Use of the DIPPR Database for Development of QSPR Correlations: Surface Tension[†]

Thomas A. Knotts, W. Vincent Wilding, John L. Oscarson, and Richard L. Rowley*

Department of Chemical Engineering, 350 CB, Brigham Young University, Provo, Utah 84602

Combination of commercial QSPR (quantitative structure−property relationship) software with an evaluated database creates a powerful tool for development of thermophysical property correlations. By using data quality codes in the DIPPR relational database, a training set of property values within a desired accuracy level can be obtained for use in QSPR regression software. Moreover, additional database queries can be used to restrict the training set to specified families or functional groups and further refine the molecular descriptors that are used to correlate the property. This provides a good basis for rapid development of QSPR correlations of known uncertainty and chemical domain. This procedure is illustrated by its application to the extension of the Macleod−Sugden (*Trans. Faraday Soc.* **1923**, *19*, 38. *Chem. Soc.* **1924**, *125*, 32.) correlation for surface tension based upon the parachor. Quayle (*Chem. Rev.* **1953**, 53, 439−591.) correlated the parachor in terms of additive atomic and structural increments but used a training set limited in temperature and scope. In this work, new molecular descriptors were selected consistent with the accuracy of the training set extracted from the DIPPR database, and their additive increments to the parachor were regressed from 8697 surface tension values of uncertainty less than 5% for 649 different compounds. This produced a correlation with an average absolute deviation (AAD) of 3.2%. This can be compared with an AAD of 6.9% using the Quayle descriptors for the same set.

Introduction

The DIPPR pure-component database, containing 44 properties for over 1700 compounds, is an *evaluated* database. While not the largest database available, its focus on evaluation of the collected data can be very useful in developing property correlations. All values are evaluated for experimental accuracy, thermodynamic consistency between multiple properties where appropriate, and consistency with known relationships and trends. Twenty-six quality control checks are used to verify these internal consistencies. Additionally, comparisons of trends for properties within families are determined to ensure that a broader agreement of the properties exists throughout the database. For example, Figure 1 illustrates the consistency for the values accepted by DIPPR for the critical temperature, T_c, for the *n*-alcohol family.

This comprehensive evaluation of properties is used to assign a single *accepted* value for constant properties and a best-fit correlation for temperature-dependent properties. Assigned accuracy levels for the data are also stored in the database as an uncertainty of <0.2%, <1%, <3%, <5%, <10%, <25%, <50%, <100%, >100%, or unknown. Additionally, values are predicted for properties for which there is no experimental value in the literature, and the same quality control checks and uncertainty assignments are made for these predicted values. DIPPR has adopted standard prediction methods, classified as primary, secondary, or tertiary, for the 44 properties in the database. The adoption and use of these methods by DIPPR are based on extensive comparisons of calculated results to experimental data. The evaluation and assessment of property values which characterize the DIPPR database as evaluated also make it a valuable tool in development of estimation methods.

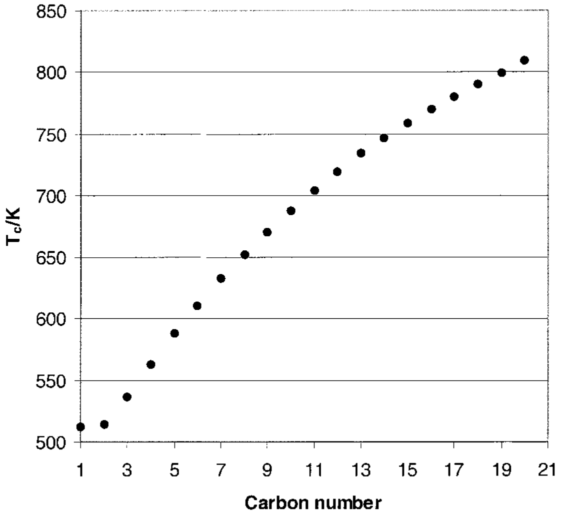

Figure 1. Internal consistency of critical temperatures for the *n*-alcohol family in the DIPPR database.

QSPR Methods

As illustrated in Figure 1, many chemical and physical properties correlate well with the molecular structure of the compound. The correlation of properties to structure has long been an aim of scientists and engineers. In recent years, efforts along this line have increased exponentially in the area of quantitative structure−property relationship (QSPR) research. In principle, the molecular structure contains all of the information which predetermines the chemical and physical properties of the compound. By this

[†] This contribution will be part of a special print edition containing papers presented at the Fourteenth Symposium on Thermophysical Properties, Boulder, CO, June 25−30, 2000.

* Corresponding author. E-mail: rowley@byu.edu. Fax: 801-378-7799.

10.1021/je000232d CCC: $20.00 © 2001 American Chemical Society
Published on Web 02/07/2001

Table 1. Main Types of Molecular Descriptors and Examples of Each[4]

type of descriptor	example
constitutional	molecular weight, number of atoms, bonds, number of rings, chemical groups
topological	Weiner index, Randic indices, Kier and Hall indices
electrostatic	partial charges, polarity indices, charged partial surface areas
geometrical	principal moments of inertia, molecular volume, solvent-accessible molecular surface
quantum-chemical	net atomic charges, dipole moment, polarization, HOMO and LUMO energies, FMO reactivity indices

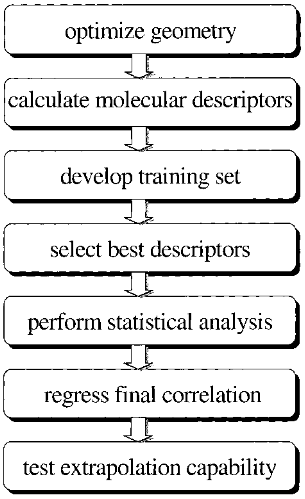
optimize geometry
calculate molecular descriptors
develop training set
select best descriptors
perform statistical analysis
regress final correlation
test extrapolation capability

Figure 2. General flowchart for development of property correlation.

statement, we mean “structure” in its fullest sense, including not only atomic arrangement and bonding but also molecular orbital and electron density information. QSPR attempts to use quantum mechanics to define the structure of the molecule, in this broadest sense, and then correlate that structure to experimental values of properties through the use of molecular descriptors. These molecular descriptors, often obtained from quantum mechanical calculations, define the overall structure at the molecular level.

Although literally hundreds of potential descriptors have been defined, Kastritzky et al.[4] list five main types of molecular descriptors and examples of each as shown in Table 1. CODESSA software, for example, contains 45 constitutional, 66 topological, 105 electronic, 8 geometric, and 76 combined descriptors.[5]

The key elements of a generalized QSPR approach for prediction of thermophysical properties are shown in Figure 2. Initially a geometry optimization is performed using appropriate energy minimization techniques in conjunction with quantum mechanics calculations. The quantum mechanics package is further used to generate the molecular descriptors from the optimized geometry and resultant wave function. One then chooses an appropriate training set (TS) of experimental data that will be used to regress coefficients for the descriptors in the correlation. Initial sensitivity analysis with commercial QSPR software can help identify those descriptors that are most significant statistically in correlating the property. This reduced set of descriptors is then used to obtain the final correlation, the linear coefficients for the descriptors being obtained from a least-squares analysis of the training set data. Some experimental data with which to later test the extrapolation capability of the new correlation should generally be withheld from the TS.

Two key aspects of this process are development of the descriptors and the appropriate TS from which to develop the correlation. The recent surge in QSPR research has enticed several companies to develop commercial QSPR software that handles the tedium of developing descriptors and doing the statistical analysis and regression. Coupled with the unique capabilities of the evaluated DIPPR database, such software can be used for rapid and accurate development of property estimation techniques. We illustrate this capability with a simple application: improvement of the Macleod−Sugden−Quayle[1−3] (MSQ) method for prediction of surface tension.

Table 2. Sugden’s Atomic and Structural Parachor Values

unit	parachor	unit	parachor	unit	parachor
C	4.8	F	25.7	three-member ring	16.7
H	17.1	Cl	54.3	four-member ring	11.6
N	12.5	Br	68.0	five-member ring	8.5
P	37.7	I	91.0	six-member ring	6.1
O	20.0	double bond	23.2	O_2 (esters)	60.0
S	48.2	triple bond	46.6		

MSQ Tension Method

A remarkably simple expression for estimation of the surface tension was proposed by Macleod.[1] Macleod expressed the surface tension, σ, as a function of the coexisting saturated liquid and vapor densities, ρ_L and ρ_V, respectively, using

$$\sigma = K(\rho_L - \rho_V)^4 \qquad (1)$$

Sugden[2] modified this expression slightly to

$$\sigma = [\mathbf{P}(\rho_L - \rho_V)]^4 \qquad (2)$$

where **P** is a temperature-independent parameter called the parachor. Sugden surveyed the then existing data for surface tensions and densities and calculated the parachors of 167 substances. Sugden assumed the parachor to be additive with respect to atomic, ring, and bond structural components. He found 2% agreement for 145 of the 167 compounds. Even though Sugden stressed atoms as the basic structural group, he recognized that oxygen atoms in esters and alcohols had to be treated differently. Sugden’s atomic and structural parachor values are given in Table 2.

Mumford and Phillips[6] detected shortcomings in Sugden’s attractively simple additivity assumption. In particular, they found considerable discrepancies as the training set was expanded to include branched chain isomers. They modified Sugden’s values based on a regression of data that included compounds with CH_x groups in structurally different environments. With the advent of better experimental techniques and higher precision surface tension measurements, Quayle[3] found that the parachor is “grossly additive, [but it] is sensitive to almost any change in structure and is particularly sensitive to any change in degree of unsaturation.” Although Quayle principally retained the atomic additivity concept, he was forced to expand his descriptors to include some structural distinction, as shown in Table 3. For example, three different values are used for H contributions depending upon the environment. Likewise, his O contributions

Table 3. Quayle's Structural Contributions to the Parachor

group	increment	group	increment
C	9.0	I	90.3
H	15.5	Se	63
H (in OH)	10.0	Si	31
H (in HN)	12.5	Al	55
CH_2 < 12 carbons	40.0	Sn	64.5
CH_2 > 12 carbons	40.3	As	54
1-methylethyl	133.3	= (terminal; aromatic)	19.1
1-methylpropyl	171.9	= (2,3 position)	17.7
2-methylpropyl	173.3	= (3,4 position)	16.3
1,1-dimethylethyl	170.4	R(C=O)R′ (R + R′ = 2)	51.3
1-methylbutyl	211.7	R(C=O)R′ (R + R′ = 3)	49.0
1-ethylpropyl	209.5	R(C=O)R′ (R + R′ = 4)	47.5
1,1-dimethylpropyl	207.5	R(C=O)R′ (R + R′ = 5)	46.3
1,2-dimethylpropyl	207.9	R(C=O)R′ (R + R′ = 6)	45.3
1,1,2-trimethylpropyl	243.5	R(C=O)R′ (R + R′ = 7)	44.1
C_6H_5	189.6	triple bond	40.6
O_2 (esters)	54.8	three-member ring	12.5
O	19.8	four-member ring	6.0
N	17.5	five-member ring	3.0
S	49.1	six-member ring	0.8
P	40.5	seven-member ring	4.0
F	26.1	sec−sec adjacency	−1.6
Cl	55.2	sec−tert adjacency	−2.0
Br	68.0	tert−tert adjacency	−4.5

depend on bond structure. Additionally, ring and bond structural descriptors are included. Though extensive data were analyzed to see the effect of environment on the parachor increment, Quayle regressed the contributions shown in Table 3 from a limited number of compounds at one or two temperatures for which the most accurate data were available.

Although other methods for estimation of σ have been developed, the good performance and simplicity of eq 2 have made it a very popular estimation technique when combined with Quayle's group contributions.[7−12] The DIPPR database also lists it as the primary estimation method.[13] Moreover, the functional form of eq 2 is consistent with an equation derivable from statistical mechanics,[14]

$$\sigma = \frac{kT}{4}\tau^{4-2g}\frac{z\zeta}{z_c}(\rho_L - \rho_V)^4 \qquad (3)$$

where k is the Boltzmann constant, T is temperature, $\tau = (1 - T_r)$, where T_r is reduced temperature (T/T_c), g is an exponent, z is the activity (see ref 12 for its definition), and ζ is a function involving the direct correlation function that is a very weak function of density. This form led Escobedo and Mansoori[12] to identify the parachor as

$$\mathbf{P} = \mathbf{P}_0(1 - T_r)^{1-g/2}T_r \exp\left(\frac{\kappa\mu_r}{T_r}\right)\zeta^{1/4} \qquad (4)$$

where $\kappa = 2\mu_c/(4kT_c)$ and μ is chemical potential, and to correlate it ultimately as a weak function of reduced temperature,

$$\mathbf{P} = \mathbf{P}_0(1 - T_r)^{0.37}T_r \exp\left(\frac{0.30066}{T_r} + 0.86442T_r^9\right) \qquad (5)$$

They used Quayle's groups to evaluate $\mathbf{P}_0$. Unfortunately, it is not clear at what temperature $\mathbf{P}_0$ should be evaluated, and in fact the Quayle groups used had been evaluated over a range of temperatures. Nevertheless, good results were reported using this correlation for 96 different compounds.

New MSQ Groups

Quayle found, when using a modest training set of the most accurate data available at the time, that the MSQ groups were dependent upon the chemical environment of the structure. That is, values for the simple atomic descriptors were affected by groups attached to them. In fact, he found that any difference in structure that tended to change the structural volume would have an impact upon the value of the parachor. The accuracy of the method can therefore be improved by defining better independent descriptors. To this end, we have defined descriptors as the smallest chemically unique group of atoms, consistent for example with the descriptors used for group contribution methods such as those by Lydersen[15] and Joback.[16] One can also make improvements in the correlation by using a broader training set so as to obtain better extrapolation of the MSQ method to new molecules at the expense of very accurate calculations for a small set of compounds. We have used both of these techniques in this work to improve the MSQ method and illustrate the capabilities of coupling an evaluated database with QSPR software.

In developing new group increments for the MSQ equation, we have been consistent with the original definition of the parachor as a temperature-independent property. Though σ values at quite different temperatures were used in the compilation of parachors reported by Quayle, no specific reference temperature for the parachor is established, and it is treated as a constant. Although eq 5 shows an explicit temperature dependence for $\mathbf{P}$, the values of $\mathbf{P}_0$ used by Escobedo and Mansoori[12] were obtained from Quayle's groups. We find it preferable to retain the temperature-independent assumption and use a broader range of temperatures in the training set. In fact, the temperature dependence of the parachor is quite small. We examined the temperature dependence of the parachor for 731 compounds and found that the average absolute deviation (AAD) of the parachor from its average value over the given temperature range was 0.79%; the corresponding average temperature range was 89 K. If the deviation from the average was divided by the specific temperature range for that compound, an AAD of 0.015%/K was obtained.

Two different training sets were chosen from which to develop the group values. The first training set was obtained from the DIPPR database using only those experimental surface tension values with an uncertainty <5%. Although the parachor is being correlated here, the experimental uncertainty in the liquid density is generally small and the uncertainty in the parachor will not, therefore, be significantly different. Additionally, DIPPR uncertainties are assigned to the entire temperature-dependent data set and represent uncertainties for the least certain values within that set. A 95% reliability was chosen to ensure that the training set obtained would have a breadth large enough that the groups would be found in many different structural environments within the molecules, but narrow enough that the accuracy of the resultant predictions would not be compromised by experimental uncertainties. This training set included 649 different compounds with a total of 8697 σ values at various temperatures. For the second training set, the allowable error was set at <1%. This training set consisted of 406 compounds and 6073 data points. Because not all of the desired groups were represented by the molecules in this training set, the allowable error was relaxed to <3% for a few families containing these specific groups. The two training sets will be referred to hereafter as the 5% TS and the 1% TS, respectively.

Corresponding values of $\mathbf{P}$ for the compounds were regressed from the experimental σ values using the accepted DIPPR correlation for the saturated liquid density

Table 4. New Descriptors and Their Increments for the Parachor

nonring C		5% TS	1% TS	nitrogen		5% TS	1% TS
CH$_3$		55.25	55.24	R−NH$_2$	primary R	44.98	45.40
>CH$_2$	$n = 1-11$	39.92	39.90	R−NH$_2$	sec R	44.63	45.85
>CH$_2$	$n = 12-20$	40.11	40.11	R−NH$_2$	tert R	46.44	46.40
>CH$_2$	$n > 20$	40.51	40.11	A-NH$_2$	attached to arom ring	46.53	43.90
>CH−		28.90	28.88	>NH	nonring	29.04	29.54
>C<		15.76	15.65	>NH	ring	31.97	33.49
=CH$_2$		49.76	49.87	>NH	in arom ring	33.92	34.12
=CH−		34.57	34.61	>N−	nonring	10.77	8.03
=C<		24.50	24.46	>N−	ring	15.71	16.05
=C=		24.76	24.53	−N=	nonring	23.24	24.44
≡CH		43.64	43.66	>N	aromatic	26.49	26.46
≡C−		28.64	28.66	HC≡N	hyd cyanide	80.94	80.94
Branch Corrections				−C≡N		65.23	66.15
per branch		−6.02	−6.02	−C≡N	aromatic	67.54	67.42
sec−sec adjacency		−2.73	−2.75				
sec−tert adjacency		−3.61	−3.72				
tert−tert adjacency		−6.10	−6.19				
nonaromatic ring C		5% TS	1% TS	nitrogen and oxygen		5% TS	1% TS
−CH$_2$−		39.21	39.53	−C=ONH$_2$	amides	93.43	93.44
>CH−		23.94	22.06	−C=ONH−	amides	73.64	73.65
>C<		7.19	5.11	−C=ON<	amides	57.05	56.33
=CH−		34.07	33.33	−NHCHO		91.69	91.69
=CH−		18.85	24.82	>NCHO		77.12	77.14
>CH−	fused ring	22.05	20.57	−N=O		64.32	64.49
Ring Corrections				−NO$_2$		73.86	72.31
three-member ring		12.67	13.12	−NO$_2$	aromatic	75.05	74.17
four-member ring		15.76	15.00				
five-member ring		7.04	7.74				
six-member ring		5.19	5.42				
seven-member ring		3.00	0.79				
aromatic ring C		5% TS	1% TS	sulfur		5% TS	1% TS
>CH		34.36	34.37	R−SH	primary R	66.89	66.87
>C−		16.07	16.08	R−SH	sec R	63.34	63.37
−C−	fused arom/arom	19.73	19.73	R−SH	tert R	65.33	65.37
−C−	fused arom/aliph	14.41	14.41	−SH	aromatic	68.30	68.24
Arom Ring Corr				−S−	nonring	51.37	51.29
ortho		−0.60	−0.60	−S−	ring	51.75	50.27
para		3.40	3.40	−S−	aromatic	51.47	52.70
meta		2.24	2.24	>S=O	nonring	72.21	72.22
Substituted Naphthalene Corr		−7.07	−7.07	>SO$_2$	nonring	93.20	93.53
				>SO$_2$	ring	90.13	88.82
oxygen		5% TS	1% TS	silicon		5% TS	1% TS
−OH	alc, primary	31.42	30.20	SiH$_4$		105.11	105.11
−OH	alc, sec	22.68	22.60	>SiH−		54.50	55.01
−OH	alc, tertiary	20.66	18.93	>Si<		44.93	44.07
−OH	phenol	30.32	19.25	>Si<	ring	28.64	29.44
−O−	nonring	20.61	20.72				
−O−	ring	21.67	20.97				
−O−	aromatic	23.54	23.43				
>C=O	nonring	47.02	46.92				
>C=O	ring	50.04	49.22				
O=CH−	aldehyde	66.06	65.96				
CHOOH	formic	94.01	93.93				
−COOH	acid	74.57	74.48				
−OCHO	formate	82.29	82.42				
−COO−	ester	64.97	64.96				
−COOCO−	acid anhyd	115.07	115.11				
−OC(=O)O−	ring	84.05	84.10				
halogen		5% TS	1% TS	other inorganics		5% TS	1% TS
−F		21.81	19.98	>PO$_4$−		115.59	115.67
−Cl		26.24	50.98	>P−		48.84	49.35
−Br		51.16	65.73	>B−		22.65	28.19
−I		54.56	90.82	>Al−		25.06	25.15
−F	aromatic	66.30	27.29	−ClO$_3$		106.03	107.87
−Cl	aromatic	70.39	54.07				
−Br	aromatic	90.84	72.07				
−I	aromatic	92.04	92.08				

and the Soave equation of state for the saturated vapor density. These **P** values were then correlated as a function of the defined descriptors. A commercial QSPR software package, called TSAR, marketed by Oxford Molecular, was used to manipulate the molecular descriptors and perform the statistical analysis and regression. The QSPR software is essentially a structurally knowledgeable spreadsheet. With groups defined as column heads and the 2D structure

Table 5. Comparison of σ Values Calculated with the New Descriptors and the Original Quayle Groups (Top Number = 5% TS Correlation; Bottom Number = 1% TS Correlation; Number in Parentheses = AAD for 5% Data Set Predicted from 1% TC Values)

family	no. of compounds	no. of values	T range/K	AAD/%		max. AD/%	
				Quayle	new	Quayle	new
hydrocarbons (nonring)	104	1617	90−503	2.10	1.24 (1.25)	29.79	11.92 (11.77)
	85	1419	118−503	2.18	1.13	17.66	11.77
hydrocarbons (ring)	28	441	239−423	3.72	1.68 (2.19)	17.22	13.77 (12.59)
	18	283	239−423	3.78	1.70	17.22	11.77
aromatics	51	976	243−673	3.78	1.84 (1.98)	12.57	12.89 (14.29)
	44	937	243−609	3.73	1.33	10.73	7.67
alcohols	60	1042	273−533	7.55	5.69 (7.74)	62.73	28.16 (51.40)
	30	615	273−508	4.96	3.24	35.79	28.42
aldehydes	5	47	283−373	2.53	2.03 (2.06)	8.26	5.39 (5.31)
	5	43	283−373	2.74	2.04	8.26	5.23
ketones	21	360	273−523	4.28	2.55 (2.63)	22.46	14.66 (14.76)
	10	246	273−500	3.58	1.59	22.46	14.76
esters	49	547	251−511	2.66	1.85 (1.93)	15.41	13.18 (12.51)
	32	465	252−473	2.29	1.47	10.51	8.60
acids	21	226	288−473	4.67	2.63 (2.71)	23.37	18.05 (16.14)
	11	167	288−453	5.22	2.04	23.30	11.09
ethers	35	345	193−523	2.85	2.32 (2.49)	19.81	20.97 (21.48)
	15	231	193−523	2.66	2.25	19.81	21.48
anhydrides and epoxides	8	98	253−473	6.43	3.76 (6.47)	49.19	36.69 (38.68)
	6	68	223−383	5.31	1.12	49.19	6.46
halides	88	930	193−477	12.70	7.05 (16.70)	64.29	33.61 (98.81)
	23	336	273−443	8.01	8.69	35.45	74.46
N-containing	111	1292	201−693	6.39	4.91 (5.27)	66.15	48.43 (65.99)
	66	671	240−499	5.72	3.73	34.88	33.13
S-containing	40	428	176−394	12.44	2.32 (4.12)	126.11	34.87 (34.60)
	36	309	231−394	12.10	1.49	126.11	13.22
Si-containing	19	275	89−399	19.74	4.63 (4.68)	43.60	18.94 (16.05)
	16	210	89−413	20.31	4.11	43.60	16.04
other inorganics	9	73	157−499	12.10	3.34 (3.34)	31.27	22.57 (17.68)
	9	73	157−499	12.10	3.34	31.27	22.69
Cumulative	649	8697	89−609	6.93	3.19 (4.33)	39.51	22.27
	406	6073		6.31	2.62	32.45	18.66

of the compounds as row heads, TSAR performed the descriptor identification and counting within the training set and regressed the values of the structural increments to the parachor.

Values obtained for the descriptor increments to the parachor are shown in Table 4. The results of the regression in terms of agreement between computed and experimental σ values are given in Table 5. The results in this table show that there is a significant improvement in the correlation with the new parameter set over that obtained with the Quayle parameters. The fact that the resultant AAD is greater than the uncertainty of the TS suggests that complete independence of descriptors with respect to varying molecular environments has not been fully achieved yet. This is particularly true of strongly polar groups such as alcohols, anhydrides, and halogens. However, the known accuracy of the TS was used in an iterative manner to ascertain which descriptors needed to be redefined in order to improve the correlation within families closer to the uncertainty inherent in the TS. Deductions about specific descriptors within families are identified below.

For hydrocarbons the 1% TS included 103 (85 nonring, 18 ring) compounds, and the correlation of σ yielded an AAD about half that produced by the Quayle coefficients. When applied in a predictive mode to the 5% data set, there is little loss in predictive capability for nonring compounds, but the AAD increases from 1.7% to 2.2% for the ring compounds. The AAD for the 5% TS is only 1.7%, suggesting a broader applicability for its parameters. Results for aromatic compounds show a pattern similar to that of the ring hydrocarbons. TS accuracy required that CH_2 contributions be divided into three groups ($n = 1-11$, $n = 12-20$, and $n > 20$) instead of the two ($n < 12$ and $n > 12$) in Quayle's table. Likewise, addition of fused ring aliphatic and aromatic groups was warranted by the TS.

Alcohols were particularly troublesome. The AAD for the 1% TS was 3.2% compared to 5.7% for the 5% TS. However, when the 1% TS values were used to predict values in the 5% data set, the AAD increased to 7.7%, indicating poor extrapolation capability. The 5% value is therefore recommended as a more general value, but this poor extrapolation ability suggests that a redefinition of the descriptor is warranted. In fact, we found that the parachor for alcohols varies considerably depending upon groups attached several C atoms away from the OH linkage. Table 4 shows that in this work we used four OH descriptors depending upon the type of linkage.

Only a few aldehydes were available in the TS, so the 5% and 1% results are similar. The 1% TS produced 1.6% AAD for ketones. Extrapolation to the 5% data set was very good, producing an AAD comparable to that obtained from the correlation of the 5% TS. This is likewise true of the esters, acids, and ethers. There is significant reduction in the AAD for the ketones and acids using the new groups compared to Quayle's groups. Again, the known accuracy of the TS required separating ring and nonring oxygen descriptors as shown in Table 4. Furthermore, it was found that the first ester (formate) and acid (formic) had to be treated with a different descriptor than the remainder of the family. Only slight improvement over Quayle's descriptors was achieved with ethers and esters.

Halides were found to be very sensitive to neighboring linkages. The TS clearly required separation of aromatic and aliphatic linkages into two groups (cf. Table 4), but additional attempts to define different descriptors as CX_n had marginal effect upon the AAD. Effects upon the

Table 6. Prediction of σ for Multifunctional Compounds Using the New Descriptors and Quayle's (The Top Number Is Predicted Using the Values from the 5% TS, and the Bottom Number Results from the 1% TS Increments)

no. of	no. of	T	AAD/%		max. deviation/%	
compounds	points	range/K	Quayle	new	Quayle	new
78	676	173−480	7.48	6.52	71.93	33.86
				8.05		33.76

parachor by multiple halogens are felt over several C−C linkages. The 5% TS gives an average value for halogen increments that can be used with some reliability. The 1% TS increments extrapolate poorly, as seen by the decay of the AAD in Table 5 when using them to predict the 5% data set.

Descriptors for nitrogen-containing compounds were divided according to chemical functionality, bond order, and ring constituency as shown in Table 4. These were warranted by the quality of the TS even though the improvement over Quayle's groups, in which there are only three divisions, is modest.

In the case of families containing S and Si, substantial improvement over the Quayle values was achieved by using functional group descriptors instead of atomic ones. The AAD was reduced by a factor of 4 or 5, in comparison to the Quayle results, for these and other families containing inorganic elements. It was necessary to separate the thiol components into primary, secondary, and tertiary groups, as was done with the alcohols. Ring structural components were also found to be distinguishable within the accuracy of the TS from the nonring groups.

As a further test of the new parameters, particularly with respect to extrapolation to other molecular environments, the new parameters were used to predict surface tension for compounds not in the 5% training set. Multifunctional compounds were reserved out of the training set for this purpose. Table 6 shows the results for the σ predictions for these multifunctional compounds. Results using parameters obtained from the 5% TS extrapolate better to this multifunctional test set than those obtained from the 1% TS, and they are moderately better than the atomic descriptors used by Quayle. This suggests that the broader structural environments included in the 5% TS aid in the ability to extrapolate to new compounds. It also suggests that additional delineation of descriptors based on the molecular environment is warranted in future work.

Summary

The availability of commercial QSPR software makes development of structurally based correlations fast and convenient. Much of the statistical work and descriptor calculations can be handled by these programs. However, an evaluated database is key to determination of accurate descriptor increments for use in the resultant correlation. An evaluated database allows training sets of known accuracy to be selected so that an optimization between elimination of experimental errors from the descriptor values and the breadth of the local descriptor environment can be achieved. By knowing the accuracy of the training set, improvements on the descriptors used can be made within the tolerance suggested by the known experimental error. A simple application of the capability created by combining the DIPPR evaluated database with a QSPR program was presented in the form of improved descriptors for the parachor used in the prediction of surface tension. Increments for the group and structural descriptors of this modified MSQ method were regressed from surface tension data for 649 compounds over a wide range of temperatures using 8697 points. The average absolute deviation from the experimental data was 3.19% compared to 6.93% with the Quayle atomic and structural increments previously available.

Acknowledgment

We thank Costa Tsonopoulos for his suggestion of refining Quayle groups.

Literature Cited

(1) Macleod, D. B. Relation between surface tension and density. *Trans. Faraday Soc.* **1923**, *19*, 38.

(2) Sugden, S. J. The variation of surface tension with temperature and some related functions. *Chem. Soc.* **1924**, *125*, 32.

(3) Quayle, O. R. The parachors of organic compounds. *Chem. Rev.*, **1953**, *53*, 439−591.

(4) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.

(5) Murugan, R.; Grendze, M.; Toomey, J. E., Jr.; Katritzky, A. R.; Karelson, M.; Lobanov, V.; Rachwal, P. Predicting physical properties from molecular structure. *CHEMTECH* **1994**, *24* (6), 17−23.

(6) Mumford, S. A.; Phillips, J. W. C. Evaluation and interpretation of parachors. *J. Chem. Soc.* **1929**, 2112−2133.

(7) Lee, S.-T.; Chien, M. C. H. A new multicomponent surface tension correlation based on scaling theory. *SPE/DOE Symp. on EOR*; SPE/DOE 12643; Tulsa, OK (Apr 15−18, 1984).

(8) Hugill, J. A.; van Welsenes, A. J. Surface tension: a simple correlation for natural gas + condensate systems. *Fluid Phase Equilib.* **1986**, *29*, 383.

(9) Gasem, K. A. M.; Dulcamara, P. B.; Dickson, B. K.; Robinson, R. L., Jr. Test of prediction methods for interfacial tensions of CO_2 and ethane in hydrocarbon solvents. *Fluid Phase Equilib.* **1989**, *53*, 39.

(10) Fanchi, J. R. Calculation of parachors for compositional simulation: an update. *SPE Reservoir Eng. J.* **1990**, 433.

(11) Ali, J. K. Prediction of parachors of petroleum cuts and pseudocomponents. *Fluid Phase Equilib.* **1994**, *95*, 383.

(12) Escobedo, J.; Mansoori, G. A. Surface tension prediction for pure fluids. *AIChE J.* **1996**, *42*, 1425−1433.

(13) *DIPPR 801 Policies and Procedures Manual*; Brigham Young University: Provo, UT, May, 2000.

(14) Boudh-Hir, M. E.; Mansoori, G. A. Statiscal mechanics basis of Macleod's formula. *J. Phys. Chem.* **1990**, *94*, 8362.

(15) Lydersen, A. L. Estimation of critical properties of organic compounds. *University Wisconsin Coll. Eng., Eng. Exp. Stn. Rept. 3* (Madison, Wis., April 1955).

(16) Joback, K. G. S.M. Thesis in chemical engineering, Massachusetts Institute of Technology, Cambridge, MA, June 1982.

Received for review August 1, 2000. Accepted November 9, 2000. We gratefully acknowledge the support of this work by the DIPPR 801 project.

JE000232D