Graph Theory and Group Contributions in the Estimation of Boiling Points

Shaomeng Wang and G. W. A. Milne*

Laboratory of Medicinal Chemistry, Developmental Therapeutics Program, Division of Cancer Treatment, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892

Gilles Klopman

Department of Chemistry, Case Western Reserve University, Cleveland, Ohio 44106

Received March 8, 1994[⊗]

Estimation of normal boiling points of organic compounds using a group contribution method is known to be unsatisfactory and an attempt has been made to improve the accuracy of the estimations by including chemical graph information in the regressions. This leads to a measurable improvement in the results obtained for both a set of 63 alcohols and also a set of 541 structurally diverse organic compounds. This new approach was found to have better predictive ability than the group contribution approach both in the cross-validation tests and in predicting the normal boiling points for 32 new compounds.

INTRODUCTION

The group contribution method has been widely used, with some success, for the prediction of physicochemical properties of organic compounds. Rekker[1,2] and Hansch[3,4] have independently used the group contribution approach in the development of "fragment addition" methods for calculation of the *n*-octanol/water partition coefficient for organic compounds. This partition coefficient, expressed as its $\log_{10}$ and called the log *P* value, is generally accepted as an indicator of the efficiency with which a chemical is absorbed into biological systems. Several other research groups have since used modifications of the group contribution approach for the calculation of log *P* values.[5−8]

Calculation of aqueous solubility of congeneric organic compounds by means of the group contribution approach has been studied,[9] and recently this approach was used,[10] with reasonable success, for the calculation of the aqueous solubilities of a large set of structurally diverse organic compounds. The approach has also been used[11] successfully to estimate the molar refractivity index of organic compounds. The group contribution effect has been used to estimate many other physical properties, including critical properties,[12,13] enthalpy of formation,[13−15] and entropy of formation.[13,15,16]

The normal boiling point is a particularly important physical property, and its estimation by the group contribution method has been the goal of a number of projects,[16] none of which however have yielded accurate results. Perhaps the best was introduced by Joback[17] who used a learning set of 438 compounds and was able to derive a correlation with a standard deviation of 17.9 K. The same approach was used to predict melting points[17] but produced even more inaccurate results.

In cases where a particular property is not affected significantly by steric factors around the contributing groups, the group contribution method can give fairly good results, but if these steric factors play a significant role, as is often the case in the estimation of boiling points and melting points, then poor results are usually obtained. The molecular topological indices derived from a molecular graph, which is an expression of a compound's structure, have also been used in attempts to correlate the structure with its physicochemical properties. The most successful correlations based on these indices have been obtained from congeneric series of compounds. Thus Randić has shown[18] that his "branching index", an expression of the degree of chain branching in a structure, correlates well with boiling point for alkanes containing fewer than eight carbons, and Kier and Hall have shown[19] that aqueous solubility as well as the normal boiling points of hydrocarbons and of aliphatic alcohols can be correlated with topological indices. They have also shown[20] that a "modified molecular graph index" which accounts for the valency or degree of unsaturation of each atom in the structure is linearly related to the log *P* value for a variety of monofunctional organic compounds, including esters, alcohols, ketones, ethers, carboxylic acids, amines, and hydrocarbons. Nirmalakhandan and Speece attempted to use molecular connectivity indices and polarizability, both calculated solely from the structure, to obtain a correlation between aqueous solubility and structure for a large set of organic structures.[21−23]

The main advantage of the group contribution approach is that it can accommodate many classes of chemicals and many different functional groups, because it treats them empirically. The method can give very good results, provided that the parameters associated with each group are well defined, the groups are independent of one another, and each group is free of unusual steric effects, such as crowding. When these conditions are met, the estimated values of physicochemical properties are often acceptable for practical applications and examples of such acceptable estimates include molar refractivity, surface area, aqueous solubility, and log *P* values of organic compounds. In the case of melting points and boiling points of organic compounds however, the group contribution method often gives disappointing results and various observations, such as those published by Randić,[18] suggest that this may be because steric effects are not accounted for.

The "molecular topological" approaches, which are based on graph theory, can give reliable predictions of some

[⊗] Abstract published in *Advance ACS Abstracts*, October 1, 1994.

0095-2338/94/1634-1242$04.50/0

Table 1. Group Parameters Used in Boiling Point Estimation

no.	groups	no. of compds	occurrence freq.	comments
1	$-CH_3$	408	835	
2	$-CH_2-$	295	1046	
3	−CH−	104	121	
4	−C−	60	95	
5	$=CH_2$	52	58	
6	=CH−	67	85	
7	=C−	54	58	
8	=C=	5	7	
9	HC≡	6	7	
10	≡C−	29	41	
11	$-CH_2-$	64	283	in ring
12	−CH−	40	54	in ring
13	−C−	7	17	in ring
14	=CH−	133	613	in ring
15	=C−	132	246	in ring
16	−F	22	131	connected to sp^3 carbon
17	−F	7	12	connected to non-sp^3 carbon
18	−Cl	31	57	connected to sp^3 carbon
19	−Cl	15	21	connected to non-sp^3 carbon
20	−Br	17	19	connected to sp^3 carbon
21	−Br	15	16	connected to non-sp^3 carbon
22	−I	6	6	connected to sp^3 carbon
23	−I	8	8	connected to non-sp^3 carbon
24	−OH	26	30	primary alcohol
25	−OH	5	5	secondary alcohol
26	−OH	12	12	tertiary alcohol
27	−OH	24	24	connected to non-sp^3 carbon, not COOH
28	−O−	45	49	
29	−O−	7	8	in ring
30	−CHO	13	14	aldehyde
31	X-COOH	10	10	X, a non-carbon atom
32	−COO−	22	25	ester
33	−CO−	51	53	acyclic carbonyl
34	−CO−	6	7	cyclic carbonyl
35	$-NH_2$	8	10	attach to a SP3 carbon
36	−NH2	8	8	attach to any atom except a sp3 carbon
37	Y-NH-X	4	4	X, Y both sp^3 carbon
38	Y-NH-X	6	6	either X or Y is a sp^3 carbon
39	−NH−	5	5	cyclic secondary amine
40	X-N-Y(-Z)	8	8	X, Y, Z all sp^3 carbon
41	X-N-Y(-Z)	6	6	at least one of X, Y, Z not a sp^3 carbon
42	X-C≡N	9	10	X sp^3 carbon
43	X−C≡N	6	6	X, not a sp^3 carbon
44	−N=	6	6	in ring
45	$-NO_2$	12	12	
46	−SH	15	15	
47	−S−	15	17	
48	X-CF3	7	13	X, a sp^3 carbon
49	X-CF3	3	3	X, not a sp^3 carbon

physical properties, particularly for families of organic compounds containing a single functional group, for example, alcohols. The use of a single molecular connectivity index to correlate properties and structure for polyfunctional, structurally diverse compounds has been much less successful, and it might be inferred that molecular connectivity indices carry useful information about steric factors in molecules but insufficient information for properties prediction.

The group contribution approach and the molecular topological approach may therefore by complementary, and, if this is the case, their combination might lead to a method for more accurate prediction of properties of organic structures, especially in the area of boiling points and melting points, where molecular shape is known to play some role. Each group which is defined for purposes of the group contribution calculations can also be assigned a local graph index, which characterizes its steric environment. The same functional group, found in two different steric environments, will thus be characterized by different graph index values. The development and use of local graph indices, either electronic or geometric in nature, in QSAR/QSPR studies have been the focus of a number of papers in recent years,[24-28] but in general, the use of local graph indices has not been extensive, perhaps because most of the molecular topological indices published since 1974[18] have been "global",[29] *i.e.*, descriptive of the whole molecule, rather than some part of it. We have combined the group contribiuon effect with a local topological index approach to the calculation of normal boiling points for a large set of structurally diverse organic compounds, and the results of this work form the subject of this paper.

Table 2. Group Parameters Defined but Absent from the Database

no.	group	no. of compds	occurrence freq.	comments
50	=C=	0	0	in a ring
51	〉C−	0	0	in a ring
52	X-OH	0	0	X=N, O, P, S
53	S	0	0	non sp^3, not in SO, SO_2
54	−SO−	0	0	not in SO_2
55	SO_2	0	0	
56	HN=	0	0	
57	−N=	0	0	
58	−N−(−)	0	0	in a ring
59	N=O	0	0	not in NO_2
60	P	0	0	sp^3 phosphorus
61	P	0	0	non-sp^3 phosphorus, not in S=P, P=O
62	−PO−	0	0	
63	=C=O	0	0	
64	CONH2	0	0	
65	CONH	0	0	
66	CON	0	0	
67	CON=	0	0	
68	S=P	0	0	

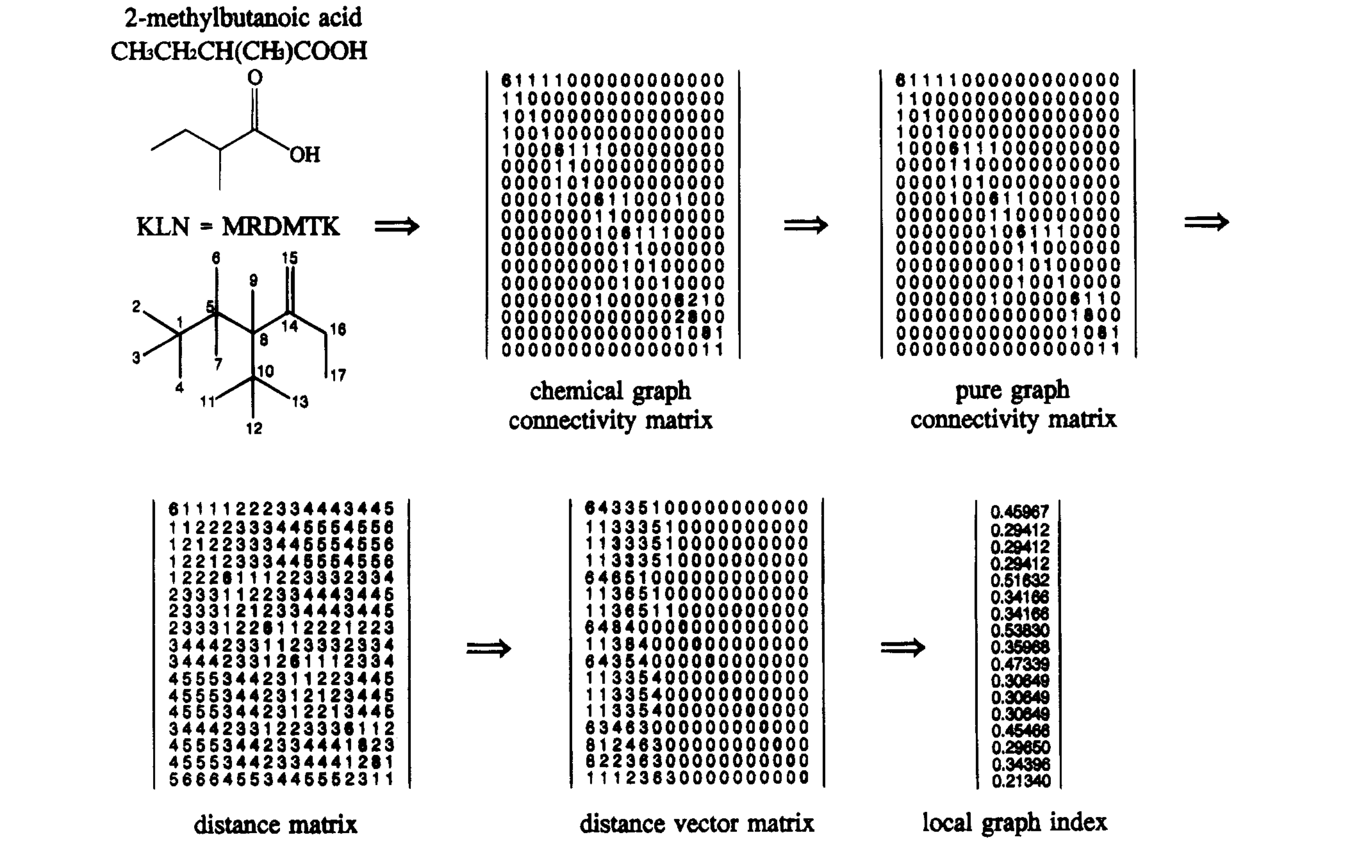
METHODS AND RESULTS

1. Group Parameters. A total of 68 chemical groups were defined. Of these, the 49 shown in Table 1 were represented in the large 541-compound database studied, and their occurrence frequencies are also given in the Table 1. A further 19 groups were defined and are given in Table 2 but were not represented in this database.

The groups that were defined consist of most of the functional groups encountered commonly in organic chemistry, together with a set of atom-centered fragments characterized by the hybridization state of the central atom, the number of hydrogen atoms attached to it, and its presence or otherwise in a ring.

The identification of these groups in the compounds of the database was carried out algorithmically. Chemical structures, in KLN code format,[30] are provided to the program, which converts the KLN code to a connection table. A ring search routine determines whether each non-hydrogen atom is part of a ring system and each group is then identified in terms of the information in Tables 1 and 2.

2. Local Graph Indices. Local graph indices are defined relative to certain atoms in a molecule rather than to the entire molecule. A graph G is undirected and unweighted and has no multiple edges between vertices (nodes) and no self-loops



```
61111000000000000
11000000000000000
10100000000000000
10010000000000000
10006111000000000
00001100000000000
00001010000000000
00001006110001000
00000001100000000
00000001061110000
00000000011000000
00000000010100000
00000000010010000
00000001000006210
00000000000002800
00000000000001081
00000000000000011
```

chemical graph connectivity matrix

```
61111000000000000
11000000000000000
10100000000000000
10010000000000000
10006111000000000
00001100000000000
00001010000000000
00001006110001000
00000001100000000
00000001061110000
00000000011000000
00000000010100000
00000000010010000
00000001000006110
00000000000001800
00000000000001081
00000000000000011
```

pure graph connectivity matrix

```
61111222334443445
11222333445554556
12122333445554556
12212333445554556
12226111223332334
23331122334443445
23331212334443445
23331226112221223
34442331123332334
34442331261112334
45553442311223445
45553442312123445
45553442312213445
34442331223336112
45553442334441823
45553442334441281
56664553445552311
```

distance matrix

```
64335100000000000
11333510000000000
11333510000000000
11333510000000000
64651000000000000
11365100000000000
11365110000000000
64840000000000000
11384000000000000
64354000000000000
11335400000000000
11335400000000000
11335400000000000
63463000000000000
81246300000000000
82236300000000000
11123630000000000
```

distance vector matrix

```
0.45967
0.29412
0.29412
0.29412
0.51632
0.34166
0.34166
0.53830
0.35968
0.47339
0.30649
0.30649
0.30649
0.45466
0.29650
0.34396
0.21340
```

local graph index

Figure 1. Matrices derived from the structure of 2-methylbutanoic acid.

at any vertex. If u and v are two vertices in the graph G, then the path p(uv) between *u* and *v* will be one of possibly several pathways, sequences of edges and nodes. All the nodes and hence the edges are distinct, and if there are *k* edges in p(*uv*) then the path is said to be of length *k*. The distance d(*uv*) is the shortest path between *u* and *v*.

If a vertex *u* is at a distance *k* from the vertex *v*, then *u* will have more effect upon *v* than a vertex *w* whose distance from v is $k + 1$. This principle, which merely states that as the number of bonds separating two atoms increases, the interaction between the atoms will decrease, underlies our definition of local graph indices. Four formulas (1−4) for the calculation of local graph indices were designed and subsequently evaluated to determine which is the best for boiling point estimation. For an atom in a molecule, i.e., a

$$\gamma(v_i) = \log_{10}\left(\sum_{j=1}^{m} (n_{ij})2^{-j}\right) \tag{1}$$

$$\gamma(v_i) = \log_{10}\left(\sum_{j=1}^{m} (n_{ij})4^{-j}\right) \tag{2}$$

$$\gamma(v_i) = \log_{10}\left(\sum_{j=1}^{m} (n_{ij})/(j+1)\right) \tag{3}$$

$$\gamma(v_i) = \log_{10}\left(\sum_{j=1}^{m} (\mathrm{n}_{ij})/(\mathrm{j}+1)^{3/2}\right) \tag{4}$$

node v_i in a graph of *m* nodes, there are n_{ij} vertices at a distance *j* from vertex v_i and any one of the above equations can be applied. In eq 1−4, $\gamma(v_i)$ is the local index related to the vertex *i*; n_{ij} is the number of vertices at a distance *j* from the node *i*; the *n*(*ij*) values are computed from the molecular graph with hydrogens preserved or suppressed.

Computation of these graph indices is done by means of a program which, for 2-methylbutanoic acid, proceeds through the steps illustrated in Figure 1. First, the chemical graph connectivity matrix is derived from the KLN code.[30] In this, and the subsequent matrices, the diagonal elements (boldfaced) represent the atoms of the molecule, appearing in any order and each identified by its atomic number. The first atom ("6") is therefore a carbon, actually C_1, which is in the terminal methyl group of the structure. This is followed by the three hydrogens attached to it and then by C_5. The off-diagonal elements represent the connections between pairs of atoms. Thus C_1 and the three hydrogens associated with it are connected by single bonds ("1"). The only nonsingle bond is in the carbonyl group, and this is found at row 15, column 15 ($i = j = 15$). The carbon ($n_{14,14} = 6$) is attached to the oxygen ($n_{15,15} = 8$) with a double bond, indicated by the 2 at $n_{14,15}$ and $n_{15,14}$.

This chemical graph connectivity matrix is converted to the pure graph connectivity matrix by simply suppressing all bond order information. Only the presence or absence of a bond is indicated in the off-diagonal elements $n_{i,j}$ ($i \neq j$) which may be 1 (atoms *i* and *j* connected) or 0 (*i* and *j* not connected).

This pure graph matrix is next transformed to the distance matrix. Here, the atoms are still identified by their atomic numbers in the diagonal elements, while the off-diagonal element $n_{i,j}$ represents the shortest path between *i* and *j*. For example, $n_1 = 6$ (again, C_1), $n_{16} = 8$ (an oxygen, O_{16}) and $n_{1,16} = 4$. This means that the shortest path between atom

1 (C_1) and atom 16 (O_{16}) involves four bonds—a fact that can be easily verified by inspection of the structure.

The distance matrix is now transformed to a distance vector matrix. The numbers in the first column of the distance vector matrix are the atomic numbers which were in the diagonal of the previous matrices. The entries in columns 2−17 of the distance vector matrix are the total number of atoms which are $(i - 1)$ bonds away from the atom in column 1 of the same row. Thus for the first row in this matrix, a carbon atom (C_1) is one bond away from four other atoms (three hydrogens and one carbon), two bonds away from three atoms, three bonds from three atoms, four bonds away from five atoms, and so on.

The distance vector matrix may be computed with all the hydrogens accounted for (hydrogen-preserved) or omitted (hydrogen-suppressed) and in either case is the source of the local graph indices that are used in this study. As an example, if eq 4 is to be used with the hydrogen-preserved distance vector matrix shown in Figure 1, then the local graph index for C_1 will be

$$\log_{10}\left(\frac{4}{(1+1)^{3/2}} + \frac{3}{(1+2)^{3/2}} + \frac{3}{(1+3)^{3/2}} + \frac{5}{(1+4)^{3/2}} + \frac{1}{(1+5)^{3/2}}\right) = \log_{10}(2.88184) = 0.459\,67$$

which is the first entry in the local graph index matrix. The local graphs indices computed with eq 1−4 define the degree of crowding (the steric environment) of an atom in a molecule expressed as a chemical graph. The differences between these indices arise from the different weights assigned to atoms in different layers of the graph with respect to a central atom.

The normal boiling point (bp) of a compound can be calculated using the group contribution effect with the help of eq 5

$$\mathrm{bp} = C_0 + \sum_{i=1}^{N}(C_iP_i) \qquad (5)$$

where bp represents the normal boiling point (in °C, the scale used in eq 8−25, C_0 is a constant derived from a regression analysis, N is the total number of group parameters defined, C_i is the coefficient of the ith group parameter and P_i is the number of occurrences of the ith group in the structure. If the group contribution effect is now combined with the local graph indices, eq 5 becomes

$$\mathrm{bp} = C_0 + \sum_{i=1}^{N}(C_iP_i(1 - d_i\bar{\gamma}_i)) \qquad (6)$$

where $\bar{\gamma}_i$ is the average of the local graph indices related to the ith parameter and d_i is the coefficient of $\bar{\gamma}_i$. The other terms have the same meaning as in eq 5.

3. Regression for 63 Aliphatic Alcohols. The procedure for calculation of normal boiling points is illustrated with the set of 63 aliphatic alcohols that was previously used in Kier and Hall in their study[19] which resulted in a correlation between the normal boiling point and a graph index. The simplest group parameters can be defined as the numbers of hydrogens, carbons, and oxygens in a molecule. Equation 7 relating these parameters to the normal boiling point is obtained:

$$\mathrm{bp} = -5.88(\mathrm{H}) + 29.19(\mathrm{C}) + 44.48(\mathrm{O}) \qquad (7)$$

For this regression, $F = 171.26$, $R^2 = 0.851$, and SD = 13.04 °C. The molecular weight of a compound is highly correlated with its boiling point. For the set of 63 compounds in Table 3, for example, the boiling point is correlated with the square root of the molecular weight, with an R^2 of 0.81 and an SD = 14.8 °C. The square root of the molecular weight is used because the normal boiling points are known to increase more slowly than the molecular weight increases. If the three parameters of eq 7 are now combined with the molecular weight (M), eq 8 results:

$$\mathrm{bp} = -4.78(\mathrm{H}) + 52.85(\mathrm{C}) + 263.32(\mathrm{O}) - 37.51(M)^{1/2} \qquad (8)$$

with $F = 124.06$, $R^2 = 0.863$, and SD = 12.49 °C. This is marginally better than eq 7 but does not represent a significant improvement. It does suggest, however, that there is a correlation between the molecular weight and the parameters of eq 7. Examination of the 63 structures in this database (Table 3) reveals that all the carbons and oxygens in these compounds are sp^3 hybridized, but the carbons can be further categorized as ring carbon atoms (CR) or chain carbon atoms (CC). The three parameters of eq 7 can thus be extended to four parameters as in eq 9:

$$\mathrm{bp} = 8.17(\mathrm{H}) + 1.04(\mathrm{CR}) + 5.73(\mathrm{CC}) + 16.71(\mathrm{O}) \qquad (9)$$

and this gives $F = 112.83$, $R^2 = 0.852$, and SD = 13.01 °C, which represents no improvement over eq 7.

With these unsatisfactory results as background, an attempt was made to include a graph index in the regression. Equation 1 with the hydrogen-suppressed graph was used to calculate graph indices for C, H, and O. The molecular weight was also included in the regression and

$$\mathrm{bp} = -44.12(\mathrm{H}) + 127.46\gamma(\mathrm{H}) + 122.11(\mathrm{C}) - 169.51\gamma(\mathrm{C}) + 214.63(\mathrm{O}) - 163.26\gamma(\mathrm{O}) - 10.11(M)^{1/2} \qquad (10)$$

eq 10 was arrived at with $F = 461.88$, $R^2 = 0.980$, and SD = 4.75 °C—a significant improvement over the preceding eqs 7−9 with the R^2 increasing from 0.86 to 0.98 and SD reduced from 12.5 to 4.75 °C. This suggests that the steric environment of each group—in this case, each atom—does indeed have an effect upon the boiling point.

To find a formula which includes graph indices and yields an optimum correlation with the normal boiling points, the local graph indices derived from eqs 1−4 using either hydrogen-suppressed or hydrogen-preserved graphs were evaluated in regressions. The results are shown in eqs 10−17 of Table 4. As can be seen from this table, the graph indices derived from eq 4 with suppressed hydrogens (eq 17) yields the best results. The R^2 is increased to 0.988, and the standard deviation is reduced from 12.5 to 3.7 °C. Equations 10−17 are all improved by the inclusion of the local graph indices, and an interesting point is that the hydrogen-suppressed graphs consistently yield better results than their hydrogen-preserved counterparts.

When the four group parameters of eq 9 are included in a correlation with graph indices derived from eqs 1−4 and

Table 3. Boiling Points (°C) Measured Experimentally and Calculated with Eq 17

no.	compd	B_{calc}	B_{exp}	Δ
1[a]	1-butanol	114.75	117.70	−2.95
2[a]	2-methylpropanol	106.18	107.90	−1.72
3	2-butanol	101.18	99.50	1.68
4[a]	1-pentanol	134.50	137.80	−3.30
5[a]	3-methylbutanol	127.92	131.20	−3.28
6[a]	2-methylbutanol	124.50	128.70	−4.20
7	2-pentanol	120.30	119.00	1.30
8	3-pentanol	117.02	115.30	1.72
9	3-methyl-2-butanol	112.69	111.50	1.19
10[a]	2-methyl-2-butanol	105.48	102.00	3.48
11[a]	2,2-dimethyl-1-propanol	114.48	113.10	1.38
12[a]	1-hexanol	154.44	157.00	−2.56
13	2-hexanol	140.13	139.90	0.23
14	3-hexanol	135.58	135.40	0.18
15	3-methyl-3-pentanol	122.26	122.40	−0.14
16	2-methyl-2–pentanol	125.31	121.40	3.91
17	2-methyl-3-pentanol	128.98	126.50	2.48
18	3-methyl-2-pentanol	131.57	134.20	−2.63
19	2,3-dimethyl-2-butanol	118.79	118.60	0.19
20	3,3-dimethylbutanol	139.25	143.00	−2.75
21	3,3-dimethyl-2-butanol	122.89	120.00	2.89
22	4-methyl-pentanol	149.26	151.80	−2.54
23	4-methyl-2-pentanol	134.67	131.70	2.97
24	2-ethylbutanol	142.33	146.50	−4.17
25[a]	cyclohexanol	163.70	161.00	2.70
26[a]	1-heptanol	174.40	176.30	−1.90
27[a]	2-methyl-2-hexanol	145.78	142.50	3.28
28[a]	3-methyl-3-hexanol	141.99	142.40	−0.41
29	3-ethyl-3-pentanol	139.44	142.50	−3.06
30	2,3-dimethyl-2-pentanol	138.85	139.70	−0.85
31[a]	2,3-dimethyl-3-pentanol	136.41	139.00	−2.59
32[a]	2,4-dimethyl-2-pentanol	141.39	133.00	8.39
33	2,4-dimethyl-3-pentanol	142.37	138.80	3.57
34	2,2-dimethyl-3-pentanol	140.31	136.00	4.31
35	3-heptanol	155.02	156.80	−1.78
36	4-heptanol	153.66	155.00	−1.34
37[a]	1–octanol	194.32	195.20	−0.88
38	2,2,3-trimethyl-3-pentanol	150.46	152.50	−2.04
39	2-octanol	180.19	179.80	0.39
40[a]	2-ethyl-hexanol	182.27	184.60	−2.33
41	1-nonanol	214.15	213.10	1.05
42	2-nonanol	200.16	198.50	1.66
43	3-nonanol	194.57	194.70	−0.13
44	4-nonanol	192.17	193.00	−0.83
45	5–nonanol	191.49	195.10	−3.61
46	2,6-dimethyl-3-heptanol	184.56	178.00	6.56
47	3,5-dimethyl-4-heptanol	181.90	187.00	−5.10
48	1,1-dimethylpentanol	180.24	192.00	−11.76
49	7-methyloctanol	211.18	206.00	5.18
50	3,5,5-trimethylhexanol	201.44	193.00	8.44
51[a]	1-decanol	233.88	230.20	3.68
52	cyclopentanol	148.20	140.85	7.35
53	cycloheptanol	182.59	185.00	−2.41
54	1-ethylcyclohexanol	169.99	166.00	3.99
55	2-ethylcyclohexanol	177.73	181.00	−3.27
56	1-methyl-cyclohexanol	160.11	155.00	5.11
57	2-methyl-cyclohexanol	166.01	165.00	1.11
58	3-methyl-cyclohexanol	168.45	174.50	−6.05
59	4-methyl-cyclohexanol	169.71	173.50	−3.79
60	1,3,5-trimethylcyclohexanol	176.37	181.00	−4.63
61[a]	ethanol	78.01	78.50	−0.49
62[a]	2-propanol	82.45	82.40	0.05
63[a]	1-propanol	95.58	97.40	−1.82

[a] Compounds also present in the 541 learning database.

Table 4. Results from Eqs 10−17

eq	F	R^2	S (°C)	g. i. eq	H preserved (P) or suppressed (S)
10	461.9	0.980	4.75	1	P
11	513.8	0.982	4.51	1	S
12	543.7	0.983	4.39	2	P
13	580.7	0.984	4.25	2	S
14	361.5	0.975	5.36	3	P
15	533.3	0.983	4.43	3	S
16	413.7	0.978	5.02	4	P
17	754.5	0.988	3.73	4	S

Table 5. Results from Eqs 18−25

eq	F	R^2	S (°C)	g. i. eq	H preserved (P) or suppressed (S)
18	368.1	0.982	4.53	1	P
19	396.3	0.983	4.37	1	S
20	559.6	0.986	4.06	2	P
21	506.7	0.987	3.87	2	S
22	361.0	0.982	4.58	3	P
23	619.3	0.989	3.51	3	S
24	434.8	0.985	4.18	4	P
25	772.1	0.991	3.14	4	S

the molecular weight term, eqs 18−25 in Table 5 are obtained. Equations 23 and 25 are somewhat improved compared to eq 17, and the standard deviation of eq 25 is reduced to approximately 3 °C. These improvements show that ring carbons and chain carbons make slightly different contributions to the boiling point. The 63 alcohols in the database are given in Table 3, along with the boiling points estimated with eq 17.

4. Models with 541 Diverse Compounds. The next step was to expand the database, abandoning the restriction to alcohols. Joback[17] has compiled a dataset of 438 organic compounds with normal boiling points, and to this was added a further 103 compounds from the CRC Handbook of Chemistry and Physics,[31] for a total of 541 compounds. These 541 compounds are listed in Appendix A. Of the 68 group parameters defined previously (see Tables 1 and 2), 49 have at least four occurrences in this dataset. A correlation was obtained between these 49 parameters (P_1−P_{49}) and the normal boiling points, and the correlation eq 26 was obtained. In eqs 26−35, bp denotes the normal boiling point in K. For eq 26, n = 541, F = 241.28, R^2 = 0.960, and SD = 17.07 K.

$$\begin{aligned} \mathrm{bp} = {} & 222.26 + 15.70P_1 + 20.98P_2 + 26.92P_3 + 28.82P_4 + 7.14P_5 + 23.76P_6 + 29.17P_7 + 17.94P_8 - \\ & 13.62P_9 + 34.48P_{10} + 22.29P_{11} + 26.63P_{12} + 32.57P_{13} + 23.13P_{14} + 31.93P_{15} - 8.42P_{16} - \\ & 9.81P_{17} - 28.94P_{18} - 25.45P_{19} + 69.67P_{20} + 52.14P_{21} + 97.83P_{22} - 91.12P_{23} + 85.66P_{24} + \\ & 81.88P_{25} + 46.78P_{26} + 74.10P_{27} + 20.48P_{28} + 24.33P_{29} - 0.15P_{30} + 6.32P_{31} - 11.83P_{32} + \\ & 36.50P_{33} + 63.44P_{34} + 54.30P_{35} + 79.03P_{36} + 39.42P_{37} + 63.52P_{38} + 62.17P_{39} + 11.88P_{40} + \\ & 42.05P_{41} + 86.19P_{42} - 55.56P_{43} + 53.49P_{44} - 114.29P_{45} + 64.45P_{46} + 61.49P_{47} + 22.38P_{48} + \\ & 53.14P_{49} \end{aligned} \quad (26)$$

Equation 26 shows that although there is a high degree of correlation between the normal boiling points and the group parameters for these compounds, the accuracy is relatively poor, and further refinements are clearly necessary if satisfactory results are to be obtained. The molecular weight, as shown earlier, is often found to be correlated with the normal boiling point.

$$\begin{aligned}\mathrm{bp} = {}& 83.06 - 6.78P_1 + 4.74P_2 + 13.42P_3 + 20.23P_4 - 10.11P_5 + 7.34P_6 + 18.85P_7 + 10.86P_8 - 13.81P_9 + 17.50P_{10} + 4{:}46P_{11} + 10.69P_{12} + 19.53P_{13} + \\ & 5.79P_{14} + 19.74P_{15} - 28.98P_{16} - 36.57P_{17} - 19.68P_{18} - 20.99P_{19} - 33.84P_{20} - 42.00P_{21} - 58.18P_{22} - 53.45P_{23} + 62.64P_{24} + 59.30P_{25} + \\ & 20.90P_{26} + 47.43P_{27} - 0.13P_{28} + 5.27P_{29} + 10.03P_{30} + 9.35P_{31} - 10.31P_{32} + 12.05P_{33} + 36.11P_{34} + 40.29P_{35} + 54.09P_{36} + 23.01P_{37} + \\ & 43.08P_{38} + 45.66P_{39} + 6.11P_{40} + 28.10P_{41} + 69.12P_{42} + 33.42P_{43} + 33.85P_{44} - 53.58P_{45} + 18.78P_{46} + 19.77P_{47} + 20.41P_{48} - 74.81P_{49} + \\ & 27.18(M)^{1/2} \quad (27)\end{aligned}$$

Inclusion of a molecular weight term in eq 26 yields eq 27 which gave $n = 541$, $F = 293.79$, $R^2 = 0.968$, and SD $= 15.36$ K. Inclusion of a molecular weight term does indeed increase the retro-fit accuracy of the boiling points, but the standard deviation, 15.36 K, is still too large.

Table 6. Results from Eqs 28–35

eq	F	R^2	S (°C)	g. i. eq	H preserved (P) or suppressed (S)
28	409.6	0.989	8.87	1	P
29	350.5	0.988	9.58	1	S
30	487.3	0.991	8.14	2	P
31	293.6	0.985	10.45	2	S
32	564.2	0.992	7.57	3	P
33	487.3	0.991	8.14	3	S
34	558.6	0.992	7.61	4	P
35	474.2	0.991	8.25	4	S

Table 7. Significance of Parameters in Eqs 27 and 32

parameter	t(27)	t(32)	parameter	t(27)	t(32)	parameter	t(32)
constant	5.75	−1.68	P_{34}	4.70	1.90	g_{18}	0.31
P_1	−1.86	1.70	P_{35}	7.24	6.22	g_{19}	1.92
P_2	3.09	2.57	P_{36}	7.65	1.67	g_{20}	−0.30
P_3	3.67	−1.80	P_{37}	2.73	3.91	g_{21}	2.12
P_4	3.14	−5.36	P_{38}	6.04	3.55	g_{22}	1.12
P_5	−2.40	0.39	P_{39}	5.69	7.87	g_{23}	0.87
P_6	2.84	2.65	P_{40}	0.86	3.11	g_{24}	−13.19
P_7	4.33	4.70	P_{41}	3.64	3.47	g_{25}	−3.17
P_8	1.43	2.53	P_{42}	11.22	10.50	g_{26}	−7.07
P_9	−2.18	1.10	P_{43}	4.22	4.79	g_{27}	−4.86
P_{10}	5.95	4.78	P_{44}	4.74	2.72	g_{28}	−0.83
P_{11}	2.26	1.84	P_{45}	6.63	5.75	g_{29}	−1.19
P_{12}	3.52	1.85	P_{46}	2.74	2.35	g_{30}	−3.00
P_{13}	3.25	−1.04	P_{47}	3.62	0.12	g_{31}	−2.17
P_{14}	2.89	2.22	P_{48}	1.92	0.79	g_{32}	−1.69
P_{15}	7.12	3.82	P_{49}	−6.81	−2.50	g_{33}	0.26
$P31_6$	−7.46	−2.34	(MW)$^{0.5}$	10.74	9.81	g_{34}	−0.84
P_{17}	−7.64	0.12	g_1		−2.91	g_{35}	−4.40
P_{18}	−3.49	−0.89	g_2		−3.17	g_{36}	−0.76
P_{19}	−3.47	−3.70	g_3		4.50	g_{37}	−3.08
P_{20}	−3.15	−3.00	g_4		6.52	g_{38}	−2.62
P_{21}	−4.11	−4.77	g_5		−1.28	g_{39}	−6.07
P_{22}	−3.56	−3.88	g_6		−2.67	g_{40}	−3.17
P_{23}	−3.55	−3.99	g_7		−2.04	g_{41}	−2.54
P_{24}	13.39	13.49	g_8		−1.67	g_{42}	−5.71
P_{25}	7.10	4.64	g_9		−2.26	g_{43}	−3.75
P_{26}	3.20	7.19	g_{10}		−4.23	g_{44}	−1.93
P_{27}	7.63	6.17	g_{11}		−2.11	g_{45}	−5.99
P_{28}	−0.03	0.16	g_{12}		−1.49	g_{46}	−2.57
P_{29}	0.87	1.65	g_{13}		1.67	g_{47}	0.86
P_{30}	1.47	2.47	g_{14}		−2.66	g_{48}	−0.81
P_{31}	1.07	0.20	g_{15}		−1.70	g_{49}	0.96
P_{32}	−1.65	0.20	g_{16}		−0.85		
P_{33}	1.71	1.09	g_{17}		−1.28		

Consequently, the local graph indices calculated from eq 1–4 with either hydrogen-preserved or hydrogen-suppressed molecular graphs are now included in the regressions in order to improve the correlation. Equations 28–35 resulting from this exercise are shown in Table 6, and it can be seen that, irrespective of the specific local graph indices that are selected, the correlation is greatly improved. The best correlation is obtained with the graph indices derived from eq 3 using a hydrogen-preserved graph, yielding eq 32. It is interesting to note that in these correlations, the graph indices derived from hydrogen-preserved graphs are consistently more effective than the hydrogen-suppressed counterparts, a result that stands in contrast to that obtained previously with the more homogeneous alcohols data set. Equation 32 gave $F = 564.2$, $R^2 = 0.992$, and SD $= 7.57$ K. The significance of each parameter in both eqs 27 and 32 was determined by its t value (student-t test), and these data are given in Table 7. From this table, it can be seen that the proportion of the parameters with significance of at least 0.9 ($t > 1.282$) in eq 27 is 47/50. In eq 32, it is 77/99. For the parameters with significance of at least 0.75 ($t > 0.674$), the proportions are 49/50 in eq 27 and 91/99 in eq 32. Thus despite the diverse set of compounds, the majority of the parameters using these equations are significant. The R^2 value of 0.992 and the SD of 7.57 K from eq 32 contrast with the values (0.968, 15.36 K) delivered by eq 27 and demonstrate that the new approach, combining group parameters with local graph index parameters, is better than the approach which used group parameters alone. The F value of 564.2 given by eq 32 is higher than that (293.8) from eq 27 which shows eq 32 to be more statistically significant than eq 27 and suggests that the improvement of

$$\begin{aligned}\mathrm{bp} = {}& -33.02 + 28.56P_1 + 24.51P_2 - 14.68P_3 - 68.81P_4 + 5.29P_5 + 25.63P_6 + 38.77P_7 + 51.39P_8 + 11.63P_9 + 40.51P_{10} + 20.05P_{11} + 36.48P_{12} - \\ & 25.62P_{13} + 20.26P_{14} + 37.42P_{15} - 23.09P_{16} + 5.31P_{17} - 10.59P_{18} - 50.74P_{19} - 60.11P_{20} - 111.74P_{21} - 121.11P_{22} - 127.47P_{23} + 174.18P_{24} + \\ & 179.12P_{25} + 193.68P_{26} + 103.09P_{27} + 1.58P_{28} + 27.17P_{29} + 24.41P_{30} + 28.60P_{31} + 3.27P_{32} + 16.39P_{33} + 99.79P_{34} + 115.24P_{35} + 92.76P_{36} + \\ & 82.68P_{37} + 136.89P_{38} + 168.58P_{39} + 59.31P_{40} + 117.80P_{41} + 99.21P_{42} + 77.90P_{43} + 86.89P_{44} + 136.93P_{45} + 41.07P_{46} + 1.63P_{47} + 13.74P_{48} - \\ & 26.80P_{49} + 37.47(M)^{1/2} - 52.92g_1 - 23.51g_2 + 45.53g_3 + 129.44g_4 - 21.91g_5 - 33.18g_6 - 30.35g_7 - 81.15g_8 - 50.49g_9 - 55.80g_{10} - 19.63g_{11} - 35.52g_{12} + \\ & 57.44g_{13} - 25.02g_{14} - 24.35g_{15} - 11.21g_{16} - 147.41g_{17} + 3.86g_{18} + 44.62g_{19} - 4.62g_{20} + 74.84g_{21} + 30.58g_{22} + 50.41g_{23} - 231.73g_{24} - \\ & 243.21g_{25} - 268.97g_{26} - 133.09g_{27} - 12.74g_{28} - 38.95g_{29} - 66.06g_{30} - 88.62g_{31} - 61.13g_{32} + 8.44g_{33} - 35.43g_{34} - 155.18g_{35} - 82.49g_{36} - \\ & 102.73g_{37} - 145.80g_{38} - 242.08g_{39} - 77.08g_{40} - 115.65g_{41} - 92.43g_{42} - 158.88g_{43} - 122.09g_{44} - 59.68g_{46} + 19.12g_{47} - 194.19g_{45} - 29.00g_{48} + \\ & 81.54g_{49} \quad (32)\end{aligned}$$

Table 8. Cross-Validation Results Using the Database of 541 Compounds

	group parameters only				group parameters and graph indices			
	$(R^2)_L$	$(S)_L$	$(R^2)_P$	$(S)_P$	$(R^2)_L$	$(S)_L$	$(R^2)_P$	$(S)_P$
1	0.961	16.94	0.986	11.03	0.993	6.92	0.992	9.01
2	0.964	16.25	0.941	24.12	0.994	6.84	0.983	12.41
3	0.963	16.30	0.944	25.68	0.994	6.78	0.986	12.60
4	0.963	16.21	0.957	27.88	0.994	6.75	0.983	14.70
5	0.961	16.69	0.967	17.76	0.994	6.85	0.986	11.93
6	0.963	16.38	0.936	24.60	0.994	6.98	0.993	8.00
7	0.962	16.71	0.960	17.49	0.994	6.94	0.983	11.01
8	0.962	16.59	0.961	20.05	0.993	7.03	0.993	8.20
9	0.961	16.89	0.976	12.23	0.994	6.95	0.987	8.75
10	0.962	16.90	0.971	12.88	0.994	6.57	0.947	16.52
11	0.964	16.21	0.933	25.26	0.993	6.96	0.992	8.33
12	0.963	16.63	0.927	18.84	0.994	6.95	0.982	9.52
13	0.962	16.64	0.959	18.99	0.993	6.97	0.993	7.89
14	0.962	16.68	0.961	18.22	0.994	6.80	0.985	13.77
15	0.961	16.85	0.973	14.67	0.994	6.98	0.990	9.67
16	0.961	16.91	0.983	10.67	0.994	6.95	0.983	9.31
17	0.963	16.37	0.936	24.10	0.994	6.83	0.978	13.04
18	0.962	16.92	0.974	12.78	0.994	6.89	0.982	10.19
19	0.962	16.66	0.965	19.01	0.993	6.92	0.989	9.85
20	0.962	16.72	0.957	16.49	0.993	7.02	0.991	7.33
21	0.963	16.27	0.961	24.42	0.993	6.88	0.992	9.67
22	0.964	16.43	0.896	23.78	0.994	6.95	0.986	8.49
23	0.961	16.83	0.979	13.96	0.993	6.92	0.992	9.06
24	0.963	16.22	0.944	25.87	0.993	6.89	0.990	10.39
25	0.962	16.58	0.960	20.66	0.994	6.85	0.986	11.28
26	0.960	16.94	0.984	13.37	0.993	6.94	0.993	9.21
27	0.960	16.93	0.985	12.21	0.993	6.96	0.992	9.08
28	0.963	16.62	0.871	22.17	0.993	6.99	0.980	8.17
29	0.961	16.91	0.982	11.05	0.993	7.03	0.987	10.50
30	0.961	16.87	0.985	10.96	0.994	6.92	0.982	12.64
31	0.962	16.73	0.957	17.93	0.993	6.06	0.993	6.99
32	0.962	16.64	0.967	18.27	0.993	6.94	0.984	11.57
33	0.963	16.51	0.925	22.28	0.994	6.96	0.979	11.36
34	0.963	16.38	0.946	22.71	0.994	6.74	0.984	12.19
35	0.962	16.81	0.963	14.37	0.993	7.04	0.992	6.02
36	0.962	16.81	0.969	14.29	0.993	7.02	0.992	7.01
37	0.963	16.15	0.948	28.52	0.993	6.80	0.991	12.01
38	0.962	16.48	0.957	21.17	0.994	6.85	0.987	11.25
39	0.962	16.58	0.952	20.44	0.993	7.02	0.993	8.44
40	0.963	16.49	0.940	22.92	0.993	7.04	0.995	7.14
41	0.962	16.76	0.955	17.49	0.993	6.94	0.974	13.86
42	0.962	16.74	0.953	16.25	0.993	7.00	0.988	7.97
43	0.962	16.94	0.969	12.60	0.993	7.03	0.991	7.20
44	0.961	17.03	0.990	8.66	0.993	6.99	0.984	9.24
45	0.963	16.48	0.951	21.40	0.993	7.08	0.997	5.14
46	0.962	16.63	0.968	18.50	0.993	7.06	0.996	6.17
47	0.961	16.92	0.983	12.94	0.994	6.91	0.988	9.71
48	0.962	16.67	0.951	19.07	0.994	6.81	0.980	11.60
49	0.962	16.87	0.973	13.56	0.993	6.98	0.987	8.85
50	0.961	16.85	0.973	14.81	0.994	6.84	0.978	13.89
51	0.962	16.49	0.943	24.75	0.993	6.91	0.986	12.42
52	0.962	16.83	0.956	17.44	0.994	6.87	0.978	12.91
53	0.962	16.84	0.971	13.95	0.993	6.97	0.989	8.62
54	0.962	16.75	0.952	15.66	0.994	6.96	0.984	8.52
55	0.962	16.73	0.961	17.06	0.994	6.92	0.985	9.71
56	0.963	16.52	0.956	22.95	0.993	7.01	0.990	8.89
57	0.963	16.37	0.947	23.41	0.994	6.75	0.982	12.72
58	0.962	16.81	0.960	14.56	0.993	6.98	0.987	8.35
59	0.963	16.66	0.924	18.96	0.994	6.93	0.985	8.93
60	0.962	16.73	0.966	17.48	0.994	6.84	0.977	12.10
mean	0.962	16.65	0.957	18.26	0.993	6.92	0.986	10.02

eq 32 over eq 27 is not due merely to the larger number of parameters used in eq 32.

5. Evaluation of the Models. The stability and prediction potential of eq 27 and 32 were examined and compared by means of cross-validation experiments. In each of the cross-validation tests, about 5% of the structures were removed from the dataset of 541 compounds, and the remaining 95% were used to establish the model for the boiling point calculation. This model was then used to predict the boiling points of the compounds in the 5% sample. The values of R^2 and SD for the dataset can be regarded as measures of the stability of the model and the cross-validated R^2 and SD obtained from the 5% test set can be used to indicate the prediction potential of the model and hence the validity of the methodology. In order to ensure that the sampling was fairly comprehensive, some 60 such tests were performed, and the results are given in Table 8. It can be seen from Table 8 that when the local graph indices are omitted from the regression, the average value of R^2 over the 60 tests was 0.962 and the standard deviation (SD) was 16.65 K. The prediction potential of these regressions was characterized by $R^2 = 0.957$ and SD = 18.26 K. When the3 local graph indices were included in the regressions, an average R^2 of 0.993 and a standard deviation (SD) of 6.92 K was obtained in the learning database and the predictive accuracy was associated with an R^2 of 0.986 and a standard deviation of 10.02 K. Thus use of the local graph indices propelled R^2 from 0.962 to 0.992 for the learning datasets and from 0.957 to 0.986 for the predictive datasets. At the same time, the standard deviations dropped to 6.92 and 10.02 K, as compared to 16.65 and 18.26 K, respectively, obtained with group contribution parameters alone. The accuracy of the boiling points estimated by means of eq 27 can be seen from Figure 2, which is a plot of the estimated *versus* the measured values. The same plot for the boiling points estimated using eq 32 is shown in Figure 3. It is clear from these two graphs that the errors in the estimates from eq 27 are greater than those arising from the use of eq 32.

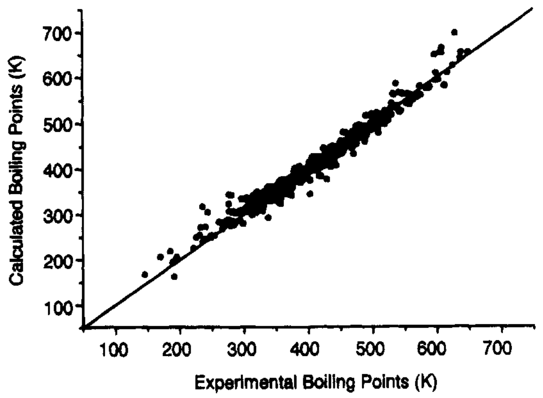


Figure 2. Boiling points calculated from eq 27 vs experimental boiling points.

The boiling points of the 63 alcohols in Table 3 were estimated using eq 32. For this set of 63 related compounds, an R^2 of 0.98 and a standard deviation (SD) of 5.40 K was obtained. If the 20 alcohols that were also in the large (541) database were removed from the dataset of 63 compounds, the boiling point prediction for the remaining 43 alcohols, which were not included in the learning set, had a predictive R^2 of 0.96 and a standard deviation of 5.75 K.

To further evaluate the predictive power of the both models (eq 27 and 32), we randomly selected 32 new compounds from the Beilstein database, excluding compounds with groups that were not in Tables 1 or 2. The boiling point values of these 32 compounds were predicted using eq 27

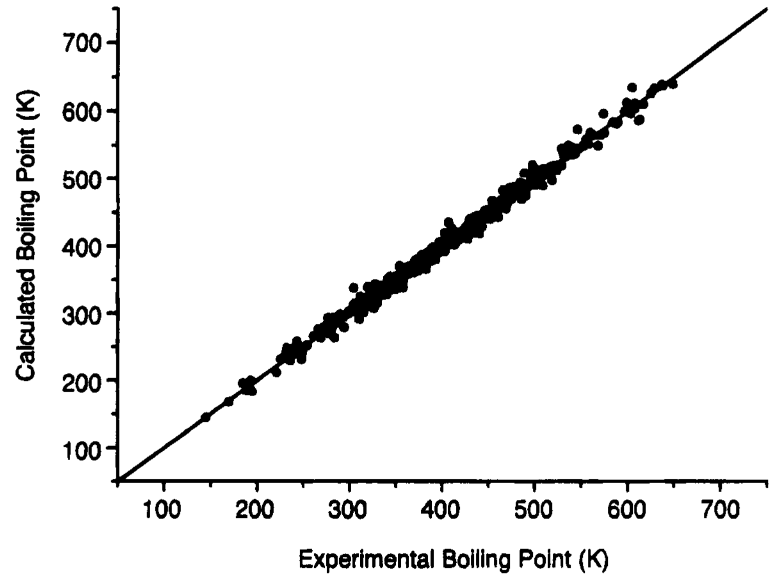
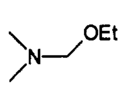
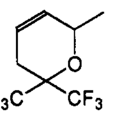


Figure 3. Boiling points calculated from eq 32 vs experimental boiling points.

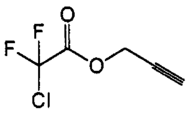
and 32, giving the results shown in Table 9. As can be seen from Table 9 both eq 27, which was based upon the group contribution parameters alone, and eq 32, which was based upon combinations of the group contribution parameters and the graph indices, were able to give fairly good predictions for the normal boiling points of these 32 new compounds. Equation 27, predicted the normal boiling points of 13 compounds with an error less than 10 K and of 18 compounds with an error less than 20 K. This equation delivered six errors larger than 35 K, and three errors larger than 40 K. The largest error, 61.42 K, predicted by eq 27 for these compounds, was associated with compound B-4805384. Equation 32 predicted the normal boiling points of 14 compounds with an error less than 10 K and of 22 compounds with an error less than 20 K. This equation gave no errors larger than 40 K. The largest error for these 32 compounds predicted by eq 32, 34.49 K, was for compound B-4652392. Thus, overall, eq 32 shows a better predictive power than eq 27.



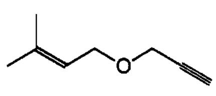
4652392



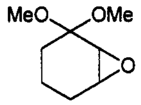
4805384



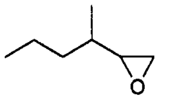
4962795



4950940



4797529



4652749

Table 9. Predictive Results Using Eq 27 and 32 for 32 New Compounds from the Beilstein Database

no.	Beilstein no.	BP_{calc}[a]	BP_{calc}[b]	BP(EXP)[c]	errors1[d]	errors2[e]
1	4652392	354.21	361.66	376.35	−41.94	−34.49
2	4805384	451.07	422.65	389.65	61.42	33.00
3	4962795	407.34	422.49	391.15	16.19	31.34
4	4950940	405.99	404.82	374.15	31.84	30.67
5	4797529	470.60	469.41	439.15	31.45	30.26
6	4652749	408.08	414.20	385.65	17.57	28.55
7	4961624	555.18	508.93	534.15	21.03	−25.22
8	4953300	451.66	436.90	413.15	38.51	23.75
9	4964763	573.74	520.84	544.65	29.09	−23.81
10	4952192	442.52	446.19	425.15	17.37	21.04
11	4967737	591.81	533.84	553.65	38.16	−19.81
12	4738729	431.65	417.28	399.15	32.50	18.13
13	4959787	515.48	509.53	492.15	23.33	17.38
14	4949735	433.81	420.85	437.15	−3.34	−16.30
15	4733109	398.47	390.38	406.65	−8.18	−16.27
16	4903024	383.34	389.71	373.65	9.69	16.06
17	4964814	573.74	522.67	537.65	36.09	−14.98
18	4857835	438.52	444.30	433.15	5.37	11.15
19	4653624	421.12	462.76	453.15	−32.03	9.61
20	4738311	402.99	455.84	446.30	−43.31	9.54
21	4953968	442.15	452.47	443.15	−1.00	9.32
22	4950344	447.66	422.10	428.15	19.51	−6.05
23	4780839	410.19	397.79	403.15	7.04	−5.36
24	4954566	388.99	378.81	384.15	4.84	−5.34
25	4849541	491.85	507.19	503.15	−11.30	4.04
26	4796687	383.28	373.25	377.15	6.13	−3.90
27	4953328	378.53	378.93	375.15	3.38	3.82
28	4959933	433.73	441.26	438.15	−4.42	3.11
29	4653190	404.88	417.73	414.65	−6.57	3.08
30	4956596	416.10	407.31	410.15	5.95	−2.84
31	4950041	424.58	404.76	404.15	20.43	0.61
32	4949821	434.17	441.55	441.15	−6.98	0.40

[a] Calculated normal boiling points (K) from eq 27. [b] Calculated normal boiling points (K) from eq 32. [c] Experimental normal boiling points (K) from Beilstein database. [d] Calculated bp from eq 27—experimental bp. [e] Calculated bp from eq 32—experimental bp.

6. Further Possible Improvements of the Models. Thus boiling point prediction by the group contribution approach is considerably improved if local graph indices are incorporated into the calculations. The combined approach fails, however, to accommodate the possibility of intramolecular interactions between groups. The boiling point of succinic acid, for example, is 508.0 K, while eq 32 predicts a value of 563.6 K, 55.6 K too high. The interaction of the two carboxyl groups in this molecule is presumably responsible, at least in part, for the fact that the boiling point is lower than predicted by eq 32. Accounting for such interactions remains a subject for future research.

The compounds from Table 9 for which the prediction errors were highest are shown below, and their structures suggest some speculation concerning the poor performance of the program. The ethoxyamine moiety in 4652392 is a residue which demands its own parameters; having none, the program fails badly. The O—C—$(CF_3)_2$ fragment in

4805384 and the epoxide rings in 4797529 and 4652749 may also require independent parameterization.

Further shortcomings in the results stems from the fact that 19 of the groups defined in Tables 1 and 2 are not represented in the database of 541 compounds that was used here, and all the 68 defined parameters in Tables 1 and 2 were constrained to organic compounds. This essentially limits the applicability of both eq 27 and 32, and it is therefore necessary to add to the database more compounds, which can support the 19 missing parameters in Table 2. In this way it may be possible to develop a model capable of predicting the normal boiling points for all classes of organic compounds. To make the model able to predict the normal boiling points of organometallic compounds, it is necessary to expand the parameter set in Tables 1 and 2 and also to include organometallic compounds in the learning set underlying the model.

CONCLUSIONS

A novel method for the estimation of the normal boiling points of organic compounds has been developed by combination of the group contribution method with an approach based upon graph theory. Inclusion of local graph indices led to a significant improvement in predictive accuracy of the group contribution method. The prediction potential of the new approach was evaluated by means of cross-validation experiments, and it was also shown that the method produced an improvement in the accuracy of the boiling point estimation for 32 new molecules compared to the results obtained with the group contribution method alone. The prediction potential of the method was also illustrated by the satisfactory calculation of the normal boiling points of 43 aliphatic alcohols that were absent from the training set.

Prediction of the normal boiling point of a molecule requires only the input of the KLN code or a connection table and involves a simple and efficient calculation which is completed rapidly. This program can serve as a general tool for boiling point prediction and operates with an accuracy of ±20 K for most of compounds tested.

ACKNOWLEDGMENT

The authors would like to thank Dr. Robert V. Henderson for his insightful discussions and invaluable suggestions.

REFERENCES AND NOTES

(1) Nys, G. C.; Rekker, R. F. Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. *Chim. Ther.* **1973**, *8*, 521–535.

(2) Rekker, R. F. *The Hydrophobic Fragmental Constant*; Elsevier: New York, 1977.

(3) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71*, 525–616.

(4) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.

(5) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. *Eur. J. Med. Chem.–Chim. Ther.* **1984**, *19*, 71–78.

(6) Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 180.

(7) Visvanadham, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

(8) Klopman, G.; Wang, S. A. Computer-Automated Structure Evaluation (CASE) Approach to Calculation of Partition Coefficients. *J. Comput. Chem.* **1991**, *12*, 1025–1032. Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log *P* Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752–781.

(9) Irmann, F. A Simple Correlation Between Water Solubility and Structure of Hydrocarbons and Halohydrocarbons. *Chem.-Ing.-Tech.* **1965**, *37*, 789–798.

(10) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474–482.

(11) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure–Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.

(12) Lyderson, A. L. *Estimation of Critical Properties of Organic Compounds*; University of Wisconsin College of English: End. Exp. Stn. Rpt 3, Madison, WI, April 1955.

(13) Lyderson, A. L. *Estimation of Critical Properties of Organic Compounds*; University of Wisconsin College English: End. Exp. Stn. Rpt 3, Madison, WI, April 1955.

(14) Yoneda, Y. An Estimation of the Thermodynamic Properties of Organic Compounds in the Ideal Gas State. I. Acyclic Compounds and Cyclic Compounds with a Ring of Cyclopentane, Cyclohexane, Benzene or Naphthalene. *Bull. Chem. Soc. Jpn.* **1979**, *52*, 1297–1314.

(15) Thinh, T. P.; Trong, T. K. Estimation of Standard Heats of Formation, ΔH_T^f, Standard Entropies of Formation, ΔS_T^f, Standard Free Energies of Formation, ΔG_T^f, and Absolute Entropies, ΔS_T of Hydrocarbons from Group Contributions: An Accurate Approach. *Can. J. Chem. Eng.* **1976**, *54*, 344.

(16) Reid, R. C.; Sherwood, T. K. *The Properties and Gases and Liquids*, 2nd. ed.; McGraw-Hill, New York, 1966; Chapter 2.

(17) Joback, K. G. M.S. Thesis in Chemical Engineering, MIT, Cambridge, MA, June 1984.

(18) Randić, M. Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(19) Hall, L. H.; Kier, L. B.; Murray, W. J. Molecular Connectivity. II. Relationship to Water Solubility and Boiling Point. *J. Pharm. Sci.* **1975**, *64*, 1974–1977.

(20) Murray, W. J.; Hall, L. H.; Kier, L. B. Molecular Connectivity. III. Relationships to Partition Coefficients. *J. Pharm. Sci.* **1975**, *64*, 1978–1980.

(21) Nirmalakhandan, N. N.; Speece, R. E. Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. *Env. Sci. Tech.* **1988**, *22*, 328–338.

(22) Nirmalakhandan, N. N.; Speece, R. E. Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. 2. Application to PNAs, PCBs, PCDDs, etc. *Env. Sci. Tech.* **1989**, *23*, 708–713.

(23) Speece, R. E. Comments on Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. 2. Application to PNAs, PCBs, PCDDs, etc. *Env. Sci. Tech.* **1990**, *24*, 927–929.

(24) Klopman, G.; Raychaudhury, C. A. Novel Approach to the Use of Graph Theory in Structure–Activity Relationship Studies. Application to the Qualitative Evaluation of Mutagenicity in a Series of Nonfused Ring Aromatic Compounds. *J. Comput. Chem.*, **1988**, *9*, 232–243.

(25) Klopman, G.; Raychaudhury, C. Vertex Indices of Molecular Graphs in Structure–Activity Relationships: A Study of the Convulsant–Anticonvulsant Activity of Barbiturates and the Carcinogenicity of Unsubstituted Polycyclic Aromatic Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 12–19.

(26) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76–82.

(27) Hall, L. H.; Kier, L. B. Determination of Topological Equivalence in Molecular Graphs from the Topological State. *Quant. Struct.–Act. Relat.* **1990**, *9*, 115–131.

(28) Filip, P. A.; Balaban, T. S.; Balaban, A. T. A New Approach for Devising Local Graph Invariants: Derived Topological Indexes with Low Degeneracy and Good Correlation Ability. *J. Math. Chem.* **1987**, *1*, 61–83.

(29) For a review, see: Katritzky, A. R.; Gordeeva, E. V. Traditional Topological Indices vs. Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835–857.

(30) Klopman, G.; McGonigal, M. Computer Simulation of Physical-Chemical Properties of Organic Molecules. 1. Molecular System Identification. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 48–52.

(31) Weast, R. C. CRC Handbook of Chemistry and Physics, 60th. ed.; CRC Press, Inc.: Boca Raton, FL, 1979.