Estimation of Normal Boiling Points from Group Contributions

S. E. Stein*

Chemical Kinetics and Thermodynamics Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899

R. L. Brown

Chappaquiddick Road, Edgartown, Massachusetts 02539

Received June 24, 1993®

A group contribution method for normal boiling point estimations was developed using a data base of 4426 diverse organic compounds. With this data set, boiling point predictions had a average absolute error of 15.5 K corresponding to a 3.2 average percent error. For a data set of 6584 other compounds, not used in deriving the method, the average absolute error was 20.4 K with a 4.3 average percent error. A vapor pressure equation was tested and used to extrapolate boiling points measured at reduced pressure. Thus this method may also be used to predict vapor pressures.

INTRODUCTION

Group contribution methods for chemical property estimation have the advantage of simplicity and generality. Their theoretical basis rests on the assumption that forces between atoms in the same or different molecules have short ranges.[1] For boiling point estimations, this assumption often gives good predictions for small, nonpolar groups. Predictions for polar compounds and those involving hydrogen bonding can be improved by increasing group sizes and considering interactions between groups.[2-4] Thirty-six different methods for estimating boiling points have been recently summarized by Horvath.[2b] One set of non-interacting group increments for normal boiling point prediction which require only chemical structure and is applicable to a wide range of compounds has been developed by Joback and Reid.[5] This set of groups was derived from a data base containing 438 organic compounds and gave an average absolute error of 12.9 K corresponding to a 3.6 average percent error.

In the present work, we extended Joback and Reid's method primarily by increasing the number of groups. We also discovered and corrected for a temperature-dependent bias. New group increment values were derived from 4426 experimental boiling points taken from the Aldrich chemical catalog.[6] For this data set, our predicted boiling points had an average absolute error of 15.5 K with a 3.2 average percent error. This is comparable to the results from the much smaller data set used by Joback and Reid. In addition we used our new group increment values to predict a set of 6584 *completely different* compounds taken from the HODOC data base.[7] For this set (which was *not* used in deriving the increment values) the predictions had an average absolute error of 20.4 K with a 4.3 average percent error. Experimental boiling points were used as delivered. The Aldrich data contained few if any bad values. Although some spurious values were evident in the HODOC data (see Discussion), critical analysis of this data base was not necessary for our purposes since it was not used in the fitting procedure.

Many of the boiling points in our data sets were measured at pressures less than 1 atm. We tested a vapor pressure equation and used it to extrapolate this data to atmospheric pressure. Incorporating this equation into the method allows one to estimate vapor pressures as well as normal boiling points.

DETERMINATION OF GROUP INCREMENT VALUES

Joback and Reid used a set of 41 groups. This set is denoted S_0. They calculated the normal boiling point T_b of a compound by adding group increment values according to the relation

$$T_b = 198.2 + \sum_i n_i g_i \qquad (1)$$

where $g_i = (\Delta T_b)_i$ is a group increment value and n_i is the number of times the group occurs in the compound. We have developed a new set S_1 by increasing the number of groups to 85. While this may seem to be a large addition, many of these new groups represent subdivisions of groups in the original set S_0 made possible by the much larger number of compounds used in the present study. New groups were defined for one of three different reasons. (1) When a clear gain in predictive accuracy would result, groups were subdivided by making finer distinctions in bonding. For example, we now distinguish whether the group –OH is attached to a primary, secondary, tertiary, or aromatic carbon or to a non-carbon atom. Set S_0 only distinguished between aliphatic and phenolic –OH. We used the same classification for the –Cl group attachment. Each of the other original halogen groups and the groups $-NH_2$, –CN, and –SH were split into two groups depending on their attachment to aliphatic or aromatic carbon. We also categorized the ring increment groups =CH– and =C< by whether or not they were in aromatic rings. These minor modifications added 16 groups to S_0. (2) Another set of new groups were defined by combining two or three of the original groups into a larger functional unit. Examples of these are –C(O)NH–, –C(O)N<, >NNO, –N=NNH–, etc. All were introduced because they produced a somewhat better fit than the uncombined groups. Many of these were represented by only a few molecules in the data sets and were included only for completeness. (3) For groups containing elements not present in S_0, a number of entirely new groups had to be introduced. Examples are >PH, >SiH–, >B–, –Se–, and >Sn<.

The first step in our fitting procedure was to take those groups in S_1 which were the same as those in S_0 or which would be formed from combinations of groups in S_0. These were given the increment values derived by Joback and Reid. We call this set S_{1a}. With it, we could calculate boiling points via eq 1 for 4286 compounds in the Aldrich data base.

* Abstract published in *Advance ACS Abstracts*, March 1, 1994.

0095-2338/94/1634-0581$04.50/0

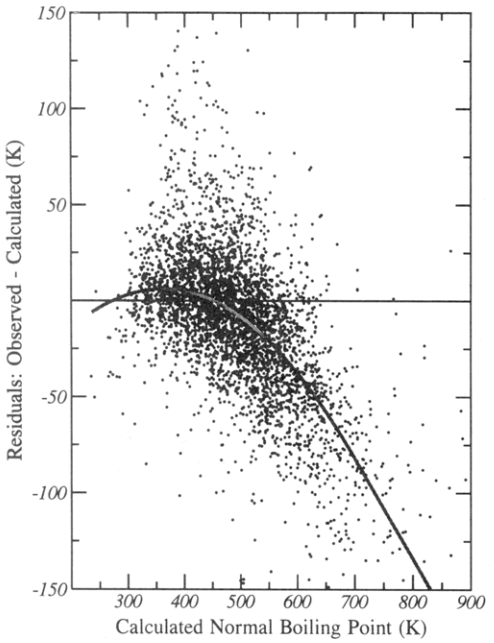

Figure 1. Scatter plot of residuals, T_b(obsd) – T_b(calcd), for the Aldrich data set as a function of T_b(calcd from eq 1) for group increment set S_{1a} (Joback–Reid increment values). The shaded curve is the correction eq 2, which is subsequently applied to boiling points calculated from eq 1.

Results using S_{1a} are shown in Figure 1. This is a scatter diagram giving the residuals $\delta = T_b(\text{obsd}) - T_b(\text{calc})$ (observed minus calculated normal boiling points) as a function of the calculated boiling points. Above 500 K, eq 1 increasingly overpredicts the boiling point. By minimizing absolute residuals we fitted a polynomial to this data and added it as a correction to the calculated values. The corrected calculated value T_b(corr) is given by

$$T_b(\text{corr}) = T_b - 94.84 + 0.5577T_b - 0.0007705T_b^2, \quad T_b \le 700\ \text{K}$$
$$= T_b + 282.7 - 0.5209T_b, \quad T_b > 700\ \text{K} \tag{2}$$

where T_b is calculated according to eq 1. Because of its quadratic term, eq 2 produces a maximum in T_b(corr) around $T_b = 1000$ K. Linearization above 700 K avoids this unrealistic behavior. Before correction, the 4280 boiling points from the Aldrich data set, calculated with group set S_{1a}, had an average absolute error of 28.9 K; after correction, this was reduced to 21.7 K.

A large number (1761) of boiling points in the Aldrich data set were reported at pressures other than 1 atm. These had to be extrapolated to 1 atm to compare with our calculated values. To do this, we used a vapor pressure equation developed by Lee and Kesler.[8] In the Appendix we discuss why we chose this equation and how we incorporated it into the method.

The next step in the fitting procedure was to obtain group increment values for those groups in S_1 which could not be derived from S_0. These are the last 14 groups shown in Table 1. We denote this set S_{1b}. For each group i in S_{1b} we selected a set of compounds from the Aldrich data which contained group i, any groups from S_{1a}, and none of the others in S_{1b}. For each compound set i we calculated the average absolute deviation $\langle|\delta|\rangle_i$ using eqs 1, 2, and A.5. This quantity is a function of the group increment values, g_i, and a vector $\mathbf{c}(i)$ denoting the compound set containing group i. This relation can be represented by the following equation

$$\langle|\delta|\rangle_i(\min\ via\ g_i) = F(g_1,...,g_i,...,g_n;\mathbf{c}(i)) \tag{3}$$

We considered F to be a function of g_i only, keeping the other increment values fixed. The value of g_i was taken to be that which minimized $\langle|\delta|\rangle_i$. This one-dimensional minimization procedure was repeated for each S_{1b} group in turn. Adding the resulting increment values to set S_{1a}, we obtained the 85 group set S_{1c} which served as the starting point for our final fitting procedure. Values for group increments in set S_{1c} are shown in parentheses in the last column of Table 1.

The final fitting procedure was simply a generalization of the one-dimensional minimizations outlined above. For each of the 85 groups in turn, only those compounds in the Aldrich set which contained that particular group were selected. Beginning with the increment values of set S_{1c}, we did one-dimensional minimizations of the $\langle|\delta|\rangle_i$ by repeatedly cycling through the set of 85 groups. During a particular cycle, each updated increment value g_i was used in the subsequent minimizations. For the $(k+1)$th cycle, we have

$$\langle|\delta|\rangle_1^{k+1} = F(g_1^{k+1},g_2^k,...,g_n^k;\mathbf{c}(1))$$
$$\langle|\delta|\rangle_2^{k+1} = F(g_1^{k+1},g_2^{k+1},...,g_n^k;\mathbf{c}(2))$$
$$\vdots$$
$$\langle|\delta|\rangle_n^{k+1} = F(g_1^{k+1},g_2^{k+1},...,g_n^{k+1};\mathbf{c}(n)) \tag{4}$$

where $n = 85$. Convergence to a constant set of g_i values was achieved after about 10 cycles. We used this fitting procedure because we lacked the computing power to fit, via absolute residual minimization, all 85 parameters at the same time. It does have the advantage of minimizing changes to the original Joback and Reid group values. We doubt that a simultaneous fit would significantly improve the average absolute error. The resulting increment values for set S_1 are given in Table 1. These would provide excellent starting values for anyone having the desire and means to do a more orthodox fitting procedure.

Statistical results using set S_1 are shown in Table 2 for each of the groups. Listed here is the number of compounds used in fitting each group value, the resulting average absolute error, standard deviation, median absolute error, and average percent error. These numbers give some indication of the accuracy of boiling point predictions for compounds containing a particular group.

Statistical results using S_1 increment values with the complete Aldrich and HODOC data sets are given in Table 3. Also shown are results for three subsets: (a) boiling points measured at 1 atm, (b) those requiring extrapolation to 1 atm, and (c) boiling points for compounds containing only the hydrocarbon groups.

DISCUSSION

Our purpose has been to expand and test a simple, general purpose group method on an unrestricted data base. Considering the large size, diversity, and unscreened nature of the two data bases, we find the results shown in Table 3 quite satisfactory. Figure 2 is a scatter plot of the Aldrich data residuals as a function of calculated boiling points using increment set S_1 and corrections with eqs 2 and A.5. Comparison with Figure 1 shows the significantly decreased scatter and elimination of drift resulting from the use of the

Table 1. Group Contributions for Normal Boiling Point to Critical Temperature Ratio, θ; Critical Pressure, P_c (bar); and Normal Boiling Point T_b (K)

structural group	$\Delta\theta$	ΔP_c	ΔT_b
carbon increments:			
$-CH_3$	0.0141[a]	−0.0012[a]	21.98 (23.58)[f]
$>CH_2$	0.0189	0.0000	24.22 (22.88)
$>C_rH_2$[b]	0.0100	0.0025	26.44 (27.15)
>CH–	0.0164	0.0020	11.86 (21.74)
$>C_rH-$	0.0122	0.0004	21.66 (21.78)
>C<	0.0067	0.0043	4.50 (18.25)
$>C_r<$	0.0042	0.0061	11.12 (21.32)
$=CH_2$	0.0113	−0.0028	16.44 (18.18)
=CH–	0.0129	−0.0006	27.95 (24.96)
$=C_rH-$	0.0082	0.0011	28.03 (26.73)
=C<	0.0117	0.0011	23.58 (24.14)
$=C_r<$	0.0143	0.0008	28.19 (31.01)
*aa*CH[c]	0.0082	0.0011	28.53 (26.73)
*aa*C–	0.0143	0.0008	30.76 (31.01)
*aaa*C	0.0143	0.0008	45.46 (31.01)
≡CH	0.0027	−0.0008	21.71 (9.20)
≡C–	0.0020	0.0016	32.99 (27.38)
oxygen increments:			
–OH	0.0741	0.0112	106.27 (92.88)
1-OH[d]	0.0741	0.0112	88.46 (92.88)
2-OH[d]	0.0741	0.0112	80.63 (92.88)
3-OH[d]	0.0741	0.0112	69.32 (92.88)
ϕ-OH[e]	0.0240	0.0184	70.48 (76.34)
–O–	0.0168	0.0015	25.16 (22.42)
$-O_r-$	0.0098	0.0048	32.98 (31.22)
–OOH	0.0909	0.0127	72.92 (115.30)
carboxyl increments:			
–CHO	0.0379	0.0030	83.38 (72.24)
>CO	0.0380	0.0031	71.53 (76.75)
$>C_rO$	0.0284	0.0028	94.76 (94.97)
–C(O)O–	0.0481	0.0005	78.85 (81.10)
$-C_r(O)O_r-$	0.0481	0.0005	172.49 (81.10)
–C(O)OH	0.0791	0.0077	169.83 (169.09)
$-C(O)NH_2$	0.0623	0.0140	230.39 (149.98)
–C(O)NH–	0.0675	0.0108	225.09 (126.92)
$-C_r(O)N_rH-$	0.0414	0.0142	246.13 (147.79)
–C(O)N<	0.0549	0.0105	142.77 (88.49)
$-C_r(O)N_r<$	0.0453	0.0102	180.22 (106.71)
nitrogen increments:			
$-NH_2$	0.0243	0.0109	61.98 (73.23)
ϕ-NH_2	0.0243	0.0109	86.63 (87.84)
>NH	0.0295	0.0077	45.28 (50.17)
$>N_rH$	0.0130	0.0114	65.50 (52.82)
>N–	0.0169	0.0074	25.78 (11.74)
$>N_r-$	0.0169	0.0074	32.77 (11.74)
>NOH	0.0910	0.0186	104.87 (104.62)
>NNO	0.0567	0.0076	184.68 (76.24)
*an*N	0.0085	0.0076	39.88 (57.55)
nitrogen increments:			
=NH	0.0255	−0.0099	73.40 (74.60)
=N–	0.0255	−0.0099	31.32 (74.60)
$=N_r-$	0.0085	0.0076	43.54 (57.55)
$=N_rN_rH-$	0.0215	0.0190	179.43 (110.37)
$-N_r=C_rRN_rH-$	0.0358	0.0198	284.16 (141.38)
–N=NNH–	0.0805	−0.0121	257.29 (199.37)
–N=N–	0.0510	−0.0198	90.87 (149.20)
–NO	0.0398	0.0002	30.91 (64.50)
$-NO_2$	0.0437	0.0064	113.99 (152.54)
–CN	0.0496	−0.0101	119.16 (125.66)
ϕ-CN	0.0496	−0.0101	95.43 (96.30)
halogen increments:			
–F	0.0111	−0.0057	0.13 (−0.03)
ϕ-F	0.0111	−0.0057	−7.81 (−0.03)
–Cl	0.0105	−0.0049	34.08 (38.13)
1-Cl[d]	0.0105	−0.0049	62.63 (38.13)
2-Cl[d]	0.0105	−0.0049	49.41 (38.13)
3-Cl[d]	0.0105	−0.0049	36.23 (38.13)
ϕ-Cl	0.0105	−0.0049	36.79 (38.13)
–Br	0.0133	0.0057	76.28 (66.86)
ϕ-Br	0.0133	0.0057	61.85 (66.86)
–I	0.0068	−0.0034	111.67 (93.84)
ϕ-I	0.0068	−0.0034	99.93 (93.84)
sulfur increments:			
–SH	0.0031	0.0084	81.71 (63.56)
ϕ-SH	0.0031	0.0084	77.49 (63.56)
–S–	0.0119	0.0049	69.42 (68.78)
$-S_r-$	0.0019	0.0051	69.00 (52.10)
>SO			154.50 (104.22)
$>SO_2$			171.58 (172.69)
>CS			106.20 (123.63)
$>C_rS$			179.26 (188.65)
phosphorus increments:			
$-PH_2$			59.11 (67.87)
>PH			40.54 (58.04)
>P–			43.75 (42.63)
>PO–			107.23 (114.09)
silicon increments:			
>SiH–			27.15 (22.07)
>Si<			8.21 (10.05)
$>Si_r<$			−12.16 (−13.59)
miscellaneous increments:			
>B–			−27.27 (−10.77)
–Se–			92.06 (90.46)
>Sn<			62.89 (58.09)

[a] Temperature ratio and critical pressure increments are from ref 5, used directly or in appropriate combinations. These were required for extrapolating boiling point data to 1 atm. [b] Atoms having the subscript *r* are in rings. [c] The symbol *a* denotes an aromatic bond. [d] Numbers 1, 2, and 3, denote attachment to primary, secondary, and tertiary carbon atoms, respectively. [e] The symbol ϕ denotes an aromatic system. [f] Values in parentheses are those for increment set S_{1c}. (The first 71 values constitute increment set S_{1a}; the last 14 are for set S_{1b}.)

new group values and corrections. The distribution of errors in both the Aldrich and HODOC sets are shown as points in Figure 3. The solid curve is the smoothed distribution of residuals for the Aldrich data; the dotted curve is that for the HODOC data. Shaded areas are Gaussian (random) distributions having the same width at half-height as the observed distributions and centered on the median errors; the wider Gaussian applies to the HODOC data.

Both Aldrich and HODOC distributions exhibit large tails of outliers relative to a Gaussian distribution. For each, 18% of the distribution is in these tails. Without these outliers, standard deviations for the Aldrich and HODOC errors would have been only 11 and 14 K, respectively. Except for a strong tendency of the method to overpredict boiling points for perfluorinated compounds, we could discern no other compound-related biases responsible for these tails.

A cursory check of the HODOC data revealed that a number of boiling points were measured at reduced pressure, but not so reported. These would produce large negative residuals. As noted elsewhere,[9] such errors are common in other data bases. Data reported at pressures other than 1 atm are presumably free from such bias since pressures are explicitly stated. Consistent with this is the similarity of the statistics shown in Table 3 for the extrapolated Aldrich and HODOC data sets. For the *un*extrapolated sets, the statistics for the HODOC data can be made virtually identical to those for the Aldrich data simply by excluding 4% of its largest negative residuals. Inspection of these very large outliers confirmed that they were largely due to the above-mentioned omission. Thus it is probable that much of the difference between the Aldrich and HODOC statistics arises from this reporting bias.

Table 2. Statistical Results of Fitting Procedure for Individual Groups

structural group	no. of compds used	av abs error (K)	std dev (K)	median abs error (K)	av percent error
carbon increments:					
$-CH_3$	2832 (5069)[a]	14 (18)[a]	22 (34)[a]	9 (11)[a]	3.0 (3.8)[a]
$>CH_2$	2200 (3739)	14 (18)	22 (33)	9 (11)	2.9 (3.7)
$>C_rH_2$	757 (1225)	16 (18)	24 (35)	11 (10)	3.3 (3.8)
>CH–	603 (1818)	14 (17)	21 (29)	9 (12)	3.0 (3.6)
$>C_rH–$	457 (902)	14 (17)	23 (33)	10 (10)	3.0 (3.6)
>C<	370 (666)	18 (25)	28 (41)	12 (16)	4.3 (6.0)
$>C_r<$	118 (308)	19 (22)	28 (42)	14 (13)	4.2 (4.7)
$=CH_2$	284 (470)	11 (13)	22 (20)	7 (9)	2.7 (3.0)
=CH–	405 (935)	13 (16)	23 (31)	8 (10)	2.7 (3.4)
$=C_rH–$	285 (532)	20 (20)	30 (36)	13 (13)	4.1 (4.1)
=C<	169 (434)	13 (19)	21 (39)	8 (11)	2.9 (4.3)
$=C_r<$	257 (494)	21 (24)	31 (48)	13 (14)	4.2 (5.0)
*aa*CH	1727 (2426)	15 (23)	24 (46)	10 (13)	2.9 (4.6)
*aa*C–	1740 (2436)	15 (23)	23 (46)	10 (13)	2.9 (4.6)
*aaa*C	110 (290)	16 (18)	28 (33)	9 (11)	2.8 (3.1)
≡CH	43 (88)	9 (13)	13 (19)	8 (9)	2.3 (3.5)
≡C–	77 (215)	10 (13)	13 (20)	8 (8)	2.3 (3.1)
oxygen increments:					
–OH	14 (31)	25 (52)	34 (81)	18 (29)	5.2 (10.7)
1-OH	370 (291)	16 (20)	21 (31)	12 (14)	3.2 (3.9)
2-OH	205 (416)	17 (20)	24 (40)	11 (13)	3.4 (4.4)
3-OH	46 (142)	13 (17)	19 (24)	8 (11)	2.6 (3.6)
ϕ-OH	179 (257)	23 (35)	30 (66)	17 (19)	4.2 (7.2)
–O–	525 (691)	16 (22)	23 (46)	11 (11)	3.3 (4.5)
$–O_r–$	217 (268)	18 (24)	24 (39)	13 (16)	3.8 (5.0)
–(O)OH	1 (1)[b]	0 (66)		0 (66)	0.0 (13.3)
carboxyl increments:					
–CHO	178 (172)	19 (21)	29 (33)	10 (13)	3.9 (4.5)
>CO	419 (517)	17 (22)	24 (46)	11 (12)	3.6 (4.8)
$>C_rO$	134 (199)	21 (29)	32 (62)	13 (12)	4.1 (6.3)
–C(O)O–	530 (903)	16 (19)	26 (33)	10 (12)	3.2 (3.8)
$–C_r(O)O_r–$	59 (55)	28 (38)	37 (55)	25 (22)	5.2 (7.8)
–C(O)OH	169 (307)	15 (28)	27 (52)	8 (14)	3.0 (5.9)
$–C(O)NH_2$	10 (25)	18 (48)	23 (81)	13 (20)	3.5 (9.6)
–C(O)NH–	19 (46)	31 (61)	51 (85)	15 (45)	5.8 (12.0)
$–C_r(O)N_rH–$	6 (10)	38 (55)	83 (69)	4 (35)	10.3 (8.8)
–C(O)N<	19 (42)	17 (27)	23 (35)	14 (19)	3.3 (5.0)
$–C_r(O)N_r<$	17 (34)	29 (47)	47 (69)	12 (31)	5.1 (9.6)
nitrogen increments:					
$–NH_2$	217 (134)	16 (16)	25 (23)	9 (10)	3.3 (3.3)
ϕ-NH_2	146 (111)	15 (30)	20 (56)	11 (15)	2.8 (5.9)
>NH	116 (189)	21 (26)	34 (49)	12 (16)	4.3 (5.7)
$>N_rH$	86 (125)	33 (31)	45 (62)	24 (20)	6.3 (6.8)
>N–	138 (176)	20 (35)	31 (67)	14 (20)	4.0 (7.1)
$>N_r–$	105 (186)	24 (32)	34 (53)	16 (19)	4.6 (6.1)
>NOH	1 (1)[b]	0 (21)		0 (21)	0.0 (4.9)
>NNO	2 (6)	4 (32)	4 (35)	4 (29)	1.0 (6.6)
*aa*N	171 (224)	17 (21)	25 (35)	12 (13)	3.5 (3.9)
=NH	3 (2)	7 (126)	11 (15)	4 (126)	1.3 (27.8)
=N–	12 (40)	10 (36)	16 (69)	6 (18)	2.0 (7.7)
$=N_r–$	62 (91)	30 (29)	46 (50)	19 (19)	5.8 (6.2)
$=N_rN_rH–$	6 (10)	16 (17)	24 (22)	6 (12)	2.9 (3.4)
$–N_r=C_rRN_rH–$	3 (4)	5 (25)	7 (34)	4 (15)	1.0 (4.9)
–N=NNH–	2 (1)[b]	29 (32)	29 (–)	29 (32)	5.2 (5.2)
–N=N–	3 (17)	39 (55)	58 (91)	18 (23)	6.5 (13.8)
–NO	4 (13)	4 (15)	5 (22)	5 (8)	1.2 (4.7)
$–NO_2$	117 (187)	19 (27)	28 (49)	12 (16)	3.6 (5.9)
–CN	129 (102)	24 (31)	36 (52)	16 (18)	5.1 (7.5)
ϕ-CN	33 (17)	12 (36)	20 (64)	6 (15)	2.3 (8.2)
halogen increments:					
–F	136 (237)	25 (31)	37 (44)	18 (23)	6.7 (8.6)
ϕ-F	163 (30)	10 (23)	16 (53)	7 (10)	2.3 (5.4)
–Cl	180 (236)	13 (19)	19 (32)	9 (13)	3.1 (4.6)
1-Cl	158 (241)	11 (14)	21 (22)	6 (10)	2.4 (3.1)
2-Cl	56 (214)	15 (16)	19 (22)	12 (12)	3.3 (3.6)
3-Cl	45 (158)	20 (30)	31 (51)	11 (19)	4.9 (7.2)
ϕ-Cl	213 (260)	13 (22)	21 (42)	9 (12)	2.6 (4.1)
–Br	227 (300)	13 (18)	20 (28)	9 (13)	2.2 (3.8)
–I	38 (56)	19 (17)	38 (25)	8 (9)	4.3 (3.8)
ϕ-I	36 (38)	9 (17)	13 (36)	7 (8)	1.7 (3.4)

Table 2 (Continued)

structural group	no. of compds used	av abs error (K)	std dev (K)	median abs error (K)	av percent error
sulfur increments:					
–SH	45 (30)	10 (17)	14 (30)	7 (9)	2.3 (3.8)
ϕ-SH	24 (6)	9 (57)	14 (80)	6 (38)	1.9 (8.5)
–S–	79 (105)	17 (29)	24 (66)	11 (14)	3.3 (7.0)
–S_r–	84 (99)	19 (20)	27 (29)	14 (15)	3.9 (4.0)
>SO	4 (2)	41 (29)	48 (39)	52 (29)	9.5 (5.7)
>SO_2	15 (5)	39 (20)	45 (25)	43 (15)	8.0 (4.3)
>CS	2 (3)	73 (200)	73 (228)	73 (200)	17.0 (60.3)
>C_rS	2 (1)[b]	38 (2)	38 (–)	38 (2)	7.1 (0.3)
phosphorus increments					
–PH_2	1 (2)[b]	0 (6)	– (6)	0 (6)	0.0 (1.8)
>PH	1 (2)[b]	0 (22)	– (23)	0 (22)	0.0 (7.8)
>P–	10 (10)	17 (33)	22 (43)	17 (30)	3.4 (8.2)
>PO–	9 (11)	25 (92)	35 (121)	22 (66)	5.1 (18.6)
silicon increments:					
>SiH–	9 (4)	27 (20)	56 (26)	6 (17)	6.7 (4.8)
>Si<	70 (75)	17 (36)	27 (61)	11 (16)	4.2 (8.1)
>Si_r<	2 (8)	9 (61)	9 (75)	9 (46)	2.0 (14.4)
miscellaneous increments:					
>B–	9 (9)	17 (40)	24 (42)	9 (44)	4.5 (9.6)
–Se–	1 (3)[b]	0 (10)	– (10)	0 (10)	0.0 (2.4)
>Sn<	6 (6)	33 (35)	63 (42)	7 (35)	7.3 (7.0)

[a] Values not in parentheses refer to Aldrich data; those in parentheses, to HODOC data. [b] When only one compound containing a particular group is present in a data set, the standard deviation is not defined.

Table 3. Statistical Results of Fitting Procedure; Set S_1 Group Increment Values

	Data Sets							
	Aldrich				HODOC			
statistic	complete set	unextrapolated	extrapolated	hydrocarbons	complete set	unextrapolated	extrapolated	hydrocarbons
no. of compds used	4426	2664	1762	374	6584	2021	4563	1054
median error (K)	0.0	0.0	0.4	–0.3	–1.0	–0.2	–1.4	4.0
av abs error (K)	15.5	13.5	18.5	13.1	20.4	23.9	18.9	13.5
std dev (K)	24.6	22.3	27.8	25.0	38.1	51.0	30.7	22.4
std dev, equiv Gaussian (K)	11.3	10.0	14.9	11.0	13.6	11.5	14.7	11.3
excess over Gaussian in tails (%)	18.5	18.2	13.7	7.3	17.1	20.6	15.0	13.7
median abs error (K)	9.7	8.3	12.6	8.2	11.9	10.7	12.4	9.1
av % error (for deg K)	3.2	3.0	3.5	2.7	4.3	5.9	3.6	2.8

It seems reasonable that a "good" correlation method would yield a Gaussian distribution of residuals when applied to sets of experimental data having random (Gaussian) error distributions. The Aldrich and most of the HODOC data likely fall into this category. Consequently, the outlier tails appearing in Figure 3 very probably arise from the method itself rather than from the data. Even though the method yields remarkably accurate boiling points for a large number of diverse compounds, the existence of these outlier tails show that it frequently fails unpredictably, and therefore should be used with caution.

We feel that the results obtained here are close to the limit of what can be achieved with a simple general purpose group method. These group values provide a simple means of roughly estimating and extrapolating boiling points and vapor pressures for a wide range of organic substances. In our view, the principal deficiency of this method is not a higher than desirable average error, but the possibility of producing a highly erroneous value. Reducing the number of "outliers" or even establishing a means of identifying structures whose estimates are (and are not) subject to large errors is needed before one can really have confidence in estimates and their stated error limits of group-based methods.

One possible way to improve predictability would be to consider interactions between groups. An example of this approach is the UNIFAC group contribution method[4] originally developed for the calculation of vapor–liquid equilibria of liquid mixtures. Parametrization of this method with larger data bases might improve its accuracy and range of applicability. One might expect that any large errors from such more realistic physical models would be easier to interpret, and then correct. Less physically-based methods such as molecular connectivity[9] represent another approach. At present, their main disadvantage appears to be a lack of generality.

ACKNOWLEDGMENT

This work was supported in part by the Environmental Protection Agency (Grant IAG-DW-13933291-01-6).

APPENDIX: EXTRAPOLATION OF EXPERIMENTAL BOILING POINTS TO 1-ATM PRESSURE

Many of the boiling points in our data sets were measured at pressures of less than 1 atm. Three different vapor pressure equations were tested for their ability to extrapolate this data to normal boiling points: Lee–Kesler,[8] Gomez–Thodes,[10] and Antoine (see ref 3, eqs 14–25). To test these methods, we used a set of 1860 compounds from the HODOC data set, each of which had boiling points reported at two different pressures. Two such boiling points, extrapolated to 1 atm, should yield the same normal boiling point. For the different

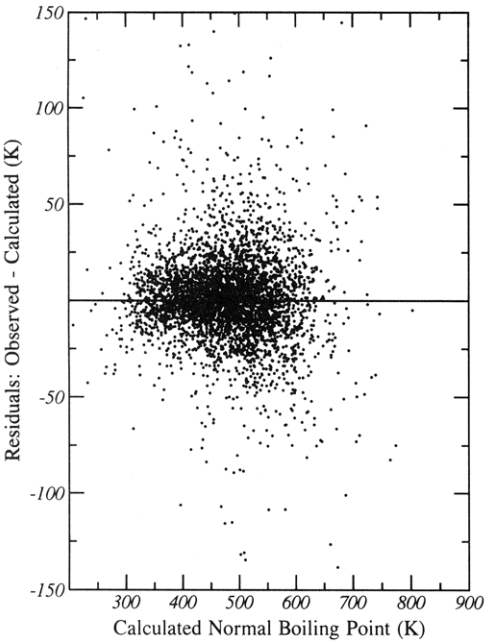

Figure 2. Scatter plot of residuals as a function of calculated boiling point for the Aldrich data set using increment set S_1, along with eq 2, which corrects for overprediction of boiling points at higher temperatures. Without the correction this figure would look like Figure 1 except that the scatter would be smaller as a result of the refitted increment values.

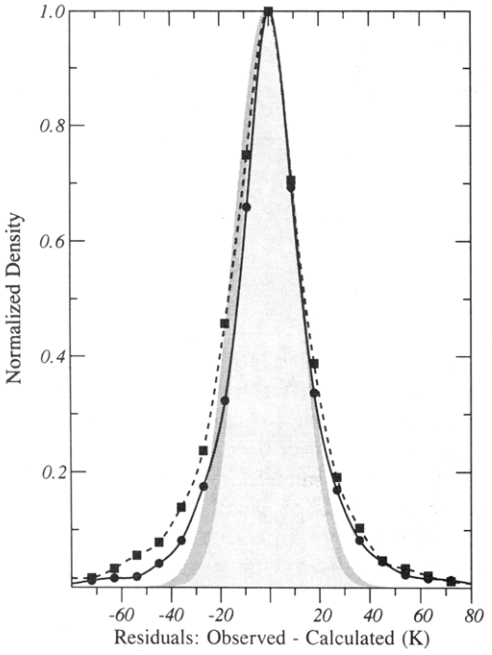

Figure 3. Histograms showing normalized frequencies of residuals, T_b(obsd) – T_b(calcd), for Aldrich (circles) and HODOC (squares) data using increment set S_1 and correction eq 2. The ordinate of each point is proportional to the number of residuals falling in a 9 K interval centered about that point. Shaded areas are Gaussian distributions having the same width at half-height as the observed distributions and centered on the median errors; the wider Gaussian applies to the HODOC data.

methods, we obtained the following average absolute differences between the two extrapolated values: Lee–Kesler, 12.0 K; Gomez–Thodos, 13.1 K; and Antoine, 18.0 K. Since the Lee–Kesler equation gave somewhat better results than the Gomez–Thodos, and was computationally simpler, we chose it as our extrapolation method.

For arbituary pressure P and temperature T, and for P = 1 atm and $T = T_e$ (extrapolated normal boiling point), The

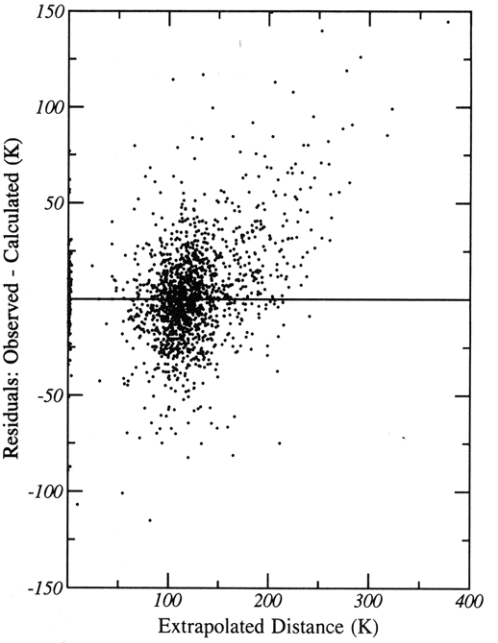

Figure 4. Residuals as a function of extrapolated distance for Aldrich data. These were calculated using increment set S_1 and correction eq 2.

Lee–Kesler equation (see ref 2, eq 7–2.6) takes the following forms:

$$\ln P = \ln P_c + f^{(0)}(T/T_c) + \omega f^{(1)}(T/T_c)$$

$$0 = \ln P_c + f^{(0)}(T_e/T_c) + \omega f^{(1)}(T_e/T_c) \qquad \text{(A.1)}$$

where P_c is the critical pressure (atm), T_c the critical temperature, and ω the acentric factor. The functions $f^{(0)}$ and $f^{(1)}$ are

$$f^{(0)}(\theta) = 5.92714 - 6.09648/\theta - 1.28862 \ln \theta + 0.169347\theta^6$$

$$F^{(1)}(\theta) = 15.2518 - 15.6875/\theta - 13.4721 \ln \theta + 0.435770\theta^6 \qquad \text{(A.2)}$$

where $\theta = T_e/T_c$. For ω we used (see ref 2, eq 2–3.4) the equation

$$\omega = -(\ln P_c + f^{(0)}(\theta))/f^{(1)}(\theta) \qquad \text{(A.3)}$$

Subtracting the second equation in (A.1) from the first, we obtain

$$f^{(0)}(T\theta/T_c) + \omega f^{(1)}(T\theta/T_c) - f^{(0)}(\theta) - \omega f^{(1)}(\theta) - \ln P = 0 \qquad \text{(A.4)}$$

We used this expression to extrapolate experimental boiling points to 1-atm pressure. Given a boiling point T, measured at pressure P, we solved eq A.4 for the extrapolated normal boiling point T_e.

Note that (A.4) depends on P_c and θ. We used the group contributions derived by Joback and Reid[5] to estimate these quantities. Their values, translated to our groups, are given in Table 1. Fourteen of our groups (set S_{1b}) did not correspond to any of the Joback and Reid groups, or combinations thereof. For these groups, we did not try to develop P_c and θ contributions. Consequently, we could not extrapolate boiling points for compounds containing these groups.

In the Aldrich data base, 1762 boiling points required extrapolation. For these compounds, the average extrapolated distance, $\langle d_e = T_e - T \rangle$ was 110 K. Figure 4 shows a plot of the residuals δ as a function of d_e. These were calculated using the final increment set S_1 and the high boiling point correction eq 2. The average value of these residuals is 2.5 K. This minor offset arises from a small excess of positive residuals for d_e values above approximately 150 K.

REFERENCES AND NOTES

(1) (a) Benson, S. W.; Buss, J. H. *J. Chem. Phys.* **1958**, *29*, 546. (b) Benson, S. W. *Thermochemical Kinetics*, 2nd ed.; Wiley: New York, 1976.

(2) (a) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids*; 4th ed.; McGraw-Hill: New York 1987. (b) Horvath, A. L. *Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds*; Elsevier: Amsterdam, The Netherlands, 1992.

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

(4) (a) Jensen, T.; Fredenslund, A.; Rasmussen, P. *Ind. Eng. Chem. Fundam.* **1981**, *20*, 239. (b) Yair, O. B.; Fredenslund, A. *Ind. Eng. Chem. Process Des. Dev.* **1983**, *22*, 433.

(5) Joback, K. G.; Reid, R. C. *Chem. Eng. Commun.* **1987**, *57*, 233.

(6) *Aldrich Handbook of Fine Chemicals*; Aldrich Chemical Co.: Milwaukee, WI, 1990.

(7) *Handbook of Data of Organic Compounds*; J. Graselli, Ed.; CRC Press: Boca Raton, FL, 1990.

(8) (a) Lee, B. I.; Kesler, M. G. *AIChE J.* **1975**, *21*, 510. (b) See ref 2a, Chapter 7, for a summary of this method.

(9) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306.

(10) (a) Gomez-Nieto, M.; Thodos, G. *Ind. Eng. Chem. Fundam.* **1977**, *16*, 254. (b) Gomez-Nieto, M.; Thodos, G. *Ind. Eng. Chem. Fundam.* **1978**, *17*, 45. (c) Gomez-Nieto, M.; Thodos, G. *Can. J. Chem. Eng.* **1977**, *55*, 445. (d) See ref 2a, p 209, for a summarization of this vapor pressure estimation method.