PERKIN 2

Using theoretical descriptions in structure activity relationships: retention indices of sulfur vesicants and related compounds

William H. Donovan and George R. Famini

US Army Edgewood Research, Development & Engineering Center SCBRD-RTC,
Bldg. E3160 Aberdeen Proving Ground, MD 21010-5423, USA

We have conducted a theoretical linear solvation energy relationship (TLSER) investigation of gas chromatographic (GC) retention indices for a series of 37 organosulfur compounds on three different columns, deriving regression equations based on descriptors obtained using the MNDO, AM1 and PM3 Hamiltonians. In all cases, satisfactory regressions based on two or three descriptors result, with molecular volume being the most important descriptor. Our results are qualitatively similar to those of Woloszyn and Jurs who, considering the same sulfur vesicant GC retention data set, applied an objective feature selection procedure to winnow descriptors from an initial set of 100, but were not able to treat any compounds containing sulfur–sulfur bonds. Our approach was free of this restriction, allowing for consideration of all 37 compounds. Only relatively small differences were obtained in the statistical quality of the regressions derived from the three Hamiltonians considered, and among the three GC columns.

Introduction

Operation Desert Storm in 1991 highlighted the need for adequate detection and identification techniques for the chemical warfare agents. Certain organosulfur compounds in particular still pose a significant threat as physical contact results in blistering of any unprotected skin or membranes. The persistency of these materials and lack of a true antidote against them are largely responsible for the continuing concern surrounding them.[1] The most widely recognized organosulfur vesicant is bis(2-chloroethyl) sulfide or mustard, and this species has been the subject of many recent experimental[1–8] and theoretical[9–14] investigations.

A number of analytical methodologies have been used to identify mustard and related derivatives. These include electron impact and chemical ionization mass spectroscopy, ion mobility spectroscopy, gas chromatography, thin-layer chromatography and high performance liquid chromatography. Of these, capillary gas chromatography provides perhaps the most efficient separations, even when the compounds are in a complex environmental or biological sample.

D'Agostino and Provost[2] determined the gas chromatography retention indices for 37 sulfur-containing compounds using three different polysiloxane-based column packings, DB-1 (100% dimethylpolysiloxane), DB-5 (95% methyl–5% diphenyl-polysiloxane) and DB-1701 (86%dimethyl–14%cyanopropyl-phenyl-polysiloxane). They calculated their retention data employing Van den Dool's equation (1),[15] where $\mathrm{RI_X}$ is the

$$\mathrm{RI_X} = 100n\,\frac{t_{\mathrm{R}(x)} - t_{\mathrm{R}(z)}}{t_{\mathrm{R}(z+n)} - t_{\mathrm{R}(z)}} + 100z \qquad (1)$$

retention index for compound X, n is the difference in carbon number between the two n-alkanes on either side of compound X, t_R is the retention time and z is the carbon number of the n-alkane immediately prior to compound X.

Woloszyn and Jurs[16] examined the use of computationally derived descriptors to correlate $\mathrm{RI_X}$ for the set of mustard derivatives considered here. They used the objective feature selection procedure to select a set of nine theoretical descriptors for use in their correlations. Their best regressions utilized four descriptors from the set of nine. Omitting all compounds in the data set containing sulfur–sulfur bonds, they were able to obtain impressive correlations with R values of 0.998. While this approach was certainly successful for treating the present sulfur mustard data set, it suffers from a lack of generality in treating other systems. Thus the descriptor selection process must be repeated for each new data set considered. Additionally, the physical meaning of the derived descriptors is not always apparent.

Our objective in the present paper is to examine the ability of the theoretical linear solvation energy relationship (TLSER) method to correlate the standard set of descriptors with the GC retention data of D'Agostino and Provost.[2] Several comprehensive descriptions of the TLSER method[17–25] are available and they allow us to be brief. This method is based on the widely successful linear solvation energy relationships (LSER) approach pioneered by Kamlet, Taft, Abraham and co-workers (KTA),[26–31] where the general model can be expressed by eqn. (2). In this formulation, the property is typically the logarithm

$$\text{property} = \text{bulk/cavity term(s)} + \text{dipolarity/polarization terms(s)} + \text{H bonding term(s)} + \text{constant} \qquad (2)$$

of a parameter that is related to a free energy. The bulk and cavity terms refer to the energy required to form a solute cavity in the solvent and to separate the solvent molecules. The dipolarity/polarization terms involve the various dipole and induced dipole interactions. Finally, the H bonding terms include acceptor hydrogen bond basicity (HBB) and donor hydrogen bond acidity (HBA) and can be considered strong dipole–dipole interactions.

KTA used an empirically based solvatochromic descriptor set to represent the terms in eqn. (2). This approach has been successfully applied to correlate more than 250 properties where solute–solvent interactions are important, and has delivered valuable insights into the nature of solute–solvent interactions.[27] Solvatochromic descriptor determination is an experimental step, although LSER parameter estimation techniques have been proposed.[32] The TLSER method replaces the solvatochromic parameters with a set of theoretical descriptors extracted from molecular orbital calculations. Thus the TLSER equation for the present study takes the form of eqn. (3). Table 1 summarizes the six TLSER descriptors used in

$$\mathrm{RI} = aV_\mathrm{mc} + b\pi_\mathrm{i} + c\varepsilon_\mathrm{B} + dq^- + e\varepsilon_\mathrm{A} + fq^+ + g \qquad (3)$$

Table 1 TLSER descriptors[a]

Symbol	Name	Definition	Units[b]	Meaning
V_{mc}	Molecular volume	Molecular volume	Å^3	Cavity/steric
π_i	Polarizability index	Polarizability/V_{mc}	None	Polarizability
ε_B	Covalent HBB	0.30 − |ΔE(h,lw)|/100	heV	Acceptor HBB
q^-	Electrostatic HBB	Maximum |(−) charge| on an atom	acu	Acceptor HBB
ε_A	Covalent HBA	0.30 − |ΔE(l,hw)|/100	heV	Donor HBA
q^+	Electrostatic HBA	Maximum (+) charge on an H atom	acu	Donor HBA

[a] HBB = hydrogen bond basicity, HBA = hydrogen bond acidity; ΔE(h,lw) = E(h) − E(lw); ΔE(l,hw) = E(l) − E(hw); E(h) = HOMO energy; E(l) = LUMO energy; E(lw) and E(hw) refer to the E(LUMO) and E(HOMO) for water, respectively; || indicate absolute values. [b] Cubic Ångströms (Å^3), hectoelectron volts (heV) and atomic charge units (acu).

this equation. These descriptors and units are the same as before,[19,20] except that the molecular van der Waals volume, V_{mc}, is in units of cubic Ångströms. We retain the use of a transformation to derive ε_A and ε_B so that these quantities are in hectoelectron volts (heV) and increase with increasing acidity and basicity, respectively. Also, π_i is the unitless polarizability index calculated by dividing the polarization volume by the molecular volume, indicating the ease with which the electron cloud may be moved or polarized. Moreover, the electrostatic basicity q^- is simply the magnitude of the most negative partial charge on an atom in the molecule, while the electrostatic acidity q^+ is the magnitude of the most positive hydrogen, as calculated according to Mulliken population analysis, in atomic charge units. Finally, g is the intercept. The coefficients a–g are determined using multilinear regression analysis to fit the data.

The four general descriptors of LSER translate into six TLSER descriptors because the hydrogen bond acidity and basicity terms of LSER are conveniently broken into covalent and electrostatic components by the molecular orbital calculations. Although it is not necessary to retain these specific descriptors should better ones be identified, for the present they constitute a convenient starting point. Indeed, an impressive variety of physicochemical and biological properties have already been successfully correlated with these six descriptors.[18–23] Typically, only two to four of the six descriptors end up being needed to correlate any given property. We emphasize that these descriptors all have a clear physical meaning that allows us to gain insights into the important chemical aspects impacting the observed property being correlated. The advantage of such an approach is that comparisons of different data sets can be conveniently made on the basis of the same set of descriptors.

Also of interest is to examine the performance of AM1 and PM3 *vis-à-vis* the MNDO Hamiltonian for this data set. The MNDO Hamiltonian has been employed almost exclusively in past TLSER reports from our laboratory,[18–24] despite the general agreement that AM1 and PM3 provide better geometries and heats of formation for most molecules.[33] Should this improved performance carry over to deliver better regressions, future TLSER studies may consider a departure from the MNDO Hamiltonian. On the other hand, should the AM1 and PM3 calculations deliver similar or worse regressions, then for the sake of consistency with past efforts, retention of the MNDO Hamiltonian is justified.

Procedure

GC retention data for the 37 sulfur mustard vesicants was taken from the work of D'Agostino and Provost.[2] Structure entry and visualization was done with PCMODEL.[34] MOPAC 6.0[33,35] was used to optimize all geometries and to obtain the TLSER descriptors employing the MNDO,[36] AM1[37] and PM3[33,38] Hamiltonians. Automated descriptor extraction from the MOPAC output files was accomplished using the in-house developed program MADCAP.[39] Not only does this procedure speed the process, it also reduces the chances for human error in the descriptor collection task. The molecular volume for the optimized geometry was determined using the algorithm of Hopfinger.[40] Multilinear regression analysis using MINITAB[41] was used to obtain the correlation equations.

The criteria used for derivation of the correlation equations were the same as before.[19–24] Namely, we attempted to minimize the number of descriptors necessary, while maximizing the correlation coefficient R and t-statistic (t-stat), and minimizing the variance inflation factor (VIF) and outliers. Only those descriptors found significant at the 95% confidence level were included in the reported regression equations. Previous TLSER studies[19–24,33] have taken outliers to be those compounds that deviate from the mean by three or more standard deviations. Adhering to this practice for the present study results in there being no outliers. However, if we change the criteria to two standard deviations, then up to three outliers are obtained. For comparison purposes, regressions are computed employing both methods.

Results and discussion

An examination of the derived descriptors and their coefficients allows for a rationalization of the mechanism involved in the GC retention, and a comparison of the effects of the different column packings. The importance of the TLSER descriptors in each correlation is indicated by the sign and magnitude of the coefficient and from the value of the t-statistic for each descriptor. Hence, by examining each equation separately, it is possible to gauge the relative importance of each descriptor in the gas/solid phase partitioning. This approach also makes it possible to identify the characteristics of the columns affecting the partitioning process. For those descriptors of low relative importance, but of sufficient significance to be included in the regression, the interpretation of the chemical meaning of such descriptors may not agree with expectation based on chemical intuition. Nevertheless, the most statistically significant descriptors as indicated by t-stat values should deliver predictions in line with chemical intuition. If this is not the case, then the present methodology has few redeeming qualities to justify its use over the classical principal component analysis methods.

Fig. 1 presents the structures of the 37 compounds considered. The relatively wide range of structures is important for ensuring a meaningful variation in the RI_X values to make a statistically significant correlation possible. Hence the set includes compounds ranging from as few as six heavy atoms (non-hydrogen) to as many as nineteen. It also includes cyclic and linear systems and compounds containing several sulfur–sulfur bonds. Fig. 2 illustrates the range of the retention indices by showing a plot of the predicted *vs.* experimental retention indices of the sulfur vesicants on the DB-1 column. Table 2 lists the values of the six MNDO-derived descriptors for each compound, together with the calculated and observed retention data for the DB-1 column, and the residuals. The regressions obtained using the MNDO, AM1 and PM3 calculated descriptors are summarized in Tables 3–5. These three tables include the values of the various coefficients given in eqn. (3) as

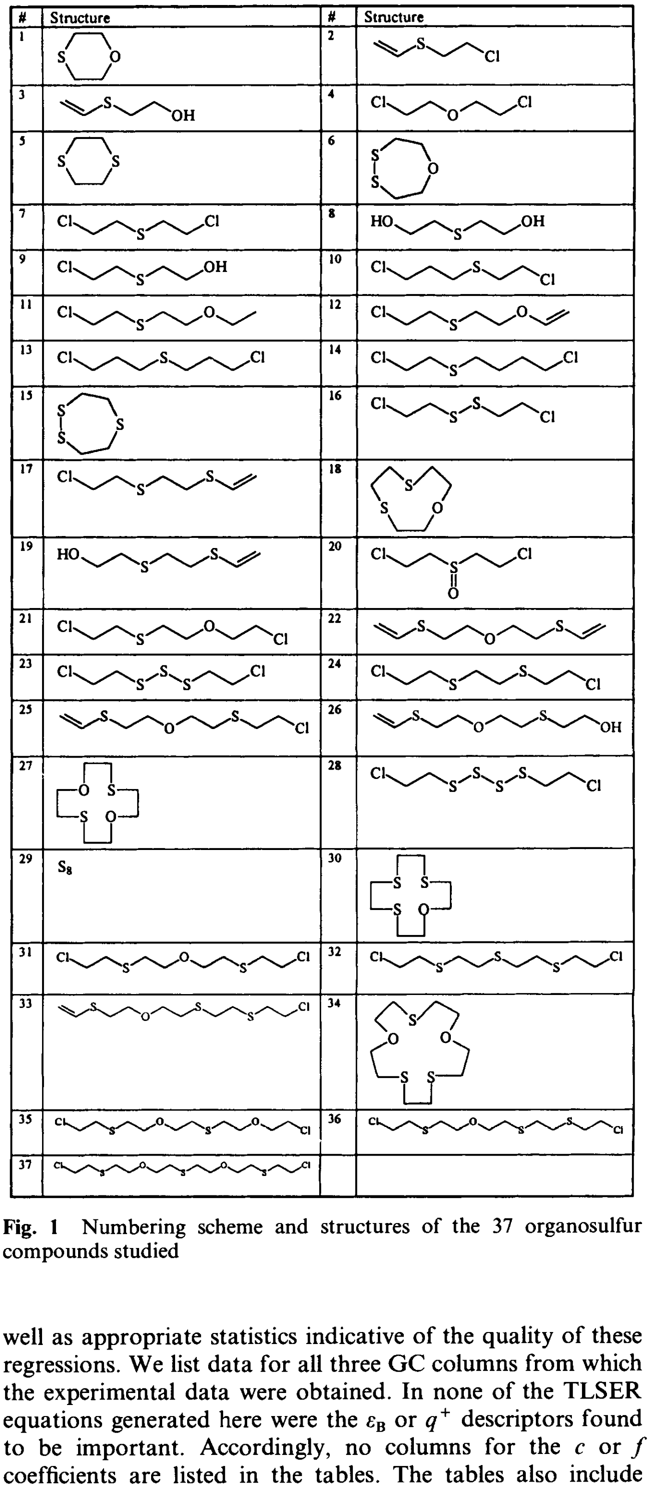

Fig. 1 Numbering scheme and structures of the 37 organosulfur compounds studied

well as appropriate statistics indicative of the quality of these regressions. We list data for all three GC columns from which the experimental data were obtained. In none of the TLSER equations generated here were the ε_B or q^+ descriptors found to be important. Accordingly, no columns for the c or f coefficients are listed in the tables. The tables also include regressions derived from data sets including all compounds having data and omitting those data points more than two standard deviations from the mean. Thus a total of eighteen regression equations were derived. No experimental data were available for compound **11** on the DB-1701 column.[2]

These tables show that only two or three of the six descriptors considered are needed to obtain good regressions with $0.966 \leqslant R \leqslant 0.982$ and $0.980 \leqslant R \leqslant 0.991$ when all compounds are included, and outliers are omitted, respectively. Descriptors that were not significant at the 95% confidence level are indicated by 'n/s' entries in the appropriate column. The t-statistics provide a relative measure of the importance of each term in the regression. The variance inflation factor (VIF) is defined as $1/(1 - R^2)$, where R is the correlation coefficient of one independent variable against the others; values close to one imply small cross correlation.[42] The standard deviations are listed in the SD row, the Fisher index is given in the F row and

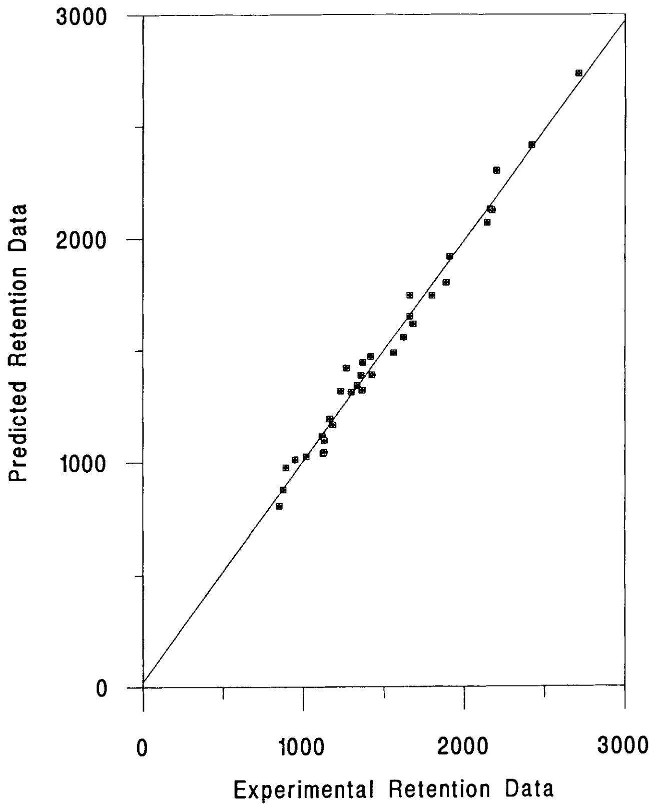

Fig. 2 Predicted *vs.* experimental retention data for sulfur vesicants and related compounds based on MNDO descriptors and DB-1 column

the number of compounds in a given regression is included in the N row. Large values of the Fisher index are an indication of more reliable correlation equations.

(a) MNDO-generated descriptors

Table 3(*a*) lists the regressions resulting from the use of MNDO-generated descriptors for each of the three columns. Table 3(*b*) gives the same regressions with outliers (>2 SD) **11**, **20** and **29** removed. Compounds **20** and **29** have unique structural features among the 37 compounds studied (sulfoxide and S_8, respectively) that may be partially responsible for the discrepancies obtained. While **11** cannot be explained by this reasoning, we note that the deviation for this compound is 2.0 SD from the mean. It may also cause some experimental difficulty as no retention data for this material was reported for the DB-1701 column.[2] In all of these cases, the TLSER equation contains three parameters: molecular volume, molecular orbital acidity and polarizability index for Table 3(*a*); molecular volume, polarizability index and molecular orbital acidity for Table 3(*b*), in that order of significance, respectively. Hence the impact of removing outliers is not trivial: there is a reversal in the order of significance of the 2nd and 3rd descriptors, the statistical quality of the regressions improve, and the number of compounds included in the analysis drops from 37 to 34 for the DB-1 and DB-5 columns, and from 36 to 34 for the DB-1701 column.

The positive sign of the volume term coefficient indicates larger molecules prefer the liquid over the gas phase. This makes physical sense as larger molecules usually have higher boiling points. The positive sign of the molecular orbital acidity coefficient indicates that better acids partition preferentially in the liquid phase. All of the column packings contain basic groups, so acids would be expected to interact with the column. Unlike the V_{mc} and ε_A coefficients, the π_i coefficient is negative, predicting that highly polarizable sulfur mustard derivatives display some preference for the gas over the liquid phase. This

Table 2 MNDO-generated TLSER descriptors, observed[a] and predicted GC retention data for 37 sulfur vesicants on a DB-1 column

Entry[b]	V_{mc}	π_i	ε_B	q^-	ε_A	q^+	Observed	Predicted	Residual
1	91.49	0.106 2	0.149 108	0.333 9	0.161 931	0.043 9	850.7	771.6	79.1
2	104.30	0.112 4	0.152 820	0.225 5	0.173 135	0.064 7	873.7	901.4	−27.7
3	99.28	0.104 2	0.155 194	0.325 9	0.169 214	0.184 8	889.6	974.5	−84.9
4	115.88	0.110 9	0.129 268	0.353 0	0.173 119	0.043 2	948.3	1032.4	−84.1
5	102.95	0.101 7	0.149 240	0.100 2	0.164 626	0.047 2	1018.7	998.0	20.7
6	110.19	0.105 5	0.148 603	0.342 5	0.177 826	0.052 4	1117.8	1154.5	−36.7
7	121.90	0.113 4	0.143 392	0.204 8	0.176 585	0.043 1	1123.8	1078.5	45.3
8	108.19	0.102 4	0.148 811	0.326 1	0.163 062	0.184 6	1130.9	1008.8	122.1
9	116.21	0.107 1	0.146 069	0.325 4	0.173 232	0.186 1	1132.3	1114.3	18.0
10	139.61	0.112 7	0.144 678	0.208 1	0.174 472	0.043 1	1168.8	1219.0	−50.2
11	150.52	0.109 8	0.146 200	0.345 7	0.173 179	0.042 1	1169.2	1356.2	−187.0
12	143.55	0.115 3	0.151 515	0.287 9	0.173 649	0.060 0	1185.3	1189.4	−4.1
13	156.29	0.112 8	0.145 993	0.206 7	0.171 778	0.037 1	1235.8	1326.8	−91.0
14	159.93	0.110 3	0.145 264	0.212 7	0.173 944	0.042 4	1269.3	1437.5	−168.2
15	121.23	0.101 4	0.146 568	0.113 9	0.179 100	0.052 5	1300.0	1350.6	−50.6
16	143.56	0.111 3	0.142 307	0.200 2	0.184 665	0.046 3	1336.3	1414.1	−77.8
17	154.65	0.112 7	0.153 923	0.224 7	0.173 816	0.063 2	1366.1	1341.1	25.0
18	146.35	0.102 8	0.151 218	0.331 7	0.164 675	0.049 3	1358.0	1352.7	5.3
19	153.13	0.104 7	0.154 863	0.325 9	0.169 496	0.184 9	1370.9	1435.3	−64.4
20	132.62	0.119 9	0.141 532	0.706 6	0.184 594	0.064 2	1367.0	1142.5	224.5
21	166.78	0.110 9	0.144 900	0.351 1	0.174 180	0.042 8	1418.3	1487.8	−69.5
22	173.49	0.115 6	0.154 894	0.349 3	0.169 475	0.062 7	1425.6	1389.0	36.6
23	161.06	0.113 5	0.141 334	0.198 5	0.189 386	0.045 7	1561.7	1582.4	−20.7
24	175.58	0.111 0	0.145 039	0.207 4	0.175 426	0.042 6	1622.7	1578.3	44.4
25	198.83	0.111 1	0.154 460	0.349 2	0.173 498	0.062 9	1660.6	1753.0	−92.4
26	188.23	0.109 7	0.154 622	0.347 7	0.169 769	0.184 9	1660.5	1641.2	19.3
27	184.82	0.106 8	0.151 422	0.338 6	0.161 612	0.035 8	1681.0	1565.0	116.0
28	184.55	0.112 2	0.142 944	0.197 8	0.191 725	0.048 0	1800.0	1843.1	−43.1
29	145.28	0.110 4	0.141 492	0.023 8	0.199 340	0.000 0	1845.0	1638.0	207.0
30	196.21	0.103 7	0.151 155	0.339 3	0.163 799	0.049 7	1884.9	1755.6	129.3
31	217.64	0.111 0	0.145 298	0.349 2	0.174 056	0.042 4	1909.9	1925.5	−15.6
32	230.36	0.109 4	0.146 181	0.208 1	0.174 806	0.042 5	2140.7	2078.3	62.4
33	246.58	0.112 8	0.154 613	0.349 4	0.173 912	0.062 8	2160.6	2137.9	22.7
34	238.78	0.105 5	0.151 952	0.340 8	0.162 348	0.041 1	2171.7	2069.4	102.3
35	260.77	0.110 6	0.146 805	0.351 5	0.173 711	0.042 6	2200.0	2303.4	−103.4
36	271.35	0.110 0	0.146 661	0.349 4	0.174 150	0.042 4	2418.8	2413.2	5.6
37	310.73	0.111 1	0.147 193	0.349 5	0.173 670	0.042 4	2711.8	2726.2	−14.4

[a] See ref. 2. [b] See Fig. 1 for structures of each compound listed.

Table 3 Regression values obtained for eqn. (3) when using MNDO descriptors and omitting outliers defined as (*a*) three or more standard deviations from the mean and (*b*) two or more standard deviations from the mean

GC column	*a*: coeff *t*-stat VIF	*b*	*d*	*e*	*g*	N	R	SD	F
(*a*)									
DB-1	8.677	−20 438	n/s	12 977	46.7	37	0.980	94.81	271
	28.18	4.64		5.77	0.11				
	1.1	1.4		1.3					
DB-5	8.992	−21 747	n/s	13 954	25.4	37	0.979	102.3	250
	27.07	4.58		5.75	0.05				
	1.1	1.4		1.3					
DB-1701	10.073	−24 085	n/s	15 891	−25.1	36	0.973	132.5	187
	23.37	3.91		5.06	0.04				
	1.1	1.4		1.3					
(*b*)									
DB-1	8.834	−21 488	n/s	8 111	968.9	34	0.991	64.10	587
	41.40	6.42		4.32	2.87				
	1.1	1.5		1.4					
DB-5	9.170	−23 435	n/s	9 004	1 027.4	34	0.991	68.83	548
	40.02	6.52		4.46	2.84				
	1.1	1.5		1.4					
DB-1701	10.423	−29 215	n/s	10 309	1 422.3	34	0.988	88.69	424
	35.30	6.31		3.97	3.05				
	1.1	1.5		1.4					

prediction conflicts with chemical intuition, however the small *t*-stat value of this descriptor indicates that this shortcoming is not catastrophic.

(*b*) AM1-generated descriptors

Tables 4(*a*) and 4(*b*) list the regressions resulting from the use of AM1-generated descriptors for each of the three columns, including all compounds and with outliers removed, respectively. Instead of obtaining a three-parameter equation, here just two parameters are needed: molecular volume and the polarizability index. Not surprisingly, the use of just two descriptors in these correlation equations results in lower quality statistics than those obtained from the corresponding

Table 4 Regression values obtained for eqn. (3) when using AM1 descriptors and omitting outliers defined as (*a*) three or more standard deviations from the mean and (*b*) two or more standard deviations from the mean

GC column	*a*: coeff *t*-stat VIF	*b*	*d*	*e*	*g*	N	R	SD	F
(*a*)									
DB-1	8.235	5267	n/s	n/s	−303	37	0.979	95.71	398
	27.57	6.31			3.48				
	1.0	1.0							
DB-5	8.523	5543	n/s	n/s	−317	37	0.977	105.1	354
	25.98	6.04			3.31				
	1.0	1.0							
DB-1701	9.547	5813	n/s	n/s	−301	36	0.966	144.8	232
	21.11	4.59			2.27				
	1.0	1.0							
(*b*)									
DB-1	8.323	5519	n/s	n/s	−339.1	35	0.984	85.3	505
	31.06	7.37			4.31				
	1.0	1.0							
DB-5	8.625	5830	n/s	n/s	−358.2	35	0.983	92.6	461
	29.65	7.17			4.19				
	1.0	1.0							
DB-1701	9.944	5446	n/s	n/s	−353.9	34	0.984	106.0	435
	29.16	3.38			2.50				
	1.0	1.0							

Table 5 Regression values obtained for eqn. (3) when using PM3 descriptors and omitting outliers defined as (*a*) three or more standard deviations from the mean and (*b*) two or more standard deviations from the mean

GC column	*a*: coeff *t*-stat VIF	*b*	*d*	*e*	*g*	N	R	SD	F
(*a*)									
DB-1	8.209	5763	383.9	n/s	−437.0	37	0.982	90.7	297
	28.90	7.82	2.12		4.38				
	1.0	1.1	1.1						
DB-5	8.494	6130	436.4	n/s	−470.5	37	0.980	98.7	270
	27.49	7.64	2.21		4.34				
	1.0	1.1	1.1						
DB-1701	9.518	7067	808.7	n/s	−604.5	36	0.976	125.0	211
	24.28	6.93	3.24		4.38				
	1.0	1.1	1.1						
(*b*)									
DB-1	8.209	5763	383.9	n/s	−437.0	37	0.982	90.7	297
	28.90	7.82	2.12		4.38				
	1.0	1.1	1.1						
DB-5	8.494	6130	436.4	n/s	−470.5	37	0.980	98.7	270
	27.49	7.64	2.21		4.34				
	1.0	1.1	1.1						
DB-1701	9.723	7270	781.0	n/s	−648.9	34	0.984	105.4	298
	28.88	8.43	3.69		5.53				
	1.0	1.1	1.1						

three descriptor MNDO-derived equations. Thus the correlation coefficients are lower and the standard deviations higher with this model than obtained with MNDO. Perhaps less expected is the observation that the π_i coefficient is found to be positive instead of negative as in the case of MNDO. Removing outliers reduces the number of compounds considered from 37 to 34 for the DB-1 and DB-5 columns, and from 36 to 34 for the DB-1701 column. In this case, compounds **20** and **22** are found to be outliers. It is not apparent why **22** is an outlier with the AM1-calculated descriptors, but we note that the deviation obtained for this compound is 2.1 SD. As found with MNDO, improvement in the statistical quality of the regressions results from removing outliers.

(*c*) PM3-generated descriptors

Tables 5(*a*) and 5(*b*) list the regressions resulting from the use of PM3-generated descriptors for each of the three columns, including all compounds and with outliers removed, respectively. In this case we obtain a three-parameter regression: molecular volume, polarizability index and electrostatic basicity are found to be important in that order. Examination of the *R* and SD values in Table 5(*a*) shows that this method delivers similar quality regressions as obtained with MNDO. However, fewer outliers are found with PM3 so that less statistical improvement is found upon removal of outliers. In fact, this model gives no outliers for the DB-1 and DB-5 columns, and just two for the DB-1701 column: **8** and **22**. In agreement with AM1, the π_i coefficient is positive. Curiously, the coefficient for q^- is also positive, suggesting that more basic HD derivatives interact more strongly with the column packings. This runs counter to expectation, but again the small effect of this parameter is noteworthy.

Implications

A comparison of the results between Tables 3–5 as computed from the MNDO, AM1 and PM3 Hamiltonians brings out several interesting points. All methods handle the DB-1 and DB-5 column retention data better than that from DB-1701. This is reflected by the lower R values and larger standard deviations obtained for the regressions involving DB-1701 data. This column packing has the most sterically hindered side chains and is the most polarizable. Thus the vesicants are retained longer by DB-1701 than the other two columns. It is possible that the longer retention on this column allows factors not fully accounted for by the present TLSER method to play a larger role in influencing retention times than in the case of DB-1 and DB-5. Also, in all regression equations, molecular volume is clearly the most important term. The next most important term according to AM1 and PM3 is the polarizability, π_i. This term has a negative coefficient with MNDO but positive coefficients with AM1 and PM3. Hence these methods predict qualitatively different effects of polarizability on the retention times. The third (and last) significant term also varies, from electrostatic basicity q^- with PM3 to none with AM1 to covalent acidity ε_A or polarizability π_i with MNDO, with and without removal of outliers, respectively. The contributions of the 2nd and 3rd terms are small and similar, making possible the reversal observed in the relative importance of these terms in the MNDO case [see Tables 3(*a*) and 3(*b*)].

It is also instructive to compare the present TLSER results with those obtained by Woloszyn and Jurs[16] (WJ). Employing a four-descriptor equation, they were able to correlate 31 out of 37 compounds with a correlation coefficient R of 0.998 for the DB-1 and DB-5 columns. With the TLSER method using the MNDO and PM3 Hamiltonians, we are able to correlate all 37 compounds with equations comprised of three descriptors, delivering R values within 0.02 of the WJ result for the same two columns. Removal of outliers results in correlating 34 of 37 compounds having R values within 0.01 of the WJ result. Similar results were obtained for the DB-1701 column although there was some minor deterioration here. The AM1-derived regressions simplify to only two descriptors. While the statistical quality of these regressions were slightly worse than those from MNDO and PM3, the use of just two descriptors can be considered a mitigating factor that partially offsets the less impressive statistics. The real advantage of the present TLSER method is that it allows for a chemically meaningful interpretation to be made from a consistent set of descriptors, that may be applied to a wide range of compounds and a vast array of properties. Although the chemical predictions obtained from the TLSER method are not always perfect, they clearly can provide useful insights about the chemistry involved in systems such as that considered in the present work.

Conclusions

This study has demonstrated that the TLSER method is able successfully to correlate the GC retention data for a wide range of sulfur vesicants on three columns. The regressions indicate that the molecular volume is the most important factor influencing the time of retention, with small volumes leading to short retention times. Similar quality regressions are obtained from all three semi-empirical Hamiltonians employed: MNDO, AM1 and PM3. This suggests that the approach of previous TLSER studies in using the MNDO Hamiltonian is reasonable. Nevertheless, similar studies employing the AM1 and PM3 Hamiltonians are clearly viable alternatives, particularly for data sets involving compounds known to be poorly described by the MNDO model. Finally, we find that the TLSER methodology delivers similar quality statistical results compared to more empirical approaches, but offers the advantage provided by a more consistent theoretical framework. The success of the TLSER method to date makes us sanguine about the prospects of it being successfully applied to more diverse systems in the future.

Acknowledgements

This work was performed while W. H. D. held a National Research Council–US Army ERDEC Research Associateship.

References

1 B. Papirmeister, A. J. Feister, S. I. Robinson and R. D. Ford, *Medical Defense Against Mustard Gas, Toxic Mechanisms and Pharmacological Implications*, CRC Press, Boca Raton, 1991.

2 P. A. D'Agostino and L. R. Provost, *J. Chromatogr.*, 1988, **436**, 399.

3 S. P. McManus, R. M. Karaman, R. Sedaghat-Harati, B. A. Hovanes, X.-T. Ding and J. M. Harris, *J. Org. Chem.*, 1993, **58**, 6466.

4 S. P. McManus, N. Neamati-Mazraeh, B. A. Hovanes, M. S. Paley and J. M. Harris, *J. Am. Chem. Soc.*, 1985, **107**, 3393.

5 Y.-C. Yang, J. R. Ward and T. Luteran, *J. Org. Chem.*, 1986, **51**, 2756.

6 Y.-C. Yang, J. R. Ward, R. B. Wilson, W. Burrows and J. S. Winterle, *Thermochim. Acta*, 1987, **114**, 313.

7 Y.-C. Yang, L. L. Szafraniec, W. T. Beaudry and J. R. Ward, *J. Org. Chem.*, 1988, **53**, 3293.

8 Y.-C. Yang, J. A. Baker and J. R. Ward, *Chem. Rev.*, 1992, **92**, 1729.

9 W. H. Donovan and G. R. Famini, *J. Phys. Chem.*, 1994, **98**, 3669.

10 P. Politzer and D. Habibollahzadeh, *J. Phys. Chem.*, 1994, **98**, 1576.

11 W. H. Donovan, G. R. Famini, J. O. Jensen and H. F. Hameka, *Phosphorus, Sulfur and Silicon*, 1993, **80**, 47.

12 H. Broch, R. Viani and D. Vasilescu, *Int. J. Quantum Chem.*, 1992, **43**, 511.

13 H. F. Hameka, S. L. Emery, G. R. Famini, J. M. Leonard and D. J. Reutter, *Phosphorus, Sulfur and Silicon*, 1990, **53**, 373.

14 C. Sosa, R. J. Bartlett, K. KuBulat and W. B. Person, *J. Phys. Chem.*, 1989, **93**, 577.

15 H. Van den Dool and P. D. Kratz, *J. Chromatogr.*, 1963, **11**, 463.

16 T. F. Woloszyn and P. C. Jurs, *Anal. Chem.*, 1992, **64**, 3059.

17 G. R. Famini and L. Y. Wilson, in *Quantitative Treatments of Solute/Solvent Interactions*, eds. P. Politzer and J. S. Murray, Elsevier, New York, 1994, pp. 213–241.

18 C. J. Cramer, G. R. Famini and A. H. Lowrey, *Acc. Chem. Res.*, 1993, **26**, 599.

19 G. R. Famini and L. Y. Wilson, *J. Chem. Soc., Perkin Trans. 2*, 1994, 1641.

20 G. R. Famini, B. C. Marquez and L. Y. Wilson, *J. Chem. Soc., Perkin Trans. 2*, 1993, 773.

21 G. R. Famini, C. A. Penski and L. Y. Wilson, *J. Phys. Org. Chem.*, 1992, **5**, 395.

22 G. R. Famini, W. P. Ashman, A. P. Mickiewicz and L. Y. Wilson, *Quant. Struct.-Act. Relat.*, 1992, **11**, 162.

23 G. R. Famini, R. J. Kassel, J. W. King and L. Y. Wilson, *Quant. Struct.-Act. Relat.*, 1991, **10**, 344.

24 L. Y. Wilson and G. R. Famini, *J. Med. Chem.*, 1991, **34**, 1668.

25 G. R. Famini, *Using Theoretical Descriptors in Quantitative Structure–Activity Relationships. V. A Review of the Theoretical Parameters*, CRDEC-TS-085, US Army Chemical Research, Development and Engineering Center, Aberdeen Proving Ground, MD, 1989.

26 M. J. Kamlet, R. M. Doherty, M. H. Abraham and R. W. Taft, *Quant. Struct.-Act. Relat.*, 1988, **7**, 71.

27 M. J. Kamlet, R. W. Taft, G. R. Famini and R. M. Doherty, *Acta Chem. Scand.*, 1987, **41**, 589.

28 M. J. Kamlet, M. A. Abraham, R. M. Doherty and R. W. Taft, *Nature (London)*, 1985, **313**, 384.

29 M. J. Kamlet, M. A. Abraham, R. M. Doherty and R. W. Taft, *J. Am. Chem. Soc.*, 1984, **106**, 464.

30 M. J. Kamlet and R. W. Taft, *Prog. Org. Chem.*, 1983, **48**, 2877.

31 M. J. Kamlet, R. W. Taft and J.-L. M. Abboud, *J. Am. Chem. Soc.*, 1977, **91**, 8325.

32 J. P. Hickey and D. R. Passino-Reader, *Environ. Sci. Technol.*, 1991, **25**, 1753.

33 J. J. P. Stewart, *J. Comput.-Aided Mol. Des.*, 1990, **4**, 1.

34 PCMODEL, Serena Software, PO Box 3076, Bloomington, IN 47402.

35 J. J. P. Stewart, *QCPE Bull.*, 1990, **10**, 86.
36 M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, 1977, **99**, 4899.
37 M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902.
38 J. J. P. Stewart, *J. Comput. Chem.*, 1989, **10**, 209.
39 J. A. Fager, MADCAP1.0: MOPAC Automatic Data Collection and Assembly Program, US Army Edgewood Research, Development & Engineering Center, Aberdeen Proving Ground, MD, 1994.
40 A. Hopfinger, *J. Am. Chem. Soc.*, 1980, **102**, 7126.
41 MINITAB, Minitab, Inc., 3081 Enterprise Dr., State College, PA 16801.
42 D. A. Belesley, E. Kuh and R. E. Welsh, *Regression Diagnostics*, Wiley, New York, 1980.

Paper 5/02804C
Received 1*st May* 1995
Accepted 28*th July* 1995