Automated Additive Modeling Techniques Applied to Thermochemical Property Estimation

GEORGE W. ADAMSON* and DAVID BAWDEN†

Postgraduate School of Librarianship and Information Science, Sheffield University,
Sheffield S10 2TN, England

Received April 22, 1980

Automated additive modeling techniques have been applied to the estimation of heats of vaporization. Computer-readable representations of chemical structures were analyzed algorithmically to derive structural groupings to be used in the model. These were then input as variables in multiple regression analyses, which gave the group contributions to be used in estimation. The method gave good estimations for various structural types and could be applied to a variety of thermochemical properties. It would be compatible with computerized data compilation activities.

INTRODUCTION

Additivity schemes for the estimation of thermochemical quantities have had wide usage over a long period .[1-3] Such schemes are entirely empirical, relying on analyses of experimental data to give values for structural fragments, which may be summed to yield estimated property values for other compounds. The structural fragments may be of constant size, in, e.g., atom- or bond-additivity schemes, or may vary in size, as in the various group additivity schemes.

The computerized perception of structural features used in such schemes for automated property estimation has been reported for various thermochemical properties[4-6] and analogously for partition coefficient.[7] The work reported here involves the automatic generation of structural descriptors of various types from a computer-readable representation of the chemical structure diagram and their use as variables in multiple regression analyses. This allows very rapid and flexible analyses of data sets, so as to determine the most appropriate model for a particular case. The procedure thus involves the generation and investigation of new additive models, rather than automation of property prediction based on an existing model. Techniques of this sort have been applied in structure-activity studies of various physicochemical and biological properties, using multiple regression[8-13] and other multivariate techniques.[14-16]

Essentially the same techniques may be used to investigate additive schemes for thermochemical property estimation. A set of compounds, for which experimental values of the relevant property are available, have their structures encoded in some computer-readable form. These representations are analyzed automatically to derive the frequencies of occurrence of structural features or fragments. The fragments are then input directly as variables in a multiple regression analysis. The resulting coefficients may be used as additive values for subsequent estimations.

There are several features about this type of analysis which enable it to usefully augment the conventional ways of deriving additive values. The use of automatic data-handling techniques makes possible examination of much larger data sets than could otherwise be easily dealt with. These techniques could also be readily integrated with computerized compilation and correlation of thermochemical data.[17] It can be advantageous to use subsets of available data, e.g., compounds containing a particular functional group, to derive values for additive estimations. This is particularly so when values for particular groups can vary according to the molecular environment: for example, $-CH_2-$ incremental values can alter with the presence of a ketonic group.[3] The members of subsets could easily be identified by substructure search.

With the aid of automated structure handling, the comparison of estimation from analyses of different sets of compounds can be made considerably easier.

The algorithms which generate structural features may easily be modified to generate varying types of fragments. This enables analyses of a data set and subsequent use of the results in estimation in terms of different kinds of structural features. This facility may be used in either of two ways. First, analyses corresponding to different additivity models, e.g., atom and bond additives, can be compared for a given data set. This could be of particular interest, if more complicated additive units, either atom or bond centered, were used. Such analyses, very time consuming if carried out nonalgorithmically, can be highly revealing of factors influencing structure-property relationships.[11,12] Second, the facility can be used within the framework of the group additivity model to investigate rapidly and conveniently the effects of varying the groups considered as additive units. This can be particularly valuable in deciding the necessity for including interaction terms in models to be used for estimation. This application is further discussed.

These techniques could, of course, be used with any additive-constitutive molecular properties, including many thermochemical properties.

METHOD

The structures of the compounds were encoded in Wiswesser Line Notation (WLN).[18] This notation is widely used in computerized chemical information systems and has been adopted for a computer-based compilation of thermochemical data.[19] It is particularly convenient, using WLN, to automatically generate structural features of the sort used in group contribution schemes: functional groups, ring systems, interaction terms, etc.

The WLN representations were fragmented by computer program to generate the structural groups to be used in the analyses. The programs used were open ended in that they did not prespecify all the groups to be used. Rather they used the features of the notation to derive relatively simple hydrocarbon fragments and functional groups,[8] and ring systems and their substituents and substituent interactions,[9,10] from the structures being analyzed. The programs were written so as to allow flexibility in choice of groups to be included. The frequencies of occurrence of these groups were the input variables in a standard stepwise multiple regression analysis program.[20] The coefficients from this analysis were the group values in the additive model and were used in subsequent estimations.

In comparing the differences between models, i.e., differences caused by choosing varying groups, the significance of the difference between the two overall regressions was assessed statistically.[21] Also the significance of each coefficient value

* To whom correspondence should be addressed at ICI Pharmaceuticals Division, Alderley Park, Macclesfield, Cheshire, England.
† Pfizer Central Research, Sandwich, Kent, England.

0095-2338/80/1620-0242$01.00/0

could be assessed from the *t* statistic calculated by the regression program. These statistical tests are of importance in evaluating the effect of altering the model.

DATA

The method was evaluated by using experimental values for heat of vaporization, a property which has been accurately measured for a variety of structural types.[17] The values used (measured in kilocalories per mole at 25 °C and 1 atm pressure) were taken from a standard compilation of thermochemical data.[3] The compounds included alkanes, alkenes, alcohols, ketones, benzenes, and pyridines, and are listed with their property values in Table I. The restriction to strictly comparable data measured under these conditions without any extrapolation or estimation procedures, although necessary to avoid ambiguities in the interpretation of correlation studies such as this, greatly reduced the amount of usable data.

RESULTS

This collection of data was analyzed in three stages. First, a set of simple aliphatic compounds was examined, to determine the success of automatic additive modeling in correlating data of this sort. Second, a set of aromatic compounds was investigated, to establish the ability of the methods to deal rapidly and systematically with complex data sets and quantitatively test alternative models. Third, a data set consisting of the first two sets combined was analyzed in order to test the method's ability to deal with a set of structurally diverse compounds.

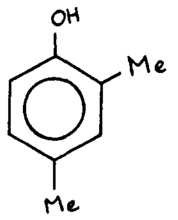
1. Aliphatic Structures. A set of 96 aliphatic structures was examined, comprising 37 alkanes, 36 alkenes, 12 alcohols, and 11 ketones. The fragment generation algorithm produced a set of 11 variables: hydrocarbon fragments, unsaturated units, and OH and C=O units,[8] which were input to the stepwise multiple regression program. The overall result was

$$R = 0.999,\ r = 0.22,\ F = 7.000,\ df = 85$$

for a range of 11.6 units property values, a highly significant correlation.

The structural groups in this analysis, together with their regression coefficients and *t* statistics, are listed in Table II. All coefficients except -CH- are significant at the 1% level. The interpretation of the coefficients, in terms of effects of carbon branching and introduction of oxygen functionality on heat of vaporization, is in accord with those noted elsewhere.[3]

In order to assess how effective the coefficients from this analysis would be likely to be for estimation, their success in quantitatively accounting for the effects of chain branching was assessed. In Table III is shown the observed difference in heat of vaporization between two pairs of isomers, contrasted with an estimation using the results of this analysis and estimations made by the Greenshield–Rossini and Laidler–Lovering structural contribution schemes.[3] The estimation from the additive model due to this analysis gives a closer agreement with the observed values than the other methods, despite its automatic derivation from a data set of diverse aliphatic compounds. *n*-Pentane and isopentane were included in the data set from which the group values were calculated, but neopentane was not. The comparison was repeated by using the "hold-out" method, with a "training set" from which all three compounds had been omitted. The *n*-pentane/isopentane difference is unaffected. The *n*-pentane/neopentane difference is worse predicted (due to the quaternary carbon coefficient with the highest standard error in the set), but is still equivalent to the Greenshields–Rossini prediction and superior to the Laidler–Lovering.



Set 1

two Me
one OH

Set 2

two Me
one OH
one OH-o-Me

Set 3

two Me
one OH
one OH-o-Me
one Me-m-Me

Set 4

two Me
one OH
one OH-o-Me
one OH-p-Me

Set 5

two Me
one OH
one OH-o-Me
one Me-m-Me
one OH-p-Me

Figure 1. Example of structural group derivation.

2. Aromatic Structures. A set of 41 aromatic compounds was examined, comprising 31 benzene derivatives and 10 pyridine derivatives. The main aim was to determine the necessity for inclusion of substituent interaction terms in the additive model. Five analyses were carried out, with these structural units included:

set 1 number and type of substituents only
set 2 as 1, also including ortho interaction terms
set 3 as 2, also including meta interaction terms
set 4 as 2, also including para interaction terms
set 5 as 3, also including para interaction terms

(The N atom of the pyridines was treated as a substituent within the ring.) An example of the structural groups derived in these ways is shown in Figure 1. It should be noted that WLN is a particularly convenient representation for deriving interaction terms of this sort, because of its use of locants explicitly denoting substituent position.

The overall results of the stepwise multiple regression analyses for these sets of groups are shown in Table IV. Highly significant correlations, as shown by the *F* values, indicate the success of these additive models. Set 2 gives a correlation significantly better at 5% than that with set 1. Set 3 gives an improved correlation, significant at the 5% level, compared with set 2. Neither set 4 nor set 5 gives a correlation significantly different from that with set 2.

The higher variable:observation ratio means that these results must be treated with caution. However, they indicate that ortho interactions exert an important effect on heat of vaporization, meta interactions are also important, while the effect of para interactions is negligible. This ortho effect has been noted previously, and ortho correction factors are incorporated in additive estimation schemes.[1-3] It appears from these results that meta interactions could also usefully be taken into account.

Diverse Structure. The two data sets described above were combined, giving a total set of 137 aliphatic, aromatic, and heteroaromatic compounds. This data set was analyzed by using two sets of structural features: Set 3A included hydrocarbon fragments, unsaturated units, simple functionalities, and whole ring systems (15 variables). Set 3B was the same as 3A, but also included terms representing substituent interactions on rings (39 variables).

The stepwise regression analysis using the groups of set 3B as variables gave a correlation superior at the 1% level to that

Table I. Data Set Used

structure number	structure	heat of vaporization, kcal/mol	structure number	structure	heat of vaporization, kcal/mol
1	$CH_3(CH_2)_2CH_3$	5.02	71	$CH_3(CH_2)_3CH=CH_2$	7.34
2	$CH_3CH(CH_3)CH_3$	4.61	72	$CH_3CH_2CH_2CH=CHCH_3$-*cis*	7.54
3	$CH_3(CH_2)_3CH_3$	6.39	73	$CH_3CH_2CH_2CH=CHCH_3$-*trans*	7.56
4	$CH_3CH_2CH(CH_3)_2$	6.03	74	$CH_3CH_2CH=CHCH_2CH_3$-*cis*	7.49
5	$CH_3(CH_2)_4CH_3$	7.54	75	$CH_3CH_2CH=CHCH_2CH_3$-*trans*	7.56
6	$CH_3(CH_2)_2CH(CH_3)_2$	7.14	76	$CH_2=C(CH_3)CH_2CH_2CH_3$	7.31
7	$CH_3CH_2CH(CH_3)CH_2CH_3$	7.24	77	$CH_2=CHCH(CH_3)CH_2CH_3$	6.85
8	$CH_3CH(CH_3)CH(CH_3)_2$	6.96	78	$CH_2=CHCH_2CH(CH_3)_2$	6.88
9	$CH_3CH_2C(CH_3)_3$	6.62	79	$CH_3CH_2CH=CH(CH_3)_2$	7.57
10	$CH_3(CH_2)_5CH_3$	8.74	80	$CH_3CH=C(CH_3)CH_2CH_3$-*cis*	7.69
11	$CH_3(CH_2)_3CH(CH_3)_2$	8.33	81	$CH_3CH=C(CH_3)CH_2CH_3$-*trans*	7.51
12	$CH_3(CH_2)_2CH(CH_3)CH_2CH_3$	8.39	82	$CH_3CH=CHCH(CH_3)_2$-*cis*	7.06
13	$CH_3CH_2CH(CH_2CH_3)_2$	8.42	83	$CH_3CH=CHCH(CH_3)_2$-*trans*	7.18
14	$CH_3(CH_2)_2C(CH_3)_3$	7.75	84	$CH_2=C(CH_3)CH(CH_3)_2$	6.99
15	$CH_3CH_2CH(CH_3)CH(CH_3)_2$	8.18	85	$CH_2=CHC(CH_3)_3$	6.38
16	$CH_3CH(CH_3)CH_2CH(CH_3)_2$	7.86	86	$CH_3C(CH_3)=C(CH_3)_2$	7.80
17	$CH_3CH_2C(CH_3)_2CH_2CH_3$	7.89	87	$CH_3(CH_2)_4CH=CH_2$	8.52
18	$CH_3C(CH_3)_2CH(CH_3)_2$	7.66	88	$CH_3CH_2CH=C(CH_3)CH_2CH_3$-*cis*	8.73
19	$CH_3(CH_2)_6CH_3$	9.92	89	$CH_3CH_2CH=C(CH_3)CH_2CH_3$-*trans*	8.58
20	$CH_3(CH_2)_4CH(CH_3)_2$	9.48	90	$CH_2=C(CH_3)CH_2CH(CH_3)_2$	7.93
21	$CH_3(CH_2)_3CH(CH_3)CH_2CH_3$	9.52	91	$CH_2=CHCH_2C(CH_3)_3$	7.47
22	$CH_3CH_2CH_2CH(CH_3)CH_2CH_2CH_3$	9.48	92	$CH_3CH(CH_3)CH=C(CH_3)_2$	8.22
23	$CH_3CH_2CH_2CH(CH_2CH_3)_2$	9.48	93	$CH_3CH=CHC(CH_3)_3$-*cis*	7.81
24	$CH_3(CH_2)_3C(CH_3)_3$	8.91	94	$CH_3CH=CHC(CH_3)_3$-*trans*	7.87
25	$CH_3CH_2CH_2CH(CH_3)CH(CH_3)_2$	9.27	95	$CH_3CH_2CH=CHC(CH_3)_3$-*cis*	8.88
26	$CH_3CH_2CH(CH_3)CH_2CH(CH_3)_2$	9.03	96	$CH_3CH_2CH=CHC(CH_3)_3$-*trans*	8.91
27	$CH_3CH(CH_3)CH_2CH_2CH(CH_3)_2$	9.05		Benzene Derivatives	
28	$CH_3CH_2CH_2C(CH_3)_2CH_2CH_3$	8.97	97	H	8.09
29	$CH_3CH_2CH(CH_3)CH(CH_3)CH_2CH_3$	9.32	98	CH_3	9.08
30	$CH_3CH_2CH(CH_2CH_3)CH(CH_3)_2$	9.21	99	CH_2CH_3	10.10
31	$CH_3CH_2C(CH_2CH_3)_2CH_3$	9.08	100	1,2-$(CH_3)_2$	10.38
32	$CH_3CH_2CH(CH_3)C(CH_3)_3$	8.82	101	1,3-$(CH_3)_2$	10.20
33	$CH_3C(CH_3)_2CH_2CH(CH_3)_2$	8.40	102	1,4-$(CH_3)_2$	10.13
34	$CH_3CH_2C(CH_3)_2CH(CH_3)_2$	8.90	103	$CH_2CH_2CH_3$	11.05
35	$CH_3CH(CH_3)CH(CH_3)CH(CH_3)_2$	9.01	104	$CH(CH_3)_2$	10.79
36	$CH_3(CH_2)_7CH_3$	11.10	105	1-CH_2CH_3,2-CH_3	11.40
37	$CH_3(CH_2)_8CH_3$	12.28	106	1-CH_2CH_3,3-CH_3	11.21
38	CH_3OH	8.94	107	1-CH_2CH_3,4-CH_3	11.14
39	CH_3CH_2OH	10.18	108	1,2,3-$(CH_3)_3$	11.73
40	$CH_3(CH_2)_2OH$	11.34	109	1,2,4-$(CH_3)_3$	11.46
41	$CH_3CH(OH)CH_3$	10.90	110	1,3,5-$(CH_3)_3$	11.35
42	$CH_3(CH_2)_3OH$	12.50	111	1-OH_3,-CH_3	14.75
43	$CH_3CH(CH_3)CH_2OH$	12.15	112	1-OH,2-CH_2CH_3	15.20
44	$CH_3CH_2CH(OH)CH_3$	11.89	113	1-OH,3-CH_2CH_3	16.30
45	$CH_3C(CH_3)_2OH$	11.14	114	1-OH,2,4-$(CH_3)_2$	15.74
46	$CH_3(CH_2)_4OH$	13.61	115	F_6	8.53
47	$CH_2CH_2CH_2CH(OH)CH_3$	12.56	116	F_5	8.65
48	$CH_3(CH_2)_5OH$	15.00	117	1,2-F_2	8.65
49	$CH_3(CH_2)_6OH$	16.20	118	1,3-F_2	8.29
50	CH_3COCH_3	7.37	119	1,4-F_2	8.51
51	$CH_3CH_2COCH_3$	8.34	120	1,2-Cl_2	11.4
52	$CH_3CH_2CH_2COCH_3$	9.14	121	1,3-Cl_2	11.1
53	$CH_3CH(CH_3)COCH_3$	8.82	122	F	8.27
54	$CH_3CH_2CH_2COCH_2CH_3$	10.01	123	Cl	9.63
55	$CH_3CH_2COCH(CH_3)_2$	9.51	124	1-CH_3, F_5	9.78
56	$CH_3CH_2COC(CH_3)_3$	10.12	125	1-CH_3,4-F	9.42
57	$CH_3CH(CH_3)COCH(CH_3)_2$	9.93	126	1-Cl,2-CH_2CH_3	11.3
58	$CH_3C(CH_3)_2COCH(CH_3)_2$	10.35	127	1-Cl,4-CH_2CH_3	11.5
59	$CH_3(CH_2)_3CO(CH_2)_3CH_3$	12.59			
60	$CH_3C(CH_3)_2COC(CH_3)_3$	10.84		Pyridine Derivatives	
61	$CH_3CH_2CH=CH_2$	4.92	128	H	9.61
62	$CH_3CH=CHCH_3$-*cis*	5.40	129	2-CH_3	10.15
63	$CH_3CH=CHCH_3$-*trans*	5.16	130	3-CH_3	10.62
64	$CH_2=C(CH_3)_2$	4.92	131	4-CH_3	10.83
65	$CH_3CH_2CH_2CH=CH_2$	6.09	132	2,3-$(CH_3)_2$	11.70
66	$CH_3CH_2CH=CHCH_3$-*cis*	6.41	133	2,4-$(CH_3)_2$	11.42
67	$CH_3CH_2CH=CHCH_3$-*trans*	6.38	134	2,5-$(CH_3)_2$	11.43
68	$CH_2=C(CH_3)CH_2CH_3$	6.18	135	2,6-$(CH_3)_2$	11.01
69	$CH_2=CHCH(CH_3)_2$	5.70	136	3,4-$(CH_3)_2$	12.38
70	$CH_3CH=C(CH_3)_2$	6.47	137	3,5-$(CH_3)_2$	12.04

Table II. Group Coefficients from Analysis of 96 Aliphatic Compounds[a]

structural feature	regression coefficient	*t* statistic (85 degrees of freedom)
$-CH_3$	1.61	32.08
$-CH_2-$	1.12	59.00
$-CH-$	0.22	2.38
$-C-$	−0.87	5.80
$-OH$	7.45	110.30
$-C=O$	3.41	40.08
$CH_2=CH-$	2.33	30.45
$-CH=CH-$	2.12	22.89
$CH_2=C-$	1.82	14.78
$-C=CH-$	1.62	10.49
$-C=C-$	1.37	4.57

[a] A *t* value of 2.63 shows a coefficient significantly different from zero at the 1% level.

Table III. Comparisons of Chain-Branching Estimation

H_V value	H_V(i) − H_V(ii)[a]	H_V(iii) − H_V(i)[a]
observed	−0.42	−1.05
these analyses	−0.41 (estimation)	−1.02 (estimation)
	−0.42 (hold out)	−0.80 (hold out)
Greenshield-Rossini	−0.19	−1.25
Laidler-Lovering	−0.32	−0.51

[a] Kilocalories/mole.

using set 3A variables. This emphasizes the importance of substituent interaction terms, even in an additive model derived from a diverse data set. The overall result from the set 3B analysis was $R = 0.997$, $r = 0.21$, $f = 445$, $df = 99$ for a range of property values of 11.7 units kcal/mol.

Of the coefficients for ring substituent interactions which are significantly different from zero at the 5% level, three are ortho and two are meta interactions. The only significant variable involving a para interaction is perfectly correlated in the data set with an ortho interaction and so is inseparable from it. This also illustrates the relative importance of the ortho meta, and para interactions.

The group values from this analysis were used to estimate heat of vaporization values for six compounds, using the hold-one-out technique. The compound under test was omitted from the set, the analysis was repeated, and the resulting coefficients were used for estimation. The estimated values were compared with the experimental values and with estimations from other techniques.[3] These were Chen's equation, using critical parameter data; Fishtine's equation, using boiling point data; and three methods based on additive structural contributions: the Laidler-Lovering method, the method of CH_2 increments, and Wright's method. The values are shown in Table V. In general the results show that, although for each compound one of the other methods gives a better estimated value, overall the additive group values derived here are consistently accurate. The CH_2-incremental technique has a similar performance, as might be expected since it is also an example of additive modeling. However, this technique is limited to analyses of homologous series, hence its inability to deal with chlorobenzene. This shows the capacity of analyses of this type, using structurally diverse data sets as input, to generate additive models with good predictive ability.

DISCUSSION

The work described above shows that automated additive modeling, using computerized techniques of chemical structure handling together with multiple regression analysis, is a feasible, and potentially valuable, aid to correlation and estimation of thermochemical properties. Models derived in this way give highly significant correlations for heat of vaporization and accurate estimations of overall property values and contributing structural effects.

Some of the group contributions derived from the diverse data set differ from those produced by the analysis of aliphatic compounds alone. The calculated differences between the pentane isomers used in Table III are almost unaffected by this. However, the estimated heats of vaporization for the individual isomers are much closer to those observed when calculated by using values obtained from the analysis of aliphatic compounds. This illustrates the point that in empirical estimations, better results are obtained within classes of similar structures, due to the constancy of environmental and proximity effects.

These methods can deal with large sets of structures and property values and employ widely used techniques for computerized structure handling and statistical analysis. They could be made compatible with computerized data compilation methods and could be used as an adjunct to such compilations: for example, for checking suspect data and providing estimations for missing values. The analyses of the aromatic data set show how an automated procedure for structural feature derivation, combined with an appropriate statistical analysis technique, allows a systematic, detailed examination of a complex data set. The rapidity and flexibility of such analyses make this feasible on a routine basis. In particular, the determination of the significance of inclusion in the model of detailed structural parameters, such as intersubstituent interaction, can be valuable. This aids the choice on a firm statistical basis of those parameters, from the many possible, which are most effective in correlation and estimation. This technique, with open-ended fragment generation applied to particular data sets, could usefully complement techniques which use computerized structure handling to automate estimation schemes with fixed sets of parameters.[5,6]

Techniques have been described which generate fragments from connection tables which would be suitable for correlation and estimation,[22,23] if the data base in use represented the chemical structures in that form. The method of analysis could also be applied to other thermodynamic properties.

ACKNOWLEDGMENT

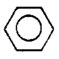
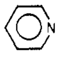
We thank Dr. I. McLure and Dr. J. B. Pedley for helpful

Table IV. Overall Regression Results for Analysis of 41 Benzene and Pyridine Derivatives[a]

structural features	no. of structural features[b]	no. of variables included[c]	degrees of freedom	multiple correlation coefficient	residual error	*F* value
set 1	9	8 + constant	32	0.991	0.29	219.23
set 2	18	17 + constant	23	0.998	0.15	337.22
set 3	26	23 + constant	17	0.999	0.10	369.01
set 4	25	23 + constant	17	0.999	0.15	369.01
set 5	33	29 + constant	11	>0.999	0.07	379.03

[a] Number of structures = 41; range of property values = 8.21. [b] I.e., no. of variables input into stepwise regression program. [c] I.e., no. of variables included by stepwise regression program.

Table V. Group Coefficients from 3B Analysis of 137 Compounds of Diverse Structure[a]

structural feature	regression coefficient	*t* statistic (99 degrees of freedom)
CH_3-	0.97	5.94
$-CH_2-$	1.12	63.56
$-\overset{\vert}{\mathrm{C}}\mathrm{H}-$	0.86	5.04
$-\overset{\vert}{\underset{\vert}{\mathrm{C}}}-$	0.40	1.21
$-OH$	6.84	81.52
$-\overset{\vert}{\mathrm{C}}{=}\mathrm{O}$	3.41	42.75
F–	0.26	2.01
Cl–	1.59	5.92
$CH_2{=}CH-$	1.68	9.11
$-CH{=}CH-$	2.11	24.45
$\mathrm{CH_2{=}}\overset{\vert}{\mathrm{C}}-$	1.81	15.72
$-\mathrm{CH{=}}\overset{\vert}{\mathrm{C}}-$	2.25	11.41
$-\underset{\vert}{\mathrm{C}}{=}\underset{\vert}{\mathrm{C}}-$	2.63	6.58
	6.74	32.21
	8.27	24.62
o-Me,ring N	-0.35	2.38
m-Me,ring N	0.20	1.33
p-Me,ring N	0.26	1.35
o-Me,-Me	0.35	2.40
m-Me,Me	0.13	1.01
p-Me,Me	0.09	0.49
o-Me,Et	0.31	1.09
m-Me,Et	0.12	0.42
p-Me,Et	0.05	0.18
m-Me,OH	−1.06	3.60
p-Me,OH	−1.17	3.61[b]
o-Me,OH[b]		
o-Et,OH	−1.73	5.85
m-Et,OH	−0.63	2.12
o-F,F	0.11	0.65
m-F,F	−0.26	1.60
p-F,F	−0.04	0.17
o-Cl,Cl	0.18	0.35
m-Cl,Cl	−0.12	0.25
o-Me,F	0.01	0.06
m-Me, F		
p-Me,F	0.16	0.65
c-Et,Cl	−0.41	1.23
p-Et,Cl	−0.22	0.64

[a] A *t* value of 2.62 shows a coefficient significantly different from zero at the 1% level. [b] Perfectly correlated structural feature.

discussions and two reviewers for suggesting improvements in presentation. D.B. thanks the Department of Education and Science (London) for the award of a postgraduate Research Studentship.

REFERENCES AND NOTES

(1) G. Janz, "Thermodynamic Properties of Organic Compounds", Academic Press, New York, 1967.

(2) S. W. Benson, "Thermochemical Kinetics", Wiley, New York, 1976.

(3) J. D. Cox and G. Pilcher, "Thermochemistry of Organic and Organometallic Compounds", Academic Press, New York, 1970.

(4) W. C. Brasie and D. W. Lion, "Chemical Structure Coding", *Chem. Eng. Prog.* **61**, 102–118 (1965).

(5) N. Jochelson et al., "The Automation of Structural Group Contribution Methods in the Estimation of Physical Properties", *J. Chem. Soc.*, **8** (2), 113–122 (1968).

(6) J. Gasteiger, "Automatic Estimation of Heats of Atomization and Heats of Reaction", *Tetradedron*, **35**, 1419–1426 (1979).

(7) J. T. Chou and P. C. Jurs, "Computer-Assisted Computation of Partition Coefficients from Molecular Structures Using Fragment Constants", *J. Chem. Inf. Comput. Sci.*, **19**, 172–178 (1979).

(8) G. W. Adamson and D. Bawden, "A Method of Structure-Activity Correlation using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **15** (4), 214–220 (1975).

(9) G. W. Adamson and D. Bawden, "An Empirical Method of Structure-Activity Correlation for Polysubstituted Cyclic Compounds using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **16** (3), 161–165 (1976).

(10) G. W. Adamson and D. Bawden, "A Substructural Analysis Method of Structure-Activity Correlation of Heterocyclic Compounds using Wiswesser Line Notation", *J. Chem. Comput. Sci.*, **17** (3), 164–171 (1977).

(11) G. W. Adamson and J. A. Bush, "Method for Relating the Structure and Properties of Chemical Compounds", *Nature (London)*, **248**, 406–408 (1974).

(12) G. W. Adamson and J. A. Bush, "The Evaluation of an Empirical Structure-Activity Relationship for Property Prediction in a structurally Diverse Group of Local Anaesthetics", *J. Chem. Soc., Perkin Trans. 1*, 168–172 (1976).

(13) K. Enslein and P. N. Craig, "A Toxicity Estimation Model", *J. Environ. Pathol. Toxicol.*, **2**, 115–121 (1978).

(14) G. W. Adamson and J. A. Bush, "A Comparison of the Performance of some similarity and dissimilarity measures in the automatic classification of chemical structures", *J. Chem. Inf. Comput. Sci.*, **15**, 55–58 (1975).

(15) K. C. Chu, R. H. Feldman, M. B. Shapiro, G. F. Hazard, R. I. Geran, "Pattern Recognition and Structure-Activity Relationship Studies", *J. Med. Chem.*, **18**, 539–545 (1975).

(16) P. C. Jurs, J. C. Chou, and M. Yuan, "Studies and Chemical Structure–Biological Activity relations using Pattern Recognition", *in* Computer-Assisted Drug Design, E. C. Olsen and R. E. Christoffersen, Eds.; *ACS. Symp. Ser.*, No. **112**, 1979.

(17) J. B. Pedley, "N.P.L. Computer Analyzed Thermochemical Data. Organic and Organometallic Compounds", School of Molecular Science, University of Sussex.

(18) E. G. Smith and P. A. Baker, "The Wiswesser Line Formula Chemical Notation", 3rd ed., Chemical Information Management Inc., Cherry Hill, NJ, 1976.

(19) J. B. Pedley, personal communication.

(20) "Statistical Analysis Mark 2 Applications Package", ICL Technical Publication 4301, International Computer Ltd., London.

(21) G. J. A. Stern, "Using ICL Survey Analysis and Mark 2", London, 1976.

(22) J. E. Crowe, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part I. Non-cyclic Fragments", *J. Chem. Soc. C*, 990–996 (1970).

(23) G. W. Adamson et al., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part V. More Detailed Cyclic Fragments", *J. Chem. Soc., Perkin Trans. 1*, 2071–2076 (1973).