Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach

James A. Platts,[†] Darko Butina,[‡] Michael H. Abraham,*[,†] and Anne Hersey[‡]

Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, U.K., and Science Development Group, Glaxo Wellcome Research and Development, Park Road, Ware SG12 0DP, U.K.

Received December 16, 1998

Additive models for the estimation of Abraham's molecular descriptors R_2, π_2^H, $\Sigma\alpha_2^H$, $\Sigma\beta_2^H$, $\Sigma\beta_2^O$, and log L^{16} have been developed. For five of the six descriptors, one set of 81 atom and functional group fragments is capable of reproducing experimentally derived results with correlation coefficients ranging from 0.95 to 0.99. However, one descriptor, $\Sigma\alpha_2^H$, required an entirely separate set of 51 fragments to be developed, resulting in a correlation coefficient of 0.97. Of particular importance is the speed of calculation (approximately 700 molecules/min), allowing so-called "high-throughput screening". Several applications of this model for molecules containing intramolecular interactions are discussed.

INTRODUCTION

The solvation parameter approach to the description of solvent−solute interactions is part of the wider field of linear free energy relations (LFERs) or quantitative structure−activity relations (QSARs). The method of Abraham[1] holds that a solvation property SP can be modeled by reducing it to a sum of specific interaction terms, i.e.,

$$\log \mathrm{SP} = c + rR_2 + s\pi_2^{\mathrm{H}} + a\Sigma\alpha_2^{\mathrm{H}} + b\Sigma\beta_2^{\mathrm{H}} + l \log L^{16} \quad (1)$$

where SP is a property of a series of solutes in a given solute system and where the solute descriptors are defined as follows:

R_2 is the excess molar refraction, i.e., the molar refraction of the solute minus the molar refraction of an alkane of equivalent volume.

π_2^H is a combined dipolarity/polarizability descriptor.

$\Sigma\alpha_2^H$ is the overall or summation solute hydrogen bond acidity.

$\Sigma\beta_2^H$ is the overall or summation solute hydrogen bond basicity.

log L^{16} is the solute gas−hexadecane partition coefficient.

Equation 1 is typically used in transport processes involving transfer of solutes from the gas phase to a condensed phase.

An analogous equation has been developed for transport processes involving two or more solution, liquid, or solid phases:

$$\log \mathrm{SP} = c + rR_2 + s\pi_2^{\mathrm{H}} + a\Sigma\alpha_2^{\mathrm{H}} + b\Sigma\beta_2^{\mathrm{H}} + vVx \quad (2)$$

where Vx is McGowan's characteristic volume[2] in units of ($cm^3mol^{-1}/100$). For transfer between water and wet solvents, such as octanol or ethyl acetate, the $\Sigma\beta_2^H$ descriptor is replaced by $\Sigma\beta_2^O$ for certain functional groups (e.g., pyridines, sulfoxides) whose basicity is found to change substantially between wet and dry solvents.[1]

The solute descriptors represent the solute effect on various solute-phase interactions. Hence the coefficients *c*, *r*, *s*, *a*, *b*, and *l* or *v* correspond to the complementary effect of the phase on these interactions. The coefficients can then be regarded as system constants, which characterize the phase and contain chemical information about the phase in question.

The system constants can be interpreted as follows. The *r*-coefficient shows the tendency of the phase to interact with solutes through π- and n-electron pairs. Usually the *r*-coefficient is positive, but for phases which contain fluorine atoms the *r*-coefficient can be negative. The *s*-coefficient gives the tendency of the phase to interact with dipolar/polarizable solutes, the *a*-coefficient denotes the hydrogen bond basicity of the phase (because acidic solutes will interact with a basic phase), and the *b*-coefficient is a measure of the hydrogen bond acidity of the phase (because basic solutes will interact with an acidic phase). The *l*-coefficient is a combination of exoergic dispersion forces that make a positive contribution to the *l*-coefficient, and an endoergic cavity term that makes a negative contribution. In the event, the dispersion interaction nearly always dominates, so that the *l*-coefficient is positive. The only phase for which a negative value is observed[3] is for solution of gases and vapors into water. Since the *l*-coefficient varies between −0.21 for water at 25 °C and +1.00 for hexadecane at 25 °C, it seems to be a useful measure of hydrophobicity of the condensed phase. Similarly, the *v*-coefficient is also a resultant of dispersion and cavity effects.

It is important to note that for gas-phase processes, the *s*-, *a*-, and *b*-coefficients must always be positive (or zero), because interactions between the phase and a solute will increase the solubility of a gaseous solute. The *r*-constant is an exception, because it is tied to hydrocarbons as a zero; hence fluoro or chloro compounds as phases may give rise to a negative *r*-constant. The coefficients in the solvation

[†] University College London.
[‡] Glaxo Wellcome Research and Development.

10.1021/ci980339t CCC: $18.00
Published on Web 07/23/1999

parameter equation are therefore not just fitting constants but must obey general chemical principles.

Two examples serve to illustrate the chemical information contained in the system constants. Partition of vapors into chloroform can be described by the equation[4]

$$\log L(\text{chl}) = 0.116 - 0.467R_2 + 1.203\pi_2^{\text{H}} + 0.138\Sigma\alpha_2^{\text{H}} + 1.432\Sigma\beta_2^{\text{H}} + 0.994 \log L^{16} \quad (3)$$

indicating that bulk chloroform interacts considerably less with π- and n-electron pairs than do bulk alkanes, is very dipolar/polarizable and a strong hydrogen bond acid, but has very little hydrogen bond basicity. The important water/1-octanol system is characterized by the following equation[5]

$$\log P(\text{oct}) = 0.088 + 0.562R_2 - 1.054\pi_2^{\text{H}} + 0.034\Sigma\alpha_2^{\text{H}} - 3.460\Sigma\beta_2^{\text{O}} + 3.814Vx \quad (4)$$

Thus, 1-octanol is revealed to be a very much weaker hydrogen bond acid and less dipolar/polarizable than water, more able to interact with π- and n-electron pairs, and of almost exactly the same hydrogen bond basicity as water. The large v-coefficient means that octanol is better able to interact with solutes by dispersion forces than is water and/or that the energy required to create a given sized cavity in octanol is less than in water.

Traditionally, the solute descriptors are derived from experimental measurements, such as changes in the infrared stretching frequency H−X upon formation of complexes B···H−X, water/solvent and gas/solvent partitions, and GC, GLC, and HPLC chromatography. While this approach delivers high-quality descriptors for most molecules, a number of disadvantages exist. First, one must physically obtain a sample of the molecule of interest. Second, certain measurements may not be suited to certain types of molecule; e.g., UV spectroscopy requires a chromophore in the molecule. Third, the process of measurement and derivation can be laborious and time-consuming, time and effort that may be better spent actually measuring the solvation property of interest. This also limits the possibility of such approaches being used in so-called high-throughput screening, the rapid evaluation of molecular properties for large libraries of compounds.

Accordingly, several methods for the estimation of solute descriptors without recourse to experimental data have been reported. *Vx* is calculated trivially by summing atomic and bond contributions.[2] R_2 can be calculated from the solute's refractive index or molar refraction.[6] Sevcik and co-workers have reported an additive scheme for the estimation of log L^{16},[7] and a neural network approach to estimating π_2^{H}.[8] The former approach simply adds contributions to log L^{16} from a given set of fragments, the contributions being derived from multivariate linear regression analysis (MLRA). The latter takes a number of structural and quantum mechanical properties as input, combining them either linearly via MLRA or nonlinearly via a feed-forward neural network. It is the former, group contribution approach we take here.

A number of such additive group contribution schemes have been proposed for various solvation properties, in addition to Sevcik's log L^{16} estimation. These include

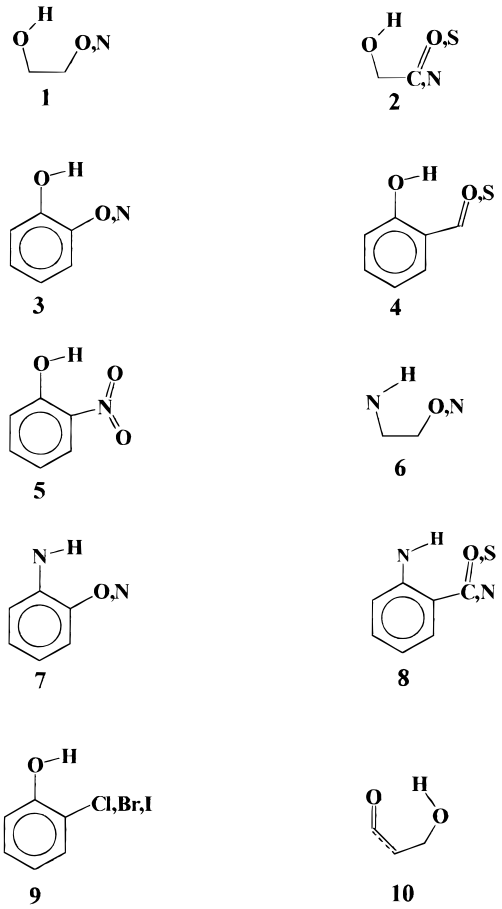


Figure 1. Hydrogen bond types employed in regressions.

Klopman et al's models of log P(oct) [9] and water solubility[10] and Ghose et al's model of log P(oct) and molar refraction.[11] The success of such an approach depends critically on the set of fragments chosen. In this work, we have taken Klopman's generally applicable, if slightly less accurate, solubility model II (Table IV, p 478 of ref 10) fragments as our starting point (Figure 1). This is because our primary aim is a reliable, robust estimation of the above descriptors for any organic molecule, not accurate predictions for molecules restricted to be similar to the training set.

METHOD OF CALCULATION

Among the requirements of a program to calculate the above descriptors are speed, generality, and ease of updating the model. Accordingly, we have taken advantage of Daylight Toolkit programs previously written by Bradshaw, Johnson, and Butina.[12] These allowed predefined structural features to be read from an external file, separate from the main program, and counts of such substructures to be multiplied by a contribution to a given molecular property, such as $\Sigma\alpha_2^{\text{H}}$, $\Sigma\beta_2^{\text{H}}$ or π_2^{H}. In this manner, key fragments may be added or modified with no change in the main calculation program.

The first 31 of Klopman's 33 fragments (i.e., excluding his alkane and hydrocarbon corrections) were defined as SMARTS strings. These definitions were checked by ensuring they reproduced the reported water solubilities for 21 molecules.[10] A C program[12] was written to read molecules as SMILES strings[13] and count the number of matches of

Table 1. Results of Regression Using Klopman's 31 Fragments

	R^2(adj)	rms error	F statistic	n	min	max
R_2	0.955	0.134	2315.1	3375	−1.37	4.62
π_2^H	0.877	0.202	673.6	2875	−0.54	4.15
$\Sigma\alpha_2^H$	0.827	0.101	569.3	3692	0.00	2.10
$\Sigma\beta_2^H$	0.877	0.137	597.9	2541	0.00	3.20
$\Sigma\beta_2^O$	0.884	0.138	640.4	2568	0.00	4.52
log L^{16}	0.986	0.275	4632.0	1947	−0.83	29.98

each of the 31 fragments within a molecule. This program was applied to all molecules contained within our in-house database of solute descriptors, a total of over 4200 molecules. The resulting counts were used as independent variables in a standard least-squares regression against all available solute descriptors (not all descriptors are present for all molecules—see Table 1 for the numbers used in each regression). All regressions employed the JMP Discovery software, published by SAS Software. Fragment contributions were omitted from the final model if their significance, as measured by a *t*-test, did not exceed 95%.

Subsequent refinement of the model used the same approach, i.e., generation of fragment counts followed by least-squares regression of these counts against database descriptors. These refinements were designed to remove structure from the residual, identifying classes of molecules modeled poorly by the initial fragments and modifying the fragment set accordingly. For example, Klopman's original definition placed all NH_2 groups in one fragment, but we have separated NH_2 groups bonded to aliphatic and aromatic atoms. Similarly, aliphatic and aromatic nitro and nitrile groups have been separated, and distinctions between cyclic and acyclic amides have been drawn. Further to these basic fragments, a number of important interactions have been identified.

For five of the six solute descriptors of interest, this approach of refining Klopman's set of fragments proved satisfactory (see below). However, the hydrogen bond acidity descriptor, $\Sigma\alpha_2^H$, proved not to be amenable to modeling using such fragments, and an entirely new set of fragments had to be developed. Initially, these consisted of the most fundamental acidic atom types (essentially OH, NH, and NH_2 in various environments). Subsequent refinement of the model proceeded as before, with corrections to and distinctions between these fundamental acid types iteratively defined until a satisfactory regression was obtained.

Table 2. Modified Fragment Descriptions

	fragments	comments	count		fragments	comments	count
1	−CH_3	sp^3	4521	42	I	any	79
2	>CH_2	sp^3	7468	43	−OC(O)−	noncyclic ester	434
3	>CH−	sp^3	1122	44	−OC(O)−	lactone	17
4	>C<	sp^3	536	45	O═P(OR)	phosphate	44
5	═CH_2	sp^2	195	46	−OC(O)O−	carbonate	4
6	═CH−	sp^2 or aromatic	8859	47	−C(O)OH	carboxylic acid	139
7	═C<	sp^2 or nonfused aromatic	5067	48	−NC(O)−	aromatic amide	34
8	C	fused aromatic	869	49	−NC(O)−	noncyclic aliphatic amide	247
9	≡C	sp	157	50	−NC(O)−	lactam	222
10	−NH_2	sp^3, connected to aliphatic	137	51	−S(O)(O)N−	sulfonamide	16
11	−NH_2	sp^3, connected to aromatic	156	52	−NC(O)N−	urea	107
12	>NH	sp^3, connected to aliphatic	96	53	−C(O)O−	carbamate	18
13	>NH	sp^3, connected to aromatic	116	54	−C(O)NC(O)−	imide	7
14	>NH	pyrrole	189	55	−C(O)C═CC(O)−	quinone	13
15	>N−	sp^3, connected to aliphatic	84	56	−CX_2−	CX_2[a]	514
16	>N−	sp^3, connected to aromatic	27	57	>CXCX<	XCCX[a]	401
17	>N−	pyrrole	76	58	steroid	fused ring system	43
18	═N	sp^2, noncyclic	24	59	H−bond 1	b	56
19	═N	sp^2, cyclic	12	60	H−bond 2		42
20	═N−	pyridine	406	61	H−bond 3		14
21	N≡C−	sp, connected to aliphatic	29	62	H−bond 4		8
22	N≡C−	sp, connected to aromatic	48	63	H−bond 5		10
23	−NO_2	connected to aliphatic	19	64	H−bond 6		10
24	−NO_2	connected to aromatic	154	65	H−bond 7		10
25	−ONO_2	nitrate	8	66	H−bond 8		20
26	−OH	any	729	67	H−bond 9		13
27	−O−	sp^3, noncyclic	823	68	>C(OH)C(OH)<	1,2-diol	23
28	−O−	sp^3, cyclic	86	69	n:n	1,2 aromatic interaction[c]	14
29	−O−	aromatic	54	70	x:x	1,2 aromatic interaction[d]	3
30	═O	sp^2	1814	71	n:c:n	1,3 aromatic interaction[e]	79
31	−S−	sp^3	116	72	x:c:x	1,3 aromatic interaction[f]	28
32	−S−	aromatic	36	73	x:c:c:x	1,4 aromatic interaction[g]	135
33	═S	sp^2	9	74	Y−c:c−Y	*ortho* interaction[h]	347
34	>S═	sp^2	14	75	Y−c:c:c−Y	meta interaction	357
35	−$OS(O_2)$−	sulfonate	7	76	Y−c:c:c:c−Y	para interaction	426
36	S	other	35	77	O═PN	phosphamide	6
37	P	any	28	78	N−c:n	2-aminopyridine	47
38	−F	connected to aliphatic	397	79	OH−C−c	benzyl alcohol	46
39	−F	connected to aromatic	156	80	O═N	*N*-oxide	6
40	Cl	any	630	81	−O−c:c−O−	1,2-dimethoxy	22
41	Br	any	220				

[a] X defined as halogen, NO_2, C≡N, or CF_3. [b] See Figure 1 for hydrogen bond definitions. [c] Pyridazine type. [d] Isoxazole type. [e] Pyrimidine type. [f] Oxazole type. [g] Pyrazine type. [h] Y defined as any heteroatom, i.e. halogen, N, O, S.

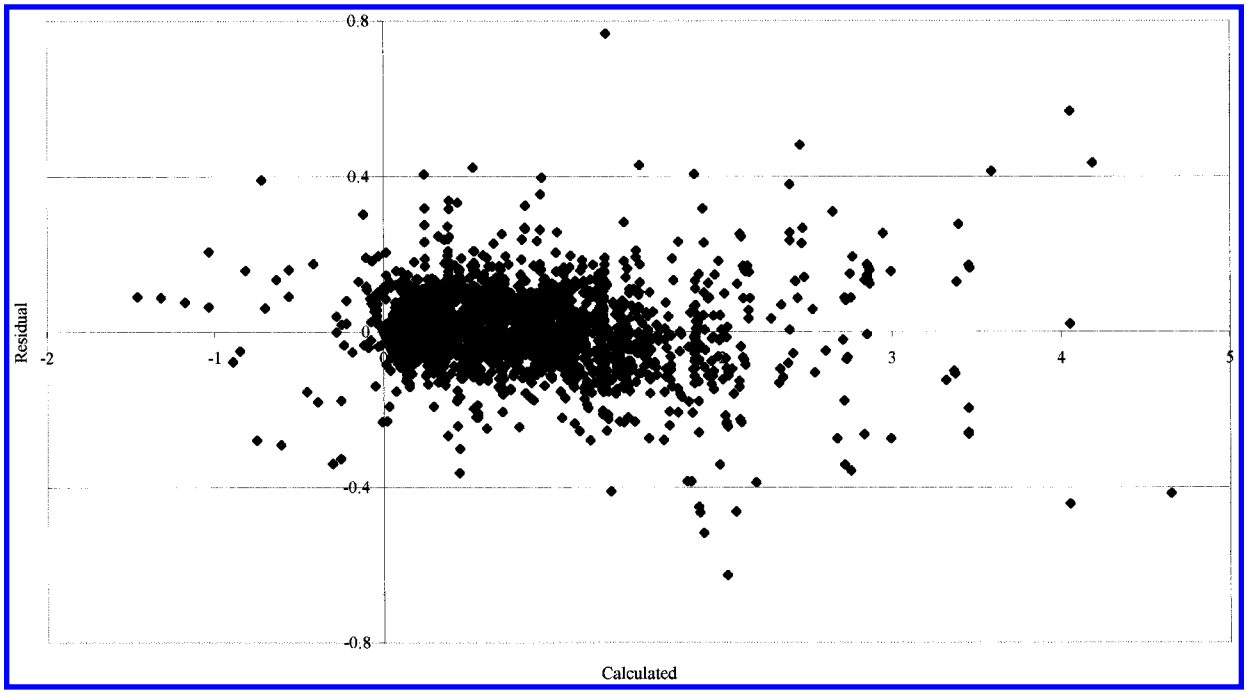


Figure 2. Residual plot for R_2 regression.

Table 3. Results of Regression Using 81 Modified Fragments

	R^2(adj)	rms error	F statistic	n
R_2	0.979	0.093	2574.4	3375
π_2^H	0.921	0.163	570.3	2875
$\Sigma\alpha_2^H$	0.881	0.083	465.7	3692
$\Sigma\beta_2^H$	0.914	0.121	417.7	2541
$\Sigma\beta_2^O$	0.919	0.123	435.6	2568
log L^{16}	0.989	0.2417	3333.5	1947

RESULTS AND DISCUSSION

Initial results of regression of Klopman's 31 fragments against the six descriptors are reported in Table 1. Excellent correlation statistics are found for R_2 and log L^{16}, precisely those descriptors found to be additive in previous studies. Thus, one could use this simple modification of Klopman's solubility model to estimate these descriptors directly. However, the remaining four descriptors are rather poorly modeled by these 31 fragments, with $\Sigma\alpha_2^H$ in particular showing unacceptably large errors. The range of descriptors covered in these and all subsequent regressions is also shown in Table 1.

Modification of these 31 fragments, as described in the previous section, led to a total of 81 fragments; these are described in Table 2. The aim of these modifications is to improve on the results presented in Table 1 without losing the general applicability of the original fragment set. To this end, we have separated some of Klopman's fragments into two or more new fragments where this was found to significantly improve the fit. We have also added several more fragments and interaction terms, each designed to be as general as possible. The number of occurrences of each fragment in the solute database is also reported.

Results of the regression of these 81 fragment counts against the data used in Table 1 are presented in Table 3. All six descriptors show an improved fit, with larger correlation coefficients and smaller root mean square (rms) errors, using this extended fragment set. However, only R_2 showed an increase in Fischer's F statistic, while the others show small decreases in this statistic. This suggests reduced significance of the fits, despite the improvement in other statistics. We assign this to the fact that in each regression several fragments are insignificant (using the t-test definition). We choose to retain these fragments in the regressions, though not in the final model, to maintain consistency.

Large correlations (≥ 0.70) between fragment counts are found for just three pairs of fragments: no. 35 (F) vs no. 49 (CX_2) with $r = 0.86$, no. 67 (ortho) vs no. 68 (meta) with $r = 0.82$, and no.15 (cyc NH) vs no. 47 (cyc amide) with $r = 0.79$. A plot of the residual errors from the fit against the calculated values of R_2 is shown in Figure 2, with no structure whatsoever apparent. This gives confidence that the model developed is not biased in any way and that all the major features of the database descriptors have been taken into account. The analogous plots for the other five descriptors show a similar lack of structure.

The coefficients and standard errors of these 81 fragments from regression against R_2, π_2^H, $\Sigma\beta_2^H$, $\Sigma\beta_2^O$, and logL^{16} are reported in Table 4 ($\Sigma\alpha_2^H$ is not included in this table for reasons discussed below). These values are simply fitting constants and as such are not particularly interesting when taken in isolation. Some revealing trends are present, however, such as the increased basicity found on moving from primary to secondary and tertiary amines and the much greater basicity of aliphatic amines over that of anilines. The methylene (CH_2) contribution to log L^{16} is almost exactly 0.5 (to within 1.1 standard deviations), as suggested by Sevcik.[6] It is pleasing to note also that the methylene contribution to the other descriptors is almost exactly zero, as are the contributions of other sp^3 carbons.

Fragment contributions to $\Sigma\alpha_2^H$ are not reported in Table 4 because the associated regression was deemed unacceptable. On first inspection the statistics of this regression (Table 3) seem reasonable, only marginally worse than those for π_2^H, $\Sigma\beta_2^H$, and $\Sigma\beta_2^O$. What is not apparent in this table is the fact that, of the 3692 data points used in the regression, over half are exactly zero. Omitting these from the regression

Table 4. Regression Coefficients for R_2, π_2^H, $\Sigma\beta_2^H$, $\Sigma\beta_2^O$, and log L^{16} [a]

	R^2		π_2^H		$\Sigma\beta_2^H$		$\Sigma\beta_2^O$		log L^{16}	
fragment	coeff	sd	coeff	sd	coeff	sd	coeff	sd	coeff	sd
1	−0.104	0.002	−0.075	0.005	0.007	0.003	0.000	0.004	0.321	0.009
2	0.000	0.001	0.000	0.001	0.000	0.001	0.000	0.001	0.499	0.002
3	0.089	0.003	0.036	0.006	0.011	0.004	0.020	0.004	0.449	0.011
4	0.187	0.006	0.071	0.013	0.037	0.009	0.047	0.009	0.443	0.025
5	−0.045	0.001	−0.085	0.013	0.019	0.009	0.024	0.002	0.244	0.021
6	0.068	0.002	0.050	0.002	0.011	0.002	0.012	0.004	0.469	0.004
7	0.180	0.002	0.101	0.005	0.000	0.004	0.000	0.003	0.624	0.008
8	0.300	0.009	0.121	0.004	0.019	0.003	0.018	0.003	0.744	0.007
9	0.040	0.010	0.034	0.016	0.028	0.011	0.032	0.011	0.332	0.029
10	0.085	0.008	0.175	0.020	0.481	0.015	0.486	0.016	0.781	0.052
11	0.163	0.011	0.383	0.017	0.275	0.013	0.326	0.012	0.949	0.034
12	0.138	0.013	0.265	0.022	0.541	0.016	0.543	0.016	0.568	0.057
13	0.192	0.012	0.311	0.025	0.415	0.021	0.426	0.019	0.912	0.062
14	−0.030	0.012	0.221	0.024	0.316	0.020	0.267	0.018	1.250	0.048
15	0.220	0.010	0.323	0.025	0.653	0.018	0.655	0.018	0.400	0.071
16	0.346	0.011	0.295	0.038	0.321	0.036	0.338	0.031	0.869	0.074
17	0.083	0.018	0.265	0.021	0.392	0.016	0.338	0.016	0.794	0.081
18	0.117	0.011	0.125	0.044	0.200	0.031	0.202	0.031	−0.235	0.110
19	0.121	0.026	0.254	0.032	0.596	0.056	0.589	0.048	−0.240	0.092
20	0.046	0.019	0.223	0.012	0.321	0.010	0.300	0.009	0.574	0.024
21	0.000	0.006	0.694	0.036	0.242	0.025	0.245	0.025	0.757	0.064
22	0.000	0.018	0.390	0.028	0.103	0.020	0.093	0.021	0.732	0.064
23	0.200	0.015	0.000	0.050	−0.476	0.034	−0.595	0.029	0.278	0.066
24	0.210	0.023	−0.231	0.024	−0.525	0.018	−0.533	0.018	0.347	0.046
25	0.000	0.076	−0.476	0.033	−0.204	0.034	−0.202	0.033	0.000	0.067
26	0.061	0.012	0.247	0.011	0.307	0.008	0.311	0.008	0.672	0.020
27	0.014	0.006	0.185	0.009	0.211	0.008	0.226	0.007	0.360	0.015
28	0.013	0.010	0.185	0.009	0.331	0.021	0.330	0.023	0.359	0.022
29	−0.125	0.004	0.000	0.026	0.047	0.019	0.060	0.018	0.057	0.079
30	−0.041	0.014	0.370	0.011	0.334	0.009	0.339	0.009	0.495	0.021
31	0.330	0.006	0.189	0.016	0.168	0.013	0.175	0.013	1.258	0.031
32	0.116	0.008	0.000	0.048	0.043	0.037	0.083	0.036	0.848	0.072
33	0.364	0.022	0.618	0.072	0.071	0.044	0.069	0.005	0.954	0.124
34	0.413	0.033	1.065	0.055	0.448	0.067	0.319	0.040	2.196	0.247
35	0.000	0.056	−0.505	0.083	−0.188	0.021	−0.190	0.031	0.000	0.097
36	0.465	0.025	0.643	0.050	0.000	0.037	0.000	0.036	0.554	0.184
37	0.295	0.024	0.703	0.113	1.183	0.080	1.189	0.081	2.051	0.301
38	−0.180	0.032	−0.042	0.012	−0.036	0.008	−0.033	0.008	−0.143	0.019
39	−0.230	0.005	0.000	0.015	0.000	0.011	0.000	0.011	−0.147	0.027
40	0.023	0.007	0.082	0.009	0.000	0.080	0.000	0.007	0.669	0.016
41	0.196	0.005	0.161	0.013	−0.011	0.007	0.000	0.009	1.097	0.024
42	0.533	0.006	0.198	0.020	0.000	0.018	0.000	0.017	1.590	0.040
43	−0.113	0.007	−0.225	0.016	−0.206	0.013	−0.223	0.012	−0.390	0.028
44	0.000	0.009	0.360	0.050	−0.214	0.038	−0.169	0.036	0.406	0.113
45	−0.100	0.008	−0.240	0.048	−0.394	0.034	−0.408	0.034	−0.483	0.122
46	0.000	0.021	−0.190	0.096	−0.267	0.069	−0.298	0.069	0.000	0.146
47	−0.192	0.015	−0.412	0.020	−0.308	0.014	−0.312	0.014	−0.369	0.057
48	0.221	0.047	−0.076	0.026	−0.095	0.021	−0.038	0.019	0.000	0.068
49	0.000	0.011	0.175	0.022	−0.287	0.016	−0.292	0.016	0.603	0.067
50	0.061	0.009	−0.100	0.019	−0.231	0.014	−0.242	0.017	0.583	0.181
51	−0.111	0.032	−0.569	0.068	−0.446	0.054	−0.443	0.054	0.000	0.155
52	−0.110	0.020	−0.553	0.039	−0.076	0.029	−0.054	0.015	0.000	0.132
53	0.000	0.054	−0.588	0.012	−0.252	0.024	−0.251	0.023	0.000	0.008
54	0.000	0.098	−0.510	0.045	−0.148	0.053	−0.149	0.046	0.000	0.018
55	0.000	0.016	−0.411	0.062	−0.051	0.012	−0.050	0.010	0.000	0.009
56	−0.017	0.004	−0.050	0.009	−0.014	0.006	−0.016	0.006	−0.111	0.014
57	0.012	0.005	0.000	0.024	0.013	0.034	0.010	0.023	0.054	0.008
58	0.285	0.024	1.029	0.046	0.267	0.044	0.218	0.033	0.488	0.178
59	0.029	0.014	−0.067	0.022	0.000	0.012	0.000	0.017	−0.072	0.029
60	0.000	0.018	−0.095	0.019	−0.068	0.024	−0.090	0.020	−0.337	0.054
61	−0.069	0.007	−0.237	0.032	−0.079	0.019	−0.122	0.022	0.000	0.033
62	0.000	0.008	−0.344	0.020	−0.387	0.025	−0.403	0.027	−0.303	0.021
63	0.000	0.013	−0.276	0.016	−0.126	0.027	−0.120	0.018	−0.364	0.054
64	0.000	0.032	−0.102	0.011	0.000	0.014	0.000	0.011	0.062	0.019
65	0.000	0.022	0.000	0.013	−0.059	0.010	−0.027	0.009	0.000	0.030
66	0.000	0.017	−0.140	0.021	−0.045	0.011	−0.069	0.011	0.169	0.031
67	−0.100	0.018	−0.120	0.015	−0.130	0.020	−0.130	0.015	−0.400	0.052
68	−0.043	0.010	0.052	0.013	0.000	0.019	−0.018	0.008	0.100	0.097
69	0.092	0.008	0.024	0.008	−0.132	0.041	−0.094	0.041	−0.179	0.067
70	−0.113	0.012	0.047	0.011	−0.157	0.070	−0.141	0.070	0.000	0.048
71	0.000	0.015	−0.040	0.022	−0.098	0.018	−0.113	0.017	0.042	0.014
72	0.052	0.005	0.087	0.036	−0.170	0.042	−0.184	0.041	0.209	0.087

Table 4 (Continued)

	R^2		π_2^H		$\sum\beta_2^H$		$\sum\beta_2^O$		log L^{16}	
fragment	coeff	sd	coeff	sd	coeff	sd	coeff	sd	coeff	sd
73	0.000	0.009	−0.051	0.015	−0.089	0.011	−0.073	0.011	−0.058	0.024
74	0.000	0.017	−0.043	0.012	0.031	0.009	0.025	0.009	−0.081	0.023
75	0.000	0.012	−0.038	0.011	−0.035	0.008	−0.033	0.008	−0.026	0.013
76	0.000	0.016	0.000	0.009	−0.023	0.006	−0.025	0.006	0.000	0.016
77	−0.080	0.009	−0.452	0.056	−0.668	0.040	−0.668	0.040	0.000	0.007
78	0.185	0.015	0.098	0.027	−0.042	0.024	−0.057	0.018	0.149	0.059
79	0.000	0.010	0.000	0.029	0.131	0.021	0.129	0.020	−0.145	0.076
80	0.000	0.008	0.434	0.053	−0.408	0.076	−0.405	0.080	0.000	0.022
81	0.000	0.014	0.380	0.076	−0.216	0.022	−0.218	0.019	0.000	0.033
intercept	0.248	0.007	0.277	0.014	0.071	0.010	0.064	0.010	0.130	0.025

[a] Values entered as 0.0 are insignificant at the 95% level. sd = standard deviation.

Table 5. Fragments and Coefficients for $\sum\alpha_2^H$

	fragment	comment	coeff	sd[a]	count
1	−OH	connected to aliphatic	0.345	0.003	471
2	−OH	phenol	0.543	0.005	234
3	−NH2	connected to aliphatic	0.177	0.008	120
4	−NH2	aniline	0.247	0.005	144
5	>NH	connected to aliphatic, noncyclic	0.087	0.008	93
6	>NH	connected to aliphatic, cyclic	0.321	0.007	108
7	>NH	aniline	0.194	0.008	121
8	>NH	pyrrole	0.371	0.010	69
9	$-CO_2H$	carboxylic acid	0.243	0.006	130
10	$-CONH_2$	primary amide	0.275	0.011	63
11	−CONH−	secondary amide, aliphatic	0.281	0.009	134
12	−CONH−	aromatic amide	−0.091	0.012	12
13	$-SO_2NH$	primary or secondary	0.356	0.018	11
14	−NHCONH−	urea	−0.165	0.013	85
15	>NCONH−	urea	−0.119	0.015	17
16	−NHCOO−	carbamate	−0.105	0.019	9
17	−NHC(=N)N<	guanidine	0.170	0.089	26
18	≡CH	alkyne	0.082	0.082	26
19	−P(OH)−	phosphoric acid	0.493	0.032	6
20	>CHX	*b*	0.019	0.003	339
21	$-CHX_2$	*b*	0.050	0.007	66
22	$>C(CO_2H)C(CO_2H)<$	1,2-diacid	−0.362	0.033	3
23	$>C(X)CO_2H$	[b]	0.118	0.015	6
24	>C(X)OH	[b]	0.100	0.027	2
25	>CX−C(OH)<	[b]	0.051	0.003	62
26	nH:x	[c]	0.194	0.024	6
27	nH:c:x	[d]	0.042	0.010	35
28	H−bond 1	[e]	−0.089	0.004	47
29	H−bond 2	[e]	−0.161	0.009	41
30	H−bond 3	[e]	−0.251	0.017	13
31	H−bond 4	[e]	−0.418	0.021	8
32	H−bond 5	[e]	−0.450	0.016	10
33	H−bond 6	[e]	−0.155	0.024	11
34	H−bond 7	[e]	0.0	0.021	8
35	H−bond 8	[e]	−0.093	0.013	19
36	H−bond 9	[e]	−0.110	0.017	13
37	H−bond 10	[e]	−0.601	0.049	9
38	8−OH quinoline	peri interaction	−0.475	0.022	7
39	3−X phenol	meta interaction	0.119	0.008	38
40	4−X phenol	para interaction	0.176	0.011	34
41	3−X aniline	meta interaction	0.080	0.009	37
42	4−X aniline	para interaction	0.084	0.010	36
43	3−X benzoic acid	meta interaction	0.085	0.015	11
44	4−X benzoic acid	para interaction	0.055	0.020	8
45	2,6-dialkyl phenol		−0.162	0.016	13
46	2,6-dialkyl aniline		−0.181	0.040	2
47	2−CX phenol		0.195	0.039	2
48	$3-CO_2H$ phenol		−0.203	0.032	3
49	$4-CO_2H$ phenol		0.096	0.015	16
50	3−C=O phenol		0.185	0.039	2
51	4−C=O phenol		0.203	0.021	7
	intercept		0.003	0.001	

[a] sd = standard deviation. [b] See Table 2 for definition of X. [c] Pyrazole type. [d] Imidazole type. [e] See Figure 1 for definitions of hydrogen bonds.

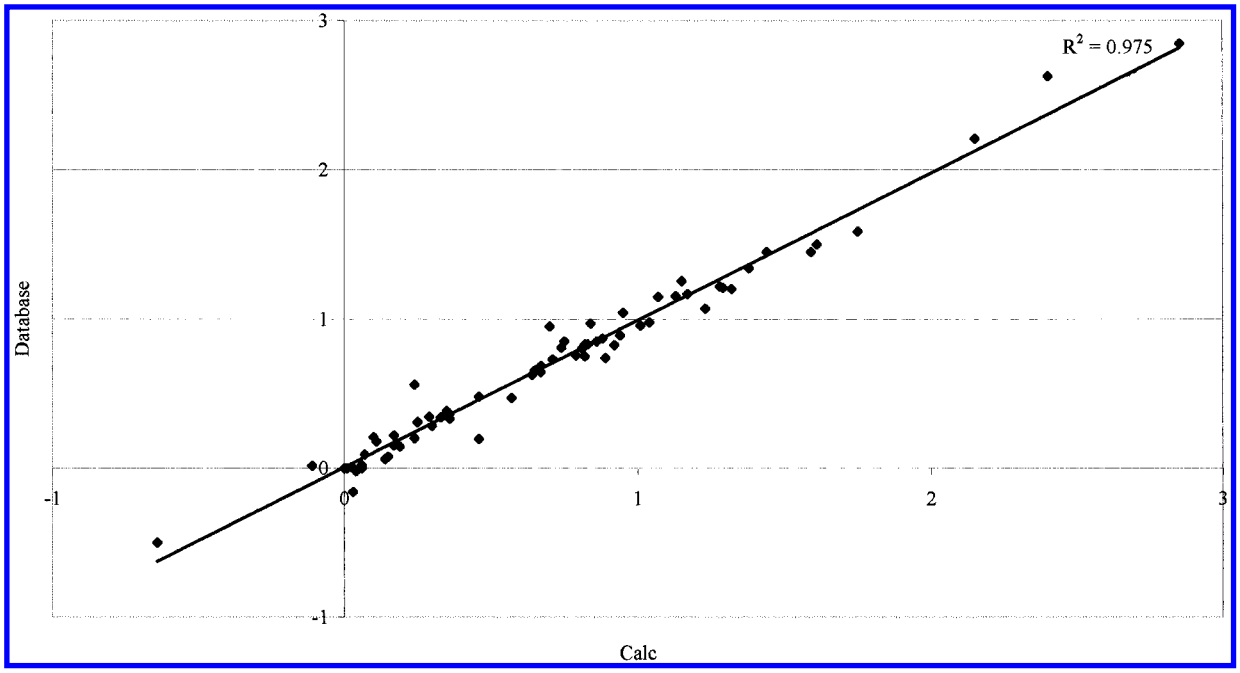


Figure 3. observed vs calculated results for R_2 test set.

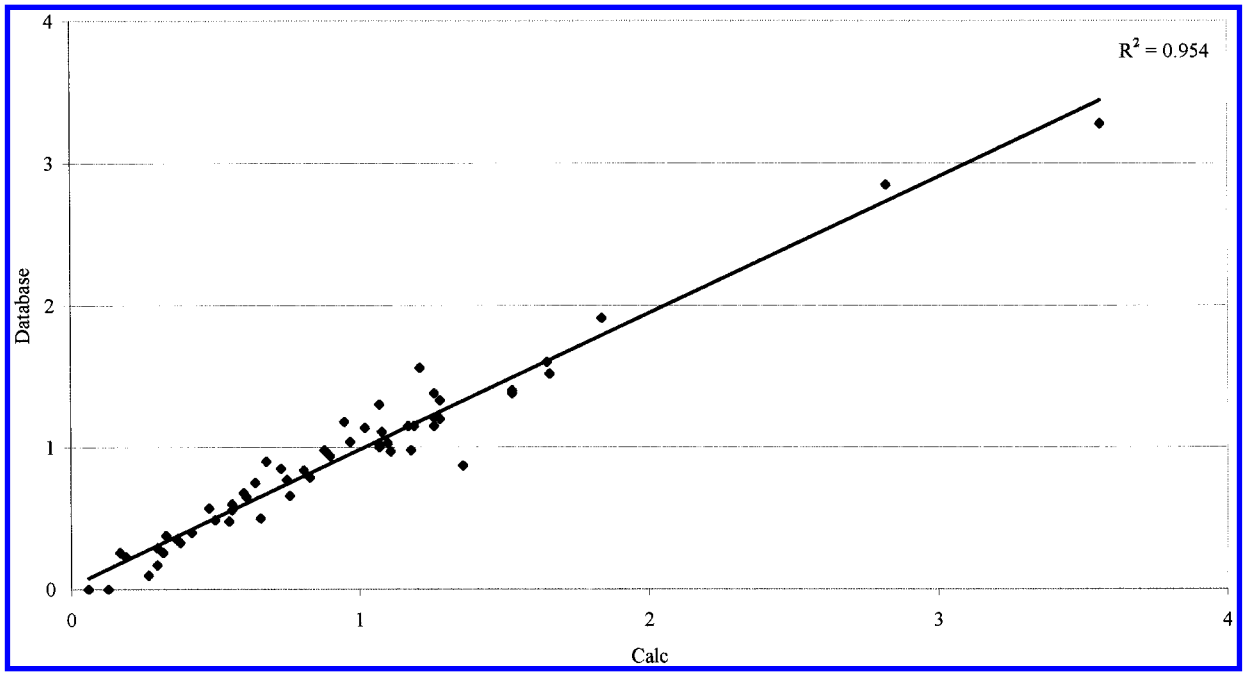


Figure 4. Observed vs calculated results for π_2^H test set.

gives rather poor statistics (R^2(adj) = 0.608, rms error = 0.142, F = 61.9, n = 1213), indicating that the Klopman model, even modified in this way, is unsuitable for the prediction of $\Sigma\alpha_2^H$.

An alternative set of fragments for the estimation of $\Sigma\alpha_2^H$ is presented in Table 5, along with their regression coefficients and standard deviations. This model resulted in much better statistics than either the original or modified Klopman models: R^2(adj) = 0.943, rms error = 0.057, F = 1408.1, and n = 3692 for all data and R^2(adj) = 0.831, rms error = 0.094, F = 130.7, and n = 1213 for all nonzero data. No inter-correlation with $r \geq 0.70$ is found for this set of fragments. As above, interesting comparisons can be drawn between coefficients, for example in the acidity of various −OH groups: phenols are almost twice as acidic as aliphatic alcohols, while carboxylic acids are stronger still. Much larger interaction corrections are found here than in the

Table 6. Results of Training and Test Regressions

	R^2 (adj)	rms error	F statistic	n
R_2 train	0.978	0.093	2528.5	3308
R_2 test	0.976	0.099		67
π_2^H train	0.922	0.164	553.5	2818
π_2^H test	0.954	0.122		57
$\Sigma\alpha_2^H$ train	0.945	0.058	1407.9	3619
$\Sigma\alpha_2^H$ test	0.943	0.055		73
$\Sigma\beta_2^H$ train	0.908	0.121	408.7	2491
$\Sigma\beta_2^H$ test	0.918	0.108		50
$\Sigma\beta_2^O$ train	0.903	0.129	387.7	2517
$\Sigma\beta_2^O$ test	0.929	0.131		51
log L^{16} train	0.990	0.240	3332.4	1908
log L^{16} test	0.994	0.180		38

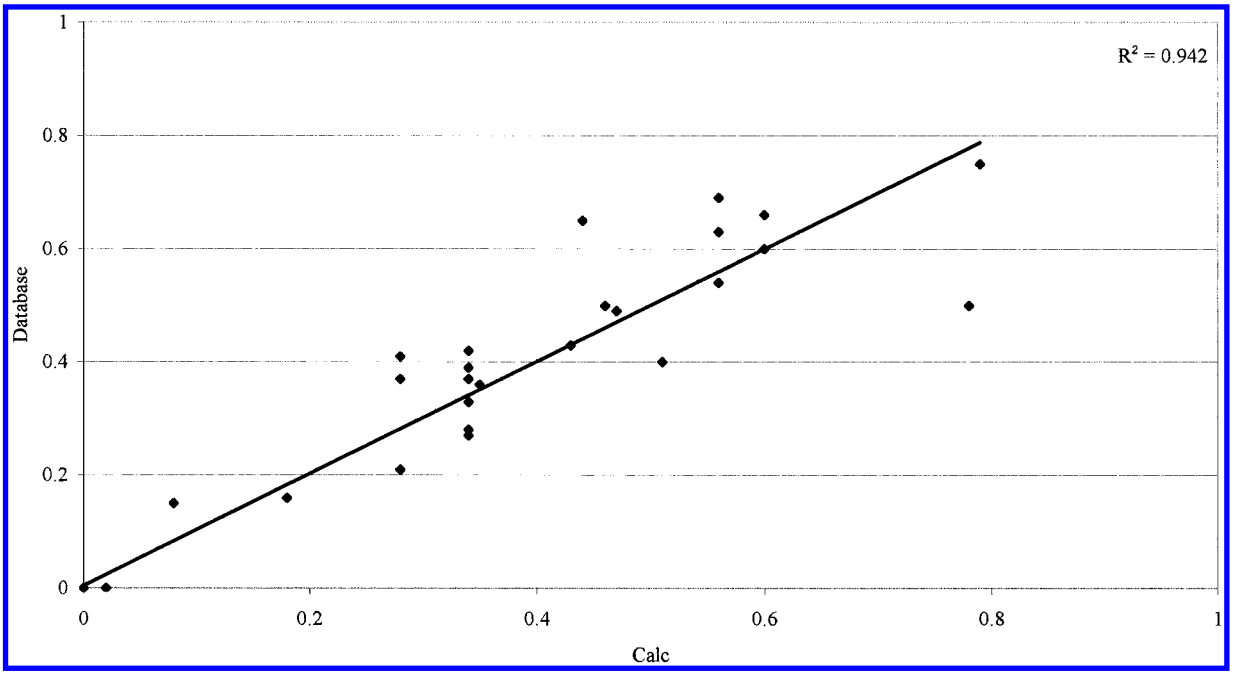


Figure 5. Observed vs calculated results for $\Sigma\alpha_2^H$ test set.

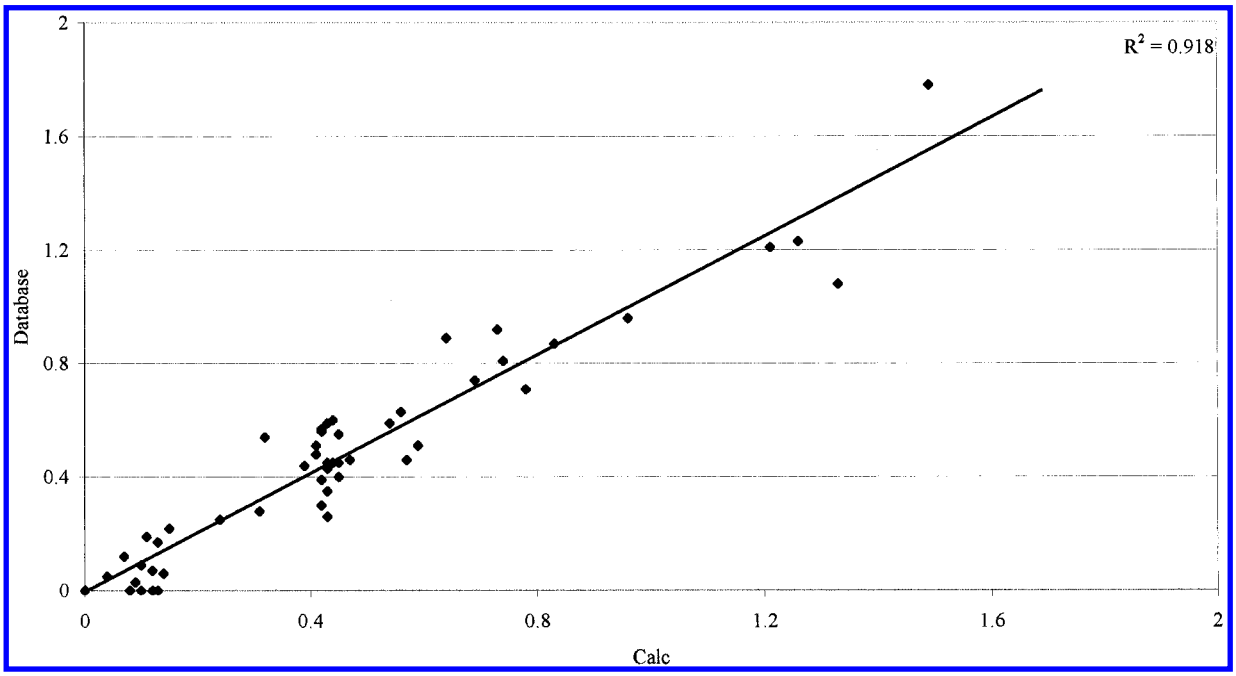


Figure 6. Observed vs calculated results for $\Sigma\beta_2^H$ test set.

previous regressions, both in aliphatic and aromatic molecules. This appears to be a result of the greater importance of intramolecular hydrogen bonding in determining $\Sigma\alpha_2^H$ than in the other descriptors discussed above.

To test the predictive power of the models developed, the training sets of data used in these regressions were split into training and test sets. These test sets were constructed by removing a random set of molecules, approximately 2% of the total, from the training set. This left sufficient data to apply the same regression techniques as above and to obtain qualitatively the same results, with test sets ranging from 40 to 70 data points. Table 6 and Figures 3−8 contain the results of this analysis. For each training set it is evident that the removal of 2% of the data points barely affects the regression statistics, with values almost identical to those reported in Table 3.

The coefficients from these training regressions were then used to predict values for the test sets. For each of the six test sets the correlation coefficient and standard deviation are very similar to that found from the training regression. The biggest deviations between training and test sets are found for π_2^H and $\Sigma\beta_2^O$; in both cases a small increase in the quality of the fit is observed in the test set. This suggests that the test set does not incorporate the largest possible spread of values for these two descriptors. However, we believe the spread of the current test sets, which may be seen in Figures 2 and 5, is sufficient for us to have confidence in the predictive power of our additive model.

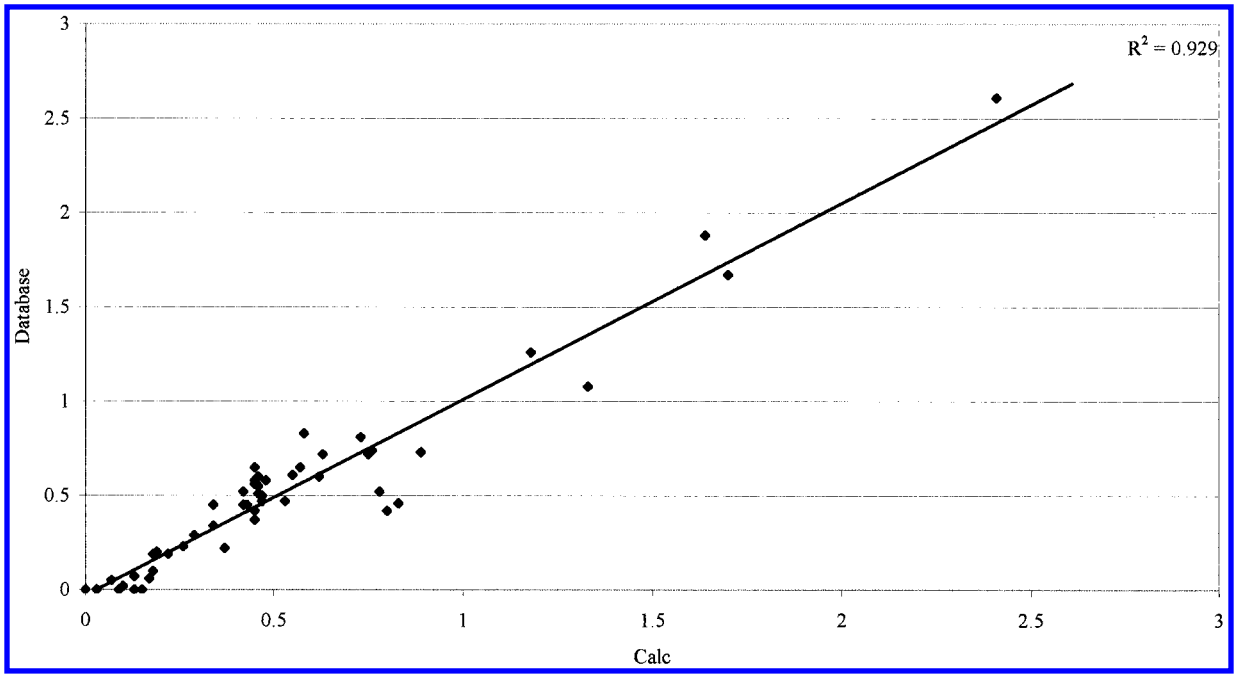


Figure 7. Observed vs calculated results for $\Sigma\beta_2^O$ test set.

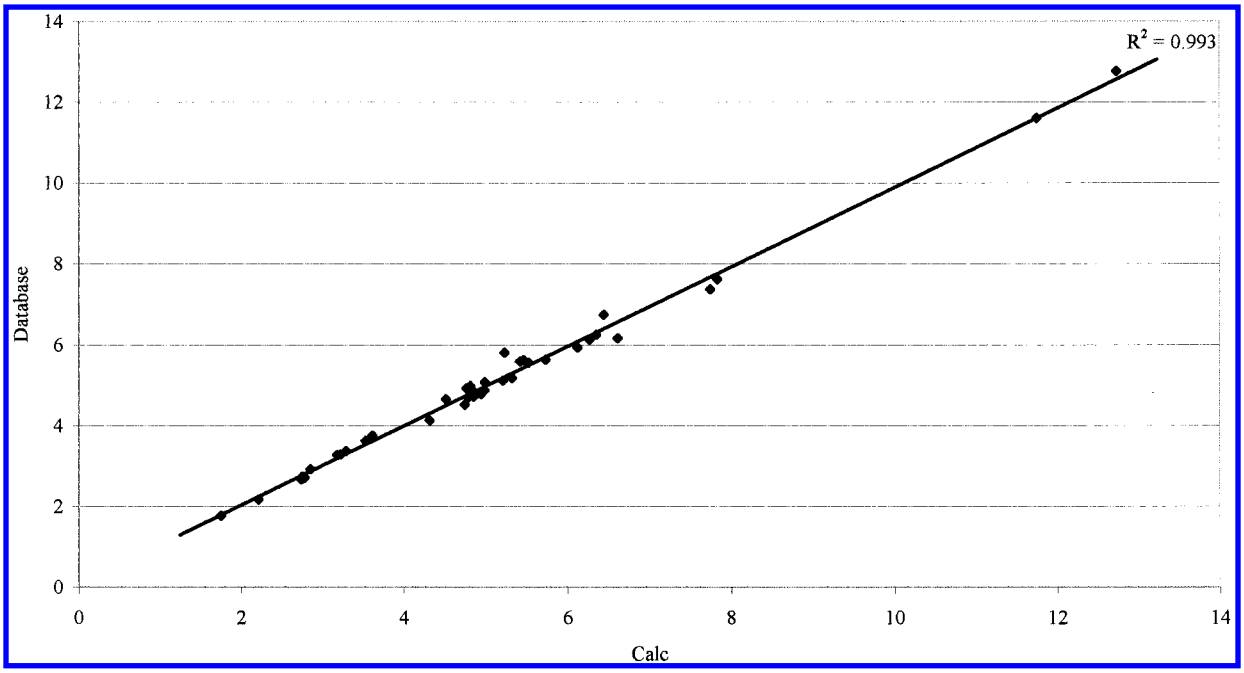


Figure 8. Observed vs calculated results for log L^{16} test set.

FURTHER DISCUSSION

The performance of the group contribution model in situations where simple additivity is likely to break down is important to the general applicability of the model. Two examples of such situations are ortho-substituted phenols and aromatic heterocycles. Table 7 contains predictions for 15 ortho-substituted phenols, along with their associated database values. In general the agreement is reasonable, though clearly some substituents such as fluorine, where the error in π_2^H is 0.2 and in $\Sigma\beta_2^H$ 0.18, are rather poorly predicted by our method. The largest rms error (0.15) is found for log L^{16}, purely a result of the much larger range of values covered here. The four other descriptors have much smaller rms errors, all of which lie between 0.02 and 0.07, approximately equal to those found for the entire training set (Table 3).

A similar analysis of aromatic heterocycles is reported in Table 8, where the overall agreement is rather better than in Table 7. Again log L^{16} has the largest rms error (0.234), while the other descriptors' rms errors range from 0.01 for $\Sigma\alpha_2^H$ to 0.09 for both scales of β. It should be pointed out, however, that individual molecules might have considerably larger errors than these rms values; e.g., $\Sigma\beta_2^H$ for imidazole is in error by 0.22, though $\Sigma\beta_2^O$ is wrong by just 0.02 for the same molecule. The unusually large rms error for log L^{16} is due mainly to the large error in the imidazole value, which may suggest that the experimentally derived figure is in error. The errors in other descriptors are better than, or

Table 7. Calculated and Experimental Descriptors for Ortho-Substituted Phenols

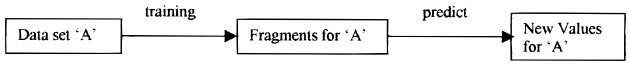
	R_2		π_2^H		$\sum\alpha_2^H$		$\sum\beta_2^H$ [a]		log L^{16}	
subst	expt	calc	expt	calc	expt	calc	expt	calc	expt	calc
Me	0.84	0.83	0.86	0.86	0.52	0.55	0.30	0.43	4.22	4.25
F	0.66	0.71	0.69	0.89	0.61	0.55	0.26	0.44	3.25	3.70
Cl	0.85	0.86	0.88	0.90	0.32	0.33	0.31	0.29	4.18	4.19
Br	1.04	1.03	0.90	0.98	0.35	0.33	0.31	0.28	4.53	4.62
I	1.36	1.37	1.00	1.01	0.40	0.33	0.35	0.29	4.96	5.11
OMe	0.84	0.78	0.91	0.77	0.22	0.29	0.52	0.6	4.45	4.52
COMe	0.95	0.97	1.14	0.99	0.20	0.13	0.47	0.38	4.94	5.06
CN	0.92	0.98	1.33	1.36	0.78	0.74	0.34	0.56	4.53	4.99
NH_2	1.11	1.03	1.10	1.04	0.60	0.64	0.66	0.59		4.79
NO_2	1.02	1.07	1.05	1.13	0.05	0.10	0.37	0.47	4.76	4.82
CO_2H	0.89	0.95	0.70	0.90	0.72	0.72	0.41	0.37		5.04
OH	0.97	0.93	1.10	0.90	0.88	0.84	0.47	0.68	4.45	4.52
CO_2Me	0.85	0.88	0.84	0.95	0.04	0.13	0.46	0.38	5.20	5.03
$CONH_2$	1.14	1.16	1.50	1.41	0.59	0.58	0.53	0.56		6.13
NHCOMe	1.05	1.10	1.56	1.54	1.09	0.87	0.79	0.79		6.80

[a] $\Sigma\beta_2^H$ and $\Sigma\beta_2^O$ are identical in all cases here.

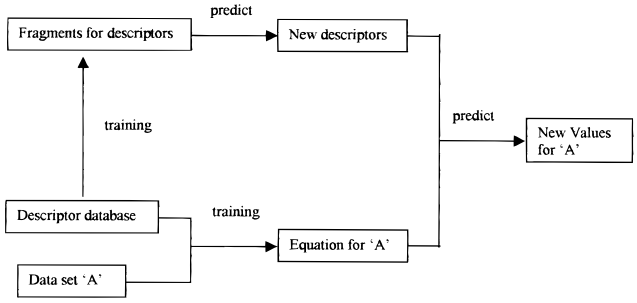
Table 8. Calculated and Experimental Descriptors for Aromatic Heterocycles

	R_2		π_2^H		$\Sigma\alpha_2^H$		$\Sigma\beta_2^H$		$\Sigma\beta_2^O$		log L^{16}	
molecule	expt	calc	expt	calc	expt	calc	expt	calc	expt	calc	expt	calc
pyrazine	0.63	0.61	0.95	0.83	0.00	0.00	0.61	0.58	0.61	0.57	2.92	3.04
pyrimidine	0.61	0.61	1.00	0.89	0.00	0.00	0.65	0.66	0.65	0.60	2.84	3.19
pyridazine	0.67	0.70	0.85	0.96	0.00	0.00	0.81	0.63	0.81	0.62	3.43	2.97
imidazole	0.71	0.47	0.85	0.79	0.42	0.42	0.78	0.56	0.50	0.48	4.02	3.34
pyrazole	0.62	0.56	1.0	0.90	0.54	0.57	0.45	0.61	0.34	0.57	3.15	3.18

Scheme 1



Scheme 2



comparable to, those found for all data. Thus our predictions for these two difficult classes of molecules, while far from perfect, are, we believe, acceptable for further application to other classes of molecules not included in the original training set.

A comparison of the method described here with the group contribution method employed by Klopman et al., Ghose et al.,[9−11] and others is in order here. The group contribution, or "process-fragment" method is conceptually very simple, as shown in Scheme 1; once fragment contributions for process "A" have been calculated, they can be used straight away to predict further values of *A*. Our "descriptor-fragment" method can be divided into two parts (see Scheme 2). First of all, a solvation equation, either (1) or (2), is constructed from the data set for process *A* and our descriptor database. Second, the descriptor database is used to obtain fragment contributions for descriptors as described above. Finally, the descriptor fragments are used to predict new descriptors that can be inserted into our equation for *A* so that further values of *A* can be predicted.

The process-fragment method has two inherent limitations. (i) The number of fragments necessary to describe any physicochemical or biochemical process is very large, and if these are to be used as independent variables in a multiple linear correlation equation, there must be a very large set of data *A* used as the dependent variable. We report 81 fragments, while others have used over 100,[10] requiring several hundred data points. There are few physicochemical processes for which this number of data points is available and no relevant biochemical processes at all. (ii) Predictions can be made only for solutes for which all the required fragments are available.

Our descriptor-fragment method overcomes both of these limitations. It has also an additional advantage that once new descriptors have been predicted, they can be used to predict solute properties in numerous systems for which solvation equations have already been constructed.

We can illustrate the above points by considering the physicochemical process of the solubility of gases and vapors in methanol, as log L^{MeOH} at 298 K.[14] Data were available for 93 solutes, leading to the equation

$$\log L^{MeOH} = -0.001 - 0.196R_2 + 1.117\pi_2^H + 3.671\sum\alpha_2^H + 1.501\sum\beta_2^H + 0.771 \log L^{16} \quad (5)$$

Although this is a reasonable number of data points for an equation with five independent variables, it is nowhere near the number required to obtain process fragments in Klopman's method. Furthermore, no log L^{MeOH} values were known for any solute with an aromatic −CO group or an aromatic −OH group, so no process fragment can be deduced for these entities. The descriptors used in eq 5 cover a wide range, as

Table 9. Range of Descriptor Values Used in Equation 5

descriptor	min	max	descriptor	min	max
R_2	−0.60	2.39	$\Sigma\beta_2^H$	0.00	0.79
π_2^H	−0.26	1.34	log L^{16}	−1.74	7.63
$\Sigma\alpha_2^H$	0.00	0.77			

Table 10. Prediction of log L^{MeOH} Values through Equation 5

solute	R_2	π_2^H	$\Sigma\alpha_2^H$	$\Sigma\beta_2^H$	log L^{16}	log L^{MeOH} (calc)
PhCHO	0.820	1.00	0.00	0.39	4.008	4.63
PhCOMe	0.818	1.01	0.00	0.48	4.501	5.15
$PhCO_2Me$	0.733	0.85	0.00	0.46	4.704	5.11
PhOH	0.805	0.89	0.60	0.30	3.766	6.41

Table 11. Calculated Descriptors and Prediction of log L^{MeOH}

solute	R_2	π_2^H	$\Sigma\alpha_2^H$	$\Sigma\beta_2^H$	log L^{16}	log L^{MeOH} (calc)
PhCHO	0.79	1.06	0.00	0.47	4.06	4.86
PhCOMe	0.80	1.03	0.00	0.47	4.54	5.20
$PhCO_2Me$	0.70	0.99	0.00	0.47	4.51	5.15
PhOH	0.83	0.88	0.55	0.44	3.77	6.41

shown in Table 9, so there is scope for prediction of further log L^{MeOH} values within this descriptor space, not only for solutes of the same nature as those used to set up the correlation but for other solutes as well. Descriptors are available for, e.g., benzaldehyde, acetophenone, methyl benzoate, and phenol and can be used to predict log L^{MeOH} values even without the aromatic −CO or aromatic −OH fragment in the original data set, see; Tables 10 and 11.

Thus our development of the descriptor-fragment method can be used, via Scheme 2, as a very general method for the rapid prediction of numerous physicochemical and biochemical properties just from structure. The application of this rapid predictive method to high-throughput screening will be discussed in a later publication. This will also act as a further test of the models' predictive ability, over and above that reported here, and will be presented in a later publication.

CONCLUSIONS

We have developed and tested additive models for six important molecular LFER descriptors, namely, R_2, π_2^H, $\Sigma\alpha_2^H$, $\Sigma\beta_2^H$, $\Sigma\beta_2^O$, and log L.[16] Five of these six, all bar $\Sigma\alpha_2^H$, are calculated from a single set of 81 atom and group fragments, while $\Sigma\alpha_2^H$ is calculated from a separate set of 51 fragments. In general, the linear fit obtained with these additive models is good, with R_2 and log L^{16} in particular giving excellent correlation. Splitting the data into training and test sets has also tested the predictive ability of such models, and is found to be almost as accurate as the full regressions.

The performance of the method in calculating descriptors for "difficult" structures, ones containing intramolecular interactions such as hydrogen bonds, has been analyzed. Variations in descriptors due to such interactions are generally found to be reproduced, though inevitably some small discrepancies are found. Arguments are put forward for the use of this method, rather than previous group contribution schemes, in cases where training data may be restricted in number or range.

ACKNOWLEDGMENT

J.A.P. is grateful to Glaxo Wellcome for a postdoctoral fellowship.

REFERENCES AND NOTES

(1) Abraham, M. H. Scales of Solute Hydrogen Bonding−Their Construction and Application to Physicochemical and Biochemical Processes. *Chem. Soc. Rev.*, **1993**, *22*, 73−83.
(2) Abraham, M. H.; McGowan, J. C. The Use of Characteristic Volumes to Measure Cavity Terms in Reversed Phase Liquid Chromatography. *Chromatographia* **1987**, *23*, 243−246.
(3) Abraham, M. H.; Andonian-Haftavan, J.; Whiting, G. S.; Leo, A. J.; Taft, R. W. Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777−1791.
(4) Abraham, M. H.; Platts, J. A.; Hersey, A.; Leo, A. J.; Taft, R. W. The correlation and estimation of gas-chloroform and water-chloroform partition coefficients by an LFER method. Submitted for publication in *J. Pharm. Sci.*
(5) Abraham, M. H.; Chadha, H. Application of a Solvation Equation to Drug Transport Properties. In *Lipophilicity in Drug Action and Toxicology*; Pliska, V., Testa, B., van de Waterbeemd, H., Eds.; VCH: Weinheim, Germany, 1996; and references cited therein.
(6) Abraham, M. H.; Whiting, G. S.; Doherty, R. M.; Shuely, W. J. Hydrogen Bonding. Part 13. A New Method for the Characterization of GLC Stationary Phases−The Laffort Data Set. *J. Chem. Soc., Perkin Trans. 2* **1990**, 1451−1460.
(7) Havelec, P.; Sevcik, J. G. K. Extended Addivity Model of Parameter log(L16). *J. Phys. Chem. Ref. Data* **1996**, *25*, 1483−1493.
(8) Svozil, D.; Sevcik, J. G. K. Neural Network Prediction of the Solvatochromic Polarity/Polarizability Parameter π_2^H. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 338−342.
(9) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer Automated log *P* Calculations Based on an Extended Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(10) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(11) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physiochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(12) Profile53: molecular profiling program originally written by J. Bradshaw; C implementation by P. Johnson; second generation, including external SMARTS definitions by D. Butina. Glaxo Wellcome: private communication.
(13) Weininger, D. SMILES, A chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(14) Abraham, M. H.; Whiting, G. S.; Carr, P. W.; Ouyang, H. Hydrogen bonding. Part 45. The solubility of gases and vapours in methanol at 298 K: An LFER study. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1385−1390.

CI980339T