Journal of Chromatography, 101 (1974) 23–31
© Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

CHROM. 7685

COMPUTER APPLICATIONS IN GAS–LIQUID CHROMATOGRAPHY

II. PREDICTION OF RETENTION VOLUME FOR HYDROCARBONS AND HETEROATOMIC ORGANIC MOLECULES USING MOLECULAR PARAMETERS AND MULTIPLE REGRESSION ANALYSIS

C. A. STREULI
Lederle Laboratories Division, American Cyanamid Company, Pearl River, N.Y. 10965 (U.S.A.)
and
M. ORLOFF*
Chemical Research Division, American Cyanamid Company, Stamford, Conn. 06904 (U.S.A.)

(First received January 18th, 1974; revised manuscript received June 11th, 1974)

SUMMARY

An equation is developed by multiple regression analysis which will predict the logarithm of the retention volume, log V_g, for both alkane and aromatic hydrocarbons, based upon molecular weight, π-electron bonding and structural considerations. The importance of each factor depends upon whether the substrate is non-polar, semi-polar or polar. With further refinement of the equation, a new form appears which can be used to predict log V_g for compounds containing heteroatoms.

INTRODUCTION

We showed in Part I[1] that the constants in the well known equation[2] for gas–liquid chromatography (GLC)

$$\log V_g = -\frac{\Delta H_s}{2.3R} \cdot \frac{1}{T} + c_1 \tag{1}$$

are readily evaluated to a reasonable degree of accuracy by a least-squares analysis of three sets of data points. The standard deviation obtained for log V_g comparing results from limited input data and complete GLC data of all the compounds tested[1] was of the order of ± 0.015 (see Table II, Part I). Both $-\Delta H_s/2.3R$ and c_1 are dependent not only on the substrate but also on the molecular structure of the eluting species. Others[3–6] have considered the complexities of the matter, but it is generally conceded that the elution volume depends upon the interacting forces of substrate and individual

[*] Present address: Organic Chemicals Division, American Cyanamid Company, Bound Brook, N.J. 08805 (U.S.A.)

molecule and that these forces are not easily described by a single parameter. Within a homologous series, we find a relation between carbon number and log V_g^2, so obviously molecular weight is one variable, but if a hydrocarbon is aromatic rather than aliphatic the retention values will differ even if the weight is the same; therefore, we might assume that π-bonding also had an effect. Even if two aromatic molecules have nearly the same weight and different electron ring densities, but differing spatial arrangements, they may elute together. Therefore spatial structure must be of some importance. The point is illustrated in Table I for three hydrocarbons.

TABLE I

VALUES OF LOG V_g ON SEMI-POLAR SUBSTRATE* FOR SEVERAL HYDROCARBONS

Compound	*Mol. wt.*	π**	*Steric type***	*T (°K)*	*log V_g*
Undecane	156.2	0.00	a	323	3.7387
				373	2.7505
				423	1.9960
2,6-Dimethylnaphthalene	156.2	3.98	b	323	4.6084
				373	3.5629
				423	2.7644
Biphenyl	154.2	4.38	c	323	4.6018
				373	3.5456
				423	2.7387

* Silicon-containing carborane polymer (Dexsil 300).

** Electron delocalization energy calculated from the Hückel molecular orbital theory.

*** a = Molecule free rotating; b = molecule planar, rigid; c = molecule mixture of rigid planar and free rotating bond.

As a trial approximation, we might assume that both $-\Delta H_s/2.3R$ and c_1 are linear functions of these variables (and perhaps others) and a multiple linear regression solution of an appropriate equation should be capable of predicting log V_g and ultimately t_g for any hydrocarbon compound. With these ideas in mind, the multiple regression analysis was approached.

EXPERIMENTAL

The experimental method and method of calculation for obtaining values of log V_g for all the hydrocarbons have been previously described[1]. Three substrates were used as detailed in Part I[1], and the regression analyses were made by a computer on data produced in the manner described.

RESULTS

After testing a number of computer models with the data for substrate 2, the equation which gave the best results had the general form:

$$\log V_g = [a\,(\text{mol. wt.}) + b\,(\pi) + c\,(\text{st.}) + d]\frac{1}{T} + [a'\,(\text{mol. wt.}) + b'\,(\pi) + c'\,(\text{st.})] \qquad (2)$$

where st. = steric factor. π values were derived from molecular orbital calculations[7] and steric values (st.) were arbitrarily set at zero for free rotating molecules and 1.00 for rigid, planar aromatics. No mixed types such as cyclohexane or biphenyl were used in the computer model. Values for the various coefficients in the equation are listed in Table II. Seventy data sets for 16 compounds were used in the evaluation. The coefficient of determination (or r^2, the correlation coefficient squared) was 1.00, the F ratio $3.7 \cdot 10^4$, indicating a significance level $>99.9\%$ for the correlation, and the standard deviation between experimental and calculated log V_g values was $\pm$ 0.038. For substrate 2, the temperature-dependent term b (π) is of little importance, but the temperature-independent term b' (π) has a large influence on the final results. If we gather all the terms that are temperature dependent and call them "M", they should correspond closely to $-\Delta H_s/2.3R$.

TABLE II

COEFFICIENTS FOR EQN. 2 FOR HYDROCARBONS ON SUBSTRATE 2

Coefficient	*Value*	*Coefficient*	*Value*
a	15.269	*a'*	−0.02104
b	23.895	*b'*	0.25532
c	314.01	*c'*	−1.1258
d	−155.85		

TABLE III

COMPARISON OF TEMPERATURE-DEPENDENT AND -INDEPENDENT TERMS IN EQNS. 1 AND 2

Compound	$\frac{-\Delta H_s}{2.3R}$	"M"	c_1	"N"
Decane	2120	2016	3.2851	2.9940
Dodecane	2546	2444	3.8479	3.5831
Tetradecane	2845	2873	4.1211	4.1743
m-Xylene	1892	1835	2.9492	2.7730
o-Xylene	1854	1835	2.7578	2.7730
p-Xylene	1909	1835	3.0156	2.7730
Naphthalene	2204	2203	2.8332	2.8914
2-Methylnaphthalene	2455	2422	3.1950	3.1411
Fluorene	2791	2809	3.4069	3.4098
Anthracene	2906	3017	3.2971	3.5194

Similarly, the sum of the temperature-independent terms, "N", should correspond to c_1. A comparison is given in Table III.

The values approximate each other in most cases, but are not exactly equal. One can also see that the equation makes no provision for isomeric compounds, such as the three xylenes or various alkane isomers. It may be concluded that eqn. 2 contains the principle, but not all, factors which control retention volume for hydrocarbons.

To determine if this model applied equally well to other substrates, data from

TABLE IV

COEFFICIENTS FOR EQN. 2 FOR TWO SUBSTRATES

Substrate	*Coefficients*						
	a	*b*	*c*	*d*	*a'*	*b'*	*c'*
Non-polar (No. 1)	16.897	−4.894	414.49	−221.39	−0.02241	0.34564	−1.3946
Polar (No. 3)	14.114	234.95	37.818	−322.12	−0.02137	−0.05519	−0.3997
	Statistical indices						
	r^2	s	*F ratio*				
Non-polar (No. 1)	1.00	± 0.028	$6.4\cdot10^4$ (significance level > 99.9%)				
Polar (No. 3)	1.00	± 0.040	$3.3\cdot10^4$ (significance level > 99.9%)				

substrate 1 (non-polar) and substrate 3 (polar) were analyzed by the computer in the same manner. Coefficients and statistically important values are listed in Table IV.

The statistics indicate that the equation gives equally good fits for all three substrates tested, and has a high probability of applying to many other substrates. As might be expected, the *b* (π) factor is very low for the non-polar substrate 1 and very high for the polar substrate 3. Data sets for the two substrates were: substrate 1, 16 compounds, 51 data points; substrate 3, 18 compounds, 75 data points.

To illustrate more clearly the relative importance of each coefficient in the equation for the various substrates, Table V was constructed. The *a* coefficient (predominating) is always the denominator; π and steric coefficients have been grouped. a'/a shows little variation for the substrates but the b/a and b'/a ratios show definite trends from non-polar to polar substrate but, oddly, in opposite directions. A trend also occurs in the steric terms. As can be seen, the elution order for a mixture of hydrocarbons depends very much on the steric and π electron density values and coefficients. This undoubtedly deteı mines the reversals in order seen when substrates and temperatures are changed.

TABLE V

IMPORTANCE OF VARIOUS COEFFICIENTS IN EQN. 2

Substrate	*Mol. wt.*	π		*Steric*		d/a
	$(a'/a)\cdot10^3$	b/a	b'/a	c/a	c'/a	
Non-polar (no. 1)	−1.33	− 0.289	0.0204	24.53	−0.0825	−13.10
Semi-polar (no. 2)	−1.38	1.56	0.0167	20.56	−0.0737	−10.20
Polar (no. 3)	−1.51	16.64	0.0039	2.68	−0.0283	−22.82

Four hydrocarbons which are neither completely free rotating nor rigidly planar were also run on all three systems. In no case could the equation be used to predict log V_g within 95% confidence limits by using a steric value of zero or 1.00. Therefore, the equation was resolved for these compounds to determine the most probable steric value. These are listed in Table VI. These values are identical within experimental error for substrate 1 and substrate 2, but values on substrate 3 are considerably different. The latter values may be in considerable error because of the relatively small values of the coefficients c and c' on substrate 3. It is also of some interest that the non-aromatic cyclics have negative values in all cases. This implies a greater difficulty in bonding to the substrate than either of the other types of molecules.

The final test for eqn. 2 was to predict log V_g values for compounds at several temperatures, and then to determine them experimentally. This was done for substrates 1 and 2 and the results are given in Table VII.

TABLE VI

AVERAGE STERIC VALUES FOR MIXED STERIC TYPES

Compound	*Substrate*		
	Semi-polar	*Non-polar*	*Polar*
Cyclohexane	−1.0 ± 0.4	−1.7 ± 0.5	−1.5 ± 0.2
Tetralin	−0.2 ± 0.1	−0.1 ± 0.1	−1.1 ± 0.1
Diphenyl	1.5 ± 0.1	1.4 ± 0.1	0.10 ± 0.1
trans-Stilbene	1.16 ± 0.04	1.13 ± 0.03	0.59 ± 0.04

TABLE VII

RELIABILITY OF PREDICTION OF LOG V_g BY EQN. 2

Compound	*T (°K)*	*Substrate 2*		*Substrate 1*	
		Predicted	*Exptl.*	*Predicted*	*Exptl.*
Nonadecane	503	2.1915	2.2195		
	513			2.3950	2.4191
Cumene	374			2.5735	2.4385
	383	2.2330	2.1258		
Pentamethylbenzene	405	2.5888	1.9868		

Both nonadecane and cumene fall within the 95% confidence limits, but the test for pentamethylbenzene fails badly. This would indicate that the molecule is either really non-planar because of the many methyl groups[8] or the π-bonding is difficult because of these groups.

The next consideration was the use of eqn. 2 to predict log V_g for substituted hydrocarbons. When molecular weight, π and steric values for substituted hydrocarbons were inserted in the equation, prediction failed badly, log V_g values being either too high or too low. This indicated that at least one other parameter must be added to the equation if the molecule contains atom(s) other than carbon and hydrogen. In other words, the equation as it now stands is a special case for hydrocarbons, much as the Pythagorean relation is a special case of the law of cosines, or Arrhenius acid–base theory a special case of Lewis–Bronsted theory.

Mathematical approach

A molecule undoubtedly interacts with a substrate as a whole, but the equation indicates that the total interaction can be broken into component parts, one set being temperature dependent and the other temperature independent. The equation was therefore augmented to the form:

$$\log V_g = [a\,(\text{mol. wt.}) + b\,(\pi) + c\,(\text{st.}) + d]\,\frac{1}{T} + [a'\,(\text{mol. wt.}) + b'\,(\pi) + c'\,(\text{st.})] + [e\,\frac{1}{T} + e']$$

Again, this is the equation of a straight line ($y = mx + b$) where $y = \log V_g$, and m and b correspond to the temperature-dependent and temperature-independent terms, including the e and e' coefficients. Moreover, the more generalized equation is (as shown) applicable to organic compounds other than hydrocarbons because of the inclusion of the e and e' which are characteristic of specific heteroatomic groups.

Solutions were obtained by least squares using the type of data illustrated in Table VIII. The coefficients for a, b, c, etc., previously given for the semi-polar substrate (substrate 2, Part I) were used for the least-squares solution.

The resulting values for e and e' for chloro compounds were −123.5 and 0.2235. The standard deviation between calculated and experimental values for log V_g was ± 0.019 and the average error in V_g was 3.3%. The standard deviation for corresponding hydrocarbons was ± 0.038 and the average error in V_g was 7.0%.

TABLE VIII

TYPICAL DATA USED TO SOLVE MODIFIED EQUATION FOR CHLORO COMPOUNDS ON THE SEMI-POLAR SUBSTRATE

Compound	*Mol. wt.*	π	*St.*	*T (°K)*	*log V_g (exptl.)*
2$^{\text{er}}$ Amyl chloride	106.6	0	0.00	363	1.7379
				368	1.6756
				373	1.6189
Chlorobenzene	112.6	2.05	1.00	351	2.3768
				363	2.2059
				363	2.2065
				373	2.0771
				373	2.0715
				383	1.9509
1-Chloronaphthalene	162.6	3.73	1.00	443	2.4852
				463	2.2489
				483	2.0289
2-Chloronaphthalene	162.6	3.73	1.00	443	2.4735
				463	2.2296
				483	2.0125

Encouraged by these results, we applied the same mathematical method to chloro compounds on both substrates 1 and 2 (see Part I) and eventually to a variety of other groupings on all three substrates. The results are given in Table IX and include many of the types of molecules commonly encountered in GLC.

The modified equation appears statistically to be correct with the single exception of the nitriles and, in the case of substrate 3, the amines. All other log V_g ratios fall within the 2*s* values for the hydrocarbons and should therefore fall within the statistical limits of the equation at the 95% confidence limit.

Two "mixed" steric types were also run and steric values calculated in the manner previously described. Diphenylamine was run only on substrate 2 and gave a steric value of 1.9 ± 0.1. Benzophenone, run on all three substrates gave the following values: substrate 1, 1.0 ± 0.1; substrate 2, 1.4 ± 0.1; substrate 3, 1.9 ± 0.2. The values for diphenyl and *trans*-stilbene on substrate 2 were 1.5 and 1.16, respectively. A multiplicity of free rotating aromatic rings seems, in general, to increase the value of both temperature-dependent and -independent terms.

A final check on the use of the equation for prediction was made on five different compounds employing the modified equation. The results are given in Table X.

In every case, calculated and experimental results derived from the data given in Part I[1] agree within two standard deviations units.

TABLE IX

e AND *e'* VALUES FOR POLAR GROUPINGS OF ORGANIC COMPOUNDS

Substrate	*Compound*	*e*	*e'*	*No. of detns.*	*Std. dev. in log V_g*	*Avg. error in V_g (%)*
2	Chloro compounds	−123.5	0.2235	15	±0.019	± 3.3
	Alcohols	322.1	−0.3901	18	0.033	6.2
	Ketones	455.7	−0.8832	18	0.039	8.6
	Pyridine types	848.1	−1.7728	9	0.069	12.5
	1^{er} and 2^{er} amines	285.8	−0.1605	12	0.015	2.1
	Nitriles	713.4	1.2439	9	0.102	20.7
	Imides	569.6	−0.7572	6	0.027	5.0
	Phenols	65.2	−0.1421	9	0.012	2.3
	Nitro compounds	161.4	−0.1104	9	0.016	2.8
1	Chloro compounds	−221.4	0.1512	9	±0.023	3.7
	Alcohols	173.5	−0.1235	18	0.032	6.0
	Ketones	151.8	−0.2780	15	0.040	8.2
	Pyridine types	146.7	−0.1844	9	0.029	5.4
	Amines	293.5	−0.2274	13	0.013	4.0
	Nitriles	667.9	−1.4059	9	0.065	12.3
	Nitro compounds	−25.6	0.2126	6	0.020	3.7
3	Chloro compounds	231.1	−0.2803	15	±0.032	6.4
	Alcohols	1000	−0.8460	15	0.070	11.6
	Ketones	1236	−2.1550	15	0.089	15.2
	Pyridine types	481.6	−0.5487	9	0.050	8.7
	1^{er} and 2^{er} amines	800	−0.7673	9	0.128	24.3
	Nitriles	1968	−3.5418	9	0.142	26.6
	Imides	1972	−3.0865	6	0.056	10.7
	Phenols	1286	−1.2795	6	0.031	6.4
	Nitro compounds	777.3	−1.0029	6	0.037	7.1

DISCUSSION

As *e* and *e'* values are drawn from a rather limited sample, their values probably contain some error. However, it appears to be established that for these substituent

TABLE X

PREDICTION OF LOG V_g BY USE OF EXPANDED EQUATION

$\log V_g = [a\,(\text{mol. wt.}) + b\,(\pi) + c\,(\text{st.}) + d + e]\,1/T + [a'\,(\text{mol. wt.}) + b'\,(\pi) + c'\,(\text{st.}) + e']$

Compound	*T (°K)*	*Substrate 2*		*Substrate 1*	
		Calc.	*Exptl.*	*Calc.*	*Exptl.*
Chlorocyclohexane	353	2.3031	2.2200		
Dodecyl alcohol	443	2.4867	2.4921		
	478			2.1869	2.2468
Methyl *p*-tolyl ketone	413	2.5108	2.4892		
4-Methylquinoline	453			2.5842	2.5697
Hexylamine	383			2.3288	2.3350

groups at least a single-valued function is sufficient to account for the presence of a heteroatom. All compounds studied carried only one substituent group. If a molecule were multiply-substituted, we do not know if the values for the substituent groups would be additive.

The trend of increase in values as substrates become more polar seems most reasonable. Substrate 1, a non-polar liquid, shows no *e* values higher than 300, with the exception of the questionable nitriles. Substrate 2 appears to have a high affinity for pyridine types, imides and ketones. This may be due to Lewis acid–base bonding by the boron atoms in the carborane structure. The very polar polyethylene glycol (substrate 3) consistently shows high values except for the chloro group, and even here the *e* value is positive instead of negative, as it is with the other two substrates.

The poor correlation shown by amines on substrate 3 may be due to a peculiarity of this substrate, as quinoline and isoquinoline elute in reverse order from that found with substrates 1 and 2. The same is true for aniline, N-methylaniline and N,N-dimethylaniline.

Unfortunately, no easy explanation can be given for the case of the nitriles. Predictability was unreliable on all three substrates. All we can say is that some factor is operating in the nitriles for which the equation does not account.

There does not seem to be any general relation between *e* and *e'* for a substrate. For instance, on substrate 1 the ratio *e*/*e'* for chloro compounds is −552, for ketones −516, but for alcohols it is −815. On substrate 2, these ratios are −1464, −546 and −1404, respectively, and on substrate 3, −824, −573 and −1182.

If a grand standard deviation for log V_g (calc.) is made for all the runs made on each substrate, the values are: substrate 2, ± 0.040; substrate 1, ± 0.034; substrate 3, ± 0.049. If the nitrile data are eliminated (as we consider they should be), the deviations are the same as or less than those for hydrocarbon alone. It therefore appears that the expanded equation is statistically justified and should be of use in many cases. We would like to emphasize again, however, that some unknown factors must exist (*e.g.*, to explain isomers) and the nitrile data emphasize the existence of an unknown and unexpected factor. Although the equation cannot be used to identify an unknown compound, it should be noted that the multiple regression analysis approach appears to be of considerable use in GLC and that with a greater understanding of molecular structure and bonding forces, we believe it should be possible to predict elution times accurately.

ACKNOWLEDGEMENT

The authors thank Mr. W. H. Muller for his help in collecting some of the data used in this paper.

REFERENCES

1 C. A. Streuli, W. H. Muller and M. Orloff, *J. Chromatogr.*, 101 (1974) 17.
2 S. Dal Nogare and R. S. Juvet, Jr., *Gas–Liquid Chromatography*, Interscience, New York, 1962, p.79.
3 E. B. Molnár, P. Mórity and J. Takács, *J. Chromatogr.*, 66 (1972) 205.
4 J. Takács, *J. Chromatogr. Sci.*, 11 (1973) 210.
5 G. D. Mitra and N. C. Saka, *Chromatographia*, 6 (1973) 93.
6 A. W. London and S. Sandler, *Anal. Chem.*, 45 (1973) 921.
7 A. Streitweiser, Jr., *Molecular Orbital Theory for Organic Chemists*, Wiley, New York, 1961.
8 L. M. Jackman and S. Sternhell, *Applications of NMR Spectroscopy in Organic Chemistry*, Pergamon Press, New York, 2nd ed., 1969, p. 170.