.4
12	propionitrile	370.5	365.9	64	3-methyl aniline[e]	476.5	479.0
13	1-nitropropane	404.3	401.6	65	4-methyl aniline[e]	473.6	478.6
14	2-nitropropane	393.4	397.2	66	carbazole[d]	627.8	626.3
15	*n*-propylamine	321.0	323.7	67	9-methyl carbazole[d]	616.8	618.2
16	trimethylamine	276.0	294.5	68	1-(2-aminoethyl)piperizine[b]	493.2	488.7
17	*cis*-crotonitrile[a]	380.6	372.7	69	1-(2-aminoethyl)piperidine	459.2	469.0
18	*n*-butyronitrile	390.8	387.9	70	*tert*-pentylamine	350.2	341.4
19	isobutyronitrile	376.8	372.2	71	*o*-methoxyaniline	498.2	486.7
20	pyrrolidine[a]	359.7	356.9	72	*m*-methoxyaniline	524.2	514.6
21	*n*-butylamine	350.6	352.0	73	*p*-methoxyaniline	514.7	503.9
22	*tert*-butylamine	317.5	317.9	74	benzylamine	457.7	447.4
23	diethylamine[a]	328.6	321.3	75	*N*-methylbutylamine	364.2	365.8
24	isobutylamine	340.9	339.6	76	2-methylbutylamine	368.7	371.4
25	*n*-pentylamine[b]	377.6	379.0	77	cyclopentylamine	380.2	386.1
26	*m*-chloroaniline	501.7	509.4	78	*p*-nitrophenol	552.2	547.6
27	*p*-chloroaniline[b]	503.7	510.6	79	allylamine	326.2	317.2
28	aniline	457.2	460.9	80	*N*-allylaniline[b]	492.2	509.4
29	3-methylpyridine[b]	417.3	406.7	81	nitrobutane	425.7	418.4
30	phenylhydrazine	516.7	495.1	82	nitrocyclopentane	453.2	437.7
31	hexanenitrile	436.8	433.0	83	*o*-nitrotoluene[b]	498.2	492.8
32	di-*n*-propylamine	382.0	384.4	84	*m*-nitrotoluene	503.7	507.6
33	3-nitrobenzotrifluoride	475.9	481.3	85	*N,N*-diethylmethylamine	337.2	337.4
34	benzonitrile	464.1	475.0	86	triethylamine	362.0	358.3
35	*p*-nitrotoluene	511.7	507.5	87	*N*-ethylbutylamine	381.2	382.0
36	*n*-heptylamine	430.1	426.8	88	*o*-bromoaniline	502.2	492.2
37	indole	526.1	529.7	89	*m*-bromoaniline	524.2	512.5
38	*n*-octylamine	452.8	448.8	90	*o*-fluoroaniline	455.7	486.3
39	isoquinoline	516.4	515.6	91	*m*-nitroaniline	557.2	560.9
40	quinoline	510.8	510.9	92	*p*-fluorobenzylamine	456.2	487.9
41	*n*-nonylamine	475.4	470.8	93	benzylamine[a]	433.2	448.1
42	tripropylamine[a]	429.7	439.1	94	3,3-dimethylpiperidine	410.2	413.6
43	*n*-decylamine	493.7	492.2	95	*N,N*-dimethyl-1,3-diaminomethane	418.2	410.3
44	diamylamine	476.1	478.7	96	2,2-dimethyl-1,3-diaminomethane	426.2	425.6
45	undecylamine	514.8	514.2	97	1-ethylpiperidine	404.2	405.1
46	*n*-dodecylamine	532.4	535.6	98	2-ethylpiperidine	416.2	417.9
47	tri-*n*-butylamine	487.2	490.6	99	2-ethylpyridine	422.2	434.8
48	acridine	619.2	610.5	100	*N*-methylcyclohexylamine	422.2	421.9
49	*n*-tetradecylamine[a]	564.5	575.2	101	2-aminoheptane	416.2	414.0
50	triamylamine[c]	516.2		102	*N*-*tert*-butylisopropylamine[a]	371.2	384.6
51	2-butylamine[a,d]	335.9	334.6	103	*n*-heptylamine	428.2	426.2
52	pyrrole[d]	402.9	403.1	104	*N*-methylhexylamine[b]	414.2	418.0

[a] Compound in external prediction set. [b] Compound in cross-validation set. [c] Compound an outlier. [d] See ref 37. [e] See ref 38. [f] Boiling point predicted by final neural network model. [g] Compounds and NBP values 1−50 taken from DIPPR Project 801 Database.[9] Compounds and NBP values 68−104 taken from 1994 Aldrich catalog.[39]

Figure 4 displays a plot of predicted vs observed boiling points for the prediction set.

Part II. N-Containing Data Subset. The 50 nitrogen-containing compounds from the DIPPR database were augmented with 54 compounds from three new sources.[37−39] A listing of the 104 compounds is given in Table 5. The same methodology detailed in the N-excluded section was also used here. The data set was split randomly into a 93-compound training set and an 11-compound external prediction set. Descriptors (129) were generated for each compound. Objective descriptor screening reduced the pool to 67 members. Since there were only 93 compounds in the training set, a vector-space descriptor analysis routine[40] was used to further reduce the descriptor pool to about 40 members.

Both the GA and SA feature selection routines were used to find good models. A model with 10 descriptors was found and analyzed in depth with the IRA routine. The RMS error was 16.9 K for the original model. This was due to malononitrile and cyanogen. These two compounds were the only dinitriles, as mentioned previously. They also exhibited very large residuals, so they were removed. The model coefficients were recalculated, and the RMS error was 11.0 K, a significant improvement.

The model was validated with the same methods used for the N-excluded data set. One outlier, triamylamine, was

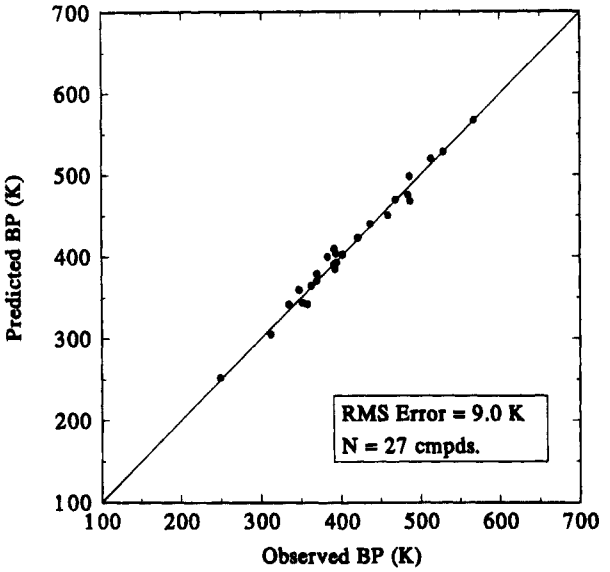

Figure 4. Plot of predicted vs observed boiling points for the N-excluded prediction set using the computational neural network model.

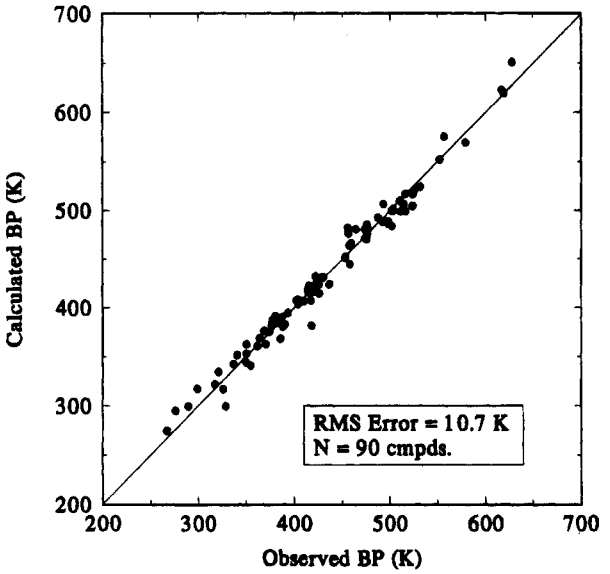

Figure 5. Plot of calculated vs observed boiling points for the training set using the N-containing linear regression model.

removed. The coefficients were recalculated, and the RMS error was 10.7 K for the training set. Table 6 shows the final linear regression model and coefficients. A plot of residual vs calculated boiling points produced no observable

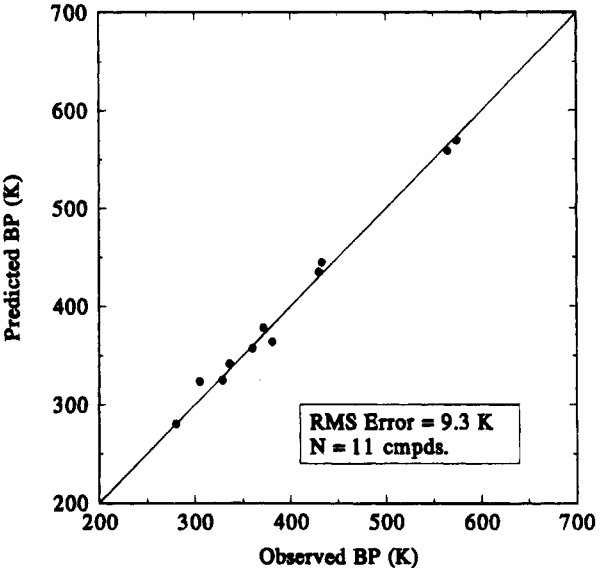

Figure 6. Plot of predicted vs observed boiling points for the prediction set using the N-containing linear regression model.

patterns. The largest pairwise r value was 0.83, and the largest VIF was 9.7, which is lower than the limiting value. Figure 5 shows a plot of calculated vs observed boiling points for the training set. The plot looks smooth, and the data fit the ideal one to one correlation line quite well. The external prediction set RMS error was 9.3 K, thus providing the final validation for the linear model. A plot of predicted vs observed boiling points for the prediction set is depicted in Figure 6.

Examination of the descriptors can lend some insight into what structural features are important to boiling point. However, as mentioned before, the descriptors themselves are not the cause of the normal boiling point for any given compound. For this data set, there are several whole molecule descriptors in the final model. Individual substructure features are not as important. This is not surprising, since the data set is composed of N-containing compounds only. The dipole moment and CPSA descriptors are encoding different aspects of intermolecular interactions. The connectivity indices MOLC 4 and V3C are encoding the degree of branching in the structure. As the branching increases, the values of MOLC 4 and V3C both increase, but the boiling point will generally decrease in value. The connectivity descriptor WTPT 3 is encoding the topological environment of heteroatoms in each structure. As the number of nitrogen atoms increases, the degree of hydrogen

Table 6. Final Linear Regression Model for the Prediction of Normal Boiling Points for the N-Containing Data Subset taken from the DIPPR Project 801 Database[9] and Refs 37−39

coeff	sd of coeff	label	descriptor definition
16.49	2.382	DPOL	dipole moment[a]
0.3673	0.05197	PNSA	partial negative surface area[b]
−346.4	24.23	RNCG	relative negative charge[b]
4.226	0.3317	RNCS	relative negative charged surface area[b]
14.20	0.4873	NAB	number of aromatic bonds
45.63	2.502	MOLC 4	path 2 molecular connectivity index[c]
−58.34	5.569	V3C	cluster 3 valence connectivity index[c]
8.349	1.234	WTPT 3	sum of all path weights from heteroatoms[d]
263.9	27.76	SADH	surface area of donatable hydrogens[e]
144.7	27.04	CHDH	charge of donatable hydrogens[e]
338.2	11.78		Y-intercept

R = 0.990; RMS error = 10.7 K; N = 90 compds; F = 397.8

[a] See ref 25. [b] See ref 26. [c] See ref 19. [d] See ref 18. [e] See text and ref 6.

Table 7. List of Unique Structural Groups and Their Coefficients for the Three Group Contribution Approaches for the N-Containing Data Subset

group		Joback	regression I	regression II
constant		198.18	219.18	134.66
$-CH_3$		23.58	8.79	−9.31
$-CH_2-$	nonring	22.88	21.95	8.75
$-CH_2-$	ring	27.15	16.58	1.89
>CH−	nonring	21.74	26.66	14.79
>CH−	ring	21.78	35.68	23.24
>C<	nonring	18.25	28.20	22.16
>C<	ring	21.32	26.98	23.44
$=CH_2$		18.18	15.02	-10.83
=CH−	nonring	24.96	5.61	3.33
=CH−	ring	26.73	26.16	12.24
=C<	ring	31.01	32.29	22.79
−F		−0.03	−18.52	−37.27
−Cl		38.13	49.43	11.47
−Br		66.86	59.93	−12.46
−OH	phenol	76.34	40.88	22.60
−O−	nonring	22.42	50.31	34.65
$-NH_2$		73.23	64.87	45.71
−NH−	nonring	50.17	69.89	53.66
−NH−	ring	52.82	73.32	55.70
>N−		11.74	43.84	33.79
=N−	ring	57.55	43.28	27.85
−CN		125.66	107.71	82.41
$-NO_2$		152.54	122.92	74.90
SQMW				19.13

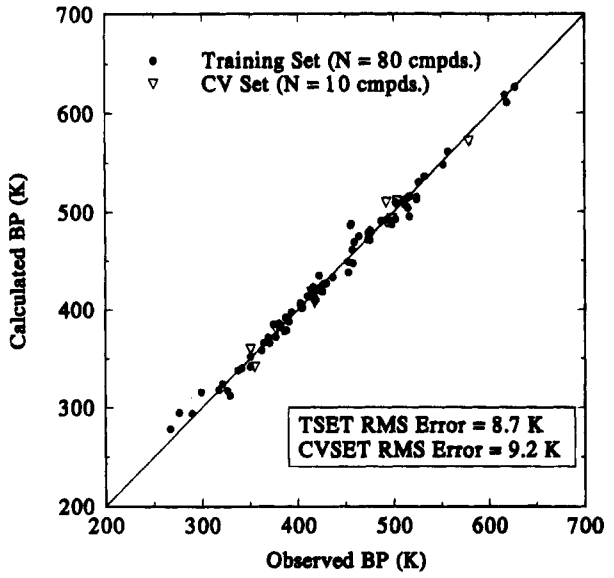

Figure 7. Plot of calculated vs observed boiling points for the N-containing training and CV sets using the computational neural network model.

bonding increases, and the two hydrogen bonding descriptors account for this. The remaining descriptor, NAB, is an indicator of the number of aromatic rings in a given structure. The ten descriptors collectively encode boiling point with a high degree of accuracy for nitrogen-containing organic compounds.

The final model for the N-containing compounds was also compared to Joback's group contribution method in the same manner as described previously. The ADAPT linear models performed better than the three variations of Joback's approach. The RMS error was 23.6 K using Joback's original coefficients. Regressing the groups against the boiling points gave a much improved RMS error of 11.7 K, which is comparable to the ADAPT linear model. Finally, the addition of the square root of molecular weight tightened

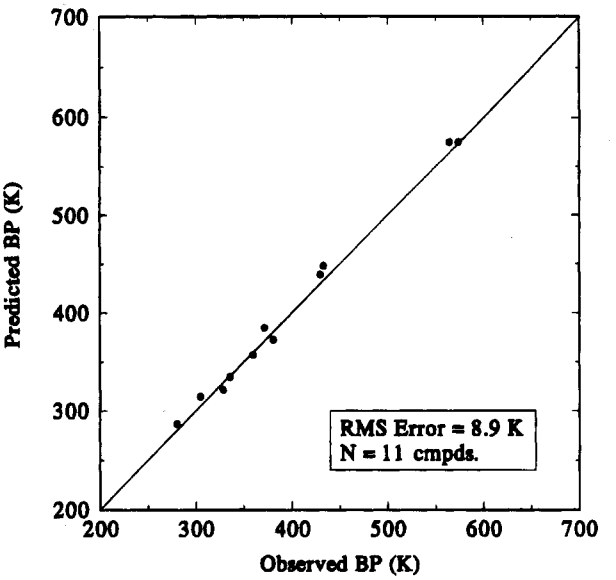

Figure 8. Plot of predicted vs observed boiling points for the N-containing prediction set using the computational neural network model.

the model further, giving an RMS error of 11.1 K, very comparable to the ADAPT model although still not as accurate. The coefficients for each of the three group contribution approaches are listed in Table 7 under the headings Joback, regression I, and regression II, respectively. Due to the small size of the set of compounds used to define these coefficient values, some of the coefficients are dependent on a very small number of occurrences of the corresponding substructure. Therefore, the coefficient values should be used with caution. While the group contribution approach (including the square root term) is comparable to the best linear model, it contains 25 adjustable coefficients as opposed to 11 in the ADAPT model.

Computational neural networks were used to improve the accuracy of the 10-descriptor linear model. Essentially the same approach described above was used here as well. The data set was split randomly into a new 80-compound training set, a 10-compound CV set, and the original 11-compound external prediction set. Since there were 80 compounds in the training set, the number of adjustable network parameters needed to be 40 or less. Therefore, a 10:3:1 network architecture (37 adjustable parameters) was chosen.

After running the automated QNET program, the best set of weights and biases was determined. The RMS error was 8.7 K for the training set and 9.2 K for the CV set. Figure 7 shows a plot of calculated vs observed boiling points for the training and CV sets. The external prediction set gave an RMS error of 8.9 K, thus validating the network model as Figure 8 details.

CONCLUSIONS

Two new models have been developed that predict the normal boiling points for heteroatom-containing organic compounds with a high degree of accuracy. The ADAPT methodology has been shown to provide more accurate calculated boiling points than Joback's group contribution approach. The two models described in this paper can be joined with the model developed in our laboratory for predicting hydrocarbon boiling points to give three methods for predicting the normal boiling points for a wide range of

compounds. The descriptors in each model can also help improve the scientific understanding of the boiling point process. Computational neural networks have proven to be valuable for reducing the RMS errors of both linear models.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

(1) Constantinou, L.; Gani, R. New Group Contribution Method for Estimating Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697.
(2) Wang, S.; Milne, G. W. A.; Klopman, G. Graph Theory and Group Contribution in the Estimation of Boiling Points. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1242.
(3) Stein, S. E.; Brown, R. L. Estimation of Normal Boiling Points from Group Contributions. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 581.
(4) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; American Chemical Society: Washington, DC, 1990.
(5) Fisher, C. H. Boiling Point Gives Critical Temperature. *Chem. Eng.* **1989**, *96*, 157.
(6) Stanton, D. T.; Egolf, L. M.; Jurs, P. C. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306.
(7) 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.
(8) Selover, T. B. DIPPR: Past-Present-Future. *AIChE Symp. Ser.* **1990**, *86*, 90.
(9) Daubert, T. E.; Danner, R. P.; Sibul, M. H.; 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.
(10) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 68.
(11) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(12) 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.
(13) Stewart, J. P. P. MOPAC 6.0, *Quantum Chemistry Program Exchange*; Indiana University, Bloomington, IN, Program 455.
(14) Stewart, J. P. P. Mopac: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.
(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. Struct.-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 Empirical 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) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238.
(28) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267.
(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) Bohachevsky, I. O.; Johnson, M. E.; Stein, M. L. Generalized Simulated Annealing for Function Optimization. *Technometrics* **1986**, *28*, 209.
(31) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77.
(32) 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. Toxicol. Chem.* **1994**, *13*, 841.
(33) 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.
(34) Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; Wiley: New York, 1980.
(35) Stanton, D. T.; Jurs, P. C. Computer-Assisted Prediction of Normal Boiling of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301.
(36) Livingstone, D. J.; Manallack, D. T. Statisitics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295.
(37) Das, A.; Frenkel, M.; Gadalla, N. A. M.; Kudchadker, S.; Marsh, K. N.; Rodgers, A. S.; Wilhoit, R. C. Thermodynamic and Thermophysical Properties of Organic Nitrogen Compounds. Part II. 1- and 2-Butanamine, 2-Methyl-1-Propanamine, 2-Methyl-2-Propanamine, Pyrrole, 1-, 2-, and 3-Methylpyrrole, Pyridine, 2-, 3-, and 4-Methylpyridine, Pyrollidine, Piperidine, Indole, Quinoline, Isoquinoline, Acridine, Carbazole, Phenanthridine, 1- and 2-Nephthalenamine, and 9-Methylcarbazole. *J. Phys. Chem. Ref. Data* **1993**, *22*, 659.
(38) Chao, J.; Gadalla, N. A. M.; Gammon, B. E.; Marsh, K. N.; Rodgers, A. S.; Somayajulu, G. R.; Wilhoit, R. C. Thermodynamic and Thermophysical Properties of Organic Nitrogen Compounds. Part I. Methanamine, Ethanamine, 1- and 2-Propanamine, Benzenamine, 2-, 3-, and 4-Methylbenzenamine. *J. Phys. Chem. Ref. Data* **1990**, *19*, 1547.
(39) Aldrich Catalog/Handbook of Fine Chemicals 1994−1995. Aldrich Chemical Co.: Milwaukee, 1994.
(40) Bradley, G. L. *A Primer of Linear Algebra*; Prentice-Hall, Inc.: New Jersey, 1975.

CI950025V