A Comparison between the Two General Sets of Linear Free Energy Descriptors of Abraham and Klamt

Andreas M. Zissimos,*,† Michael H. Abraham,† Andreas Klamt,‡ Frank Eckert,‡ and John Wood§

Department of Chemistry, University College London, 20 Gordon Street, London, UK, WC1H OAJ,
COSMOlogic GmbH and Co. KG, Burscheider Str. 515, Leverkusen 51381, Germany,
and Statistical Sciences, GlaxoSmithKline, NFSP, Third Ave, Harlow, Essex, CM19 5AN, UK

Received April 23, 2002

Two sets of molecular descriptors, the five experimental Abraham, and the five COSMOments of Klamt's COSMO-RS, have been compared for a data set of 470 compounds. Both sets are considered as almost complete sets of LFER. The two sets of descriptors are shown to exhibit a large overlap as far as their chemical content. The chemical information however is distributed differently in each set with the Abraham set incorporating extra information in the excess molar refraction descriptor E. Regression equations have been constructed to predict the experimental Abraham descriptors from theoretically calculated COSMOments. The chemical interpretation of these equations is however difficult because of the lack of clustering which characterizes the distribution of chemical information through the two sets of descriptors. The predictability of the regression equations is tested successfully using a reasonably large set of data, and the method is compared to recent attempts to calculate the Abraham descriptors from various theoretical bases.

INTRODUCTION

Quantitative structure−property relationships (QSPR) and linear free energy relationships (LFER) have proved to be useful tools for the analysis of solvation phenomena. These relationships involve the use of a number of descriptors that describe properties of a solute molecule and, as an extension, the behavior of the solute molecule in solution. One general method for the construction of QSPRs starts with the generation of a very large number of molecular descriptors for each compound. Some method of descriptor reduction is then used to reduce the number of descriptors typically to around 5−10. The reduced set may then be used linearly or nonlinearly in a QSPR. For example, in the ADEPT routine of Jurs et al.[1−4] 210 molecular descriptors were calculated and then reduced to nine descriptors that were used to correlate aqueous solubility.[3] In the CODESSA program of Katritzky et al.[5−7] no less than 800 molecular descriptors were calculated and reduced to five or six descriptors for the correlation of a number of physicochemical properties.[6] Not surprisingly, good correlations of training sets are invariably obtained. However in the latest solubility work of Katritzky et al.[8] a consistent set of descriptors is used in all the systems studied. Validation of predictive capability through external prediction sets (test sets) is not always carried out, but Mitchell and Jurs[3] used a 32-compound test set in their study of aqueous solubility. Of course, the "best" set of reduced descriptors for the correlation of any given property is very unlikely to be the same as the best set for the correlation of any other property. This leads to one disadvantage of the "plethora of descriptors" method, namely that it is not possible to carry out a term-by-term comparison of QSPRs for two systems, even if the systems are chemically closely related.

One method that avoids the latter disadvantage is to construct QSPRs, and LFERs, through the use of a small number of predetermined molecular descriptors. The same small set of descriptors is used for the correlation of various properties, so that an exact comparison can be made between correlation equations. There are a number of procedures that have been used to obtain small sets of descriptors. The original work of Kamlet and Taft and co-workers[9,10] has shown that it is indeed possible to define a rather small number of descriptors that could be combined in a linear way for the correlation of solute properties. After considerable preliminary work,[11,12] Abraham and co-workers succeeded in constructing a new and more rigorous set of five solute descriptors,[13−16] specified as follows, with the original nomenclature in parentheses. E (R_2) is an excess molar refraction that is obtained from refractive index for solutes that are liquid at 20 °C. For solids, the refractive index of the hypothetical liquid at 20 °C can be calculated, or E can be obtained by the summation of fragments or substructures. S (π^H) is the dipolarity/polarizability that can be obtained from gas liquid chromatographic measurements on polar stationary phases or more generally from water/solvent partitions. A ($\Sigma\alpha^H_2$) and B ($\Sigma\beta^H_2$) are the overall or effective hydrogen bond acidity and basicity that are most easily obtained from water−solvent partitions, and V (V_x) is the McGowan characteristic volume[17] that can easily be calculated from bond and atom contributions.[13] The range of solutes for which descriptors are currently available is now quite large and encompasses compounds as far apart as helium, hydrogen, nitrogen, etc. on one hand and drugs, environmental pollutants, and pesticides on the other.

[*] Corresponding author fax: +44-020-7679-7463; e-mail: a.zissimos@ucl.ac.uk.
[†] University College London.
[‡] COSMOlogic GmbH and Co. KG.
[§] GlaxoSmithKline.

10.1021/ci025530o CCC: $22.00
Published on Web 10/01/2002

Table 1. Comparison of Correlation Sets of Compounds Used by Various Workers

	Abraham DB		Dearden[22,23]		Platts[24−26]		Sevcic[18,19]		this work	
Abraham descriptor	N	min/max	N	min/max	N	min/max	N	min/max	N	min/max
E	4167	−1.38/4.62							470	−0.55/3.26
S	3631	−1.34/5.60			98	−0.25/1.58	333	0−1[a]	470	−0.25/2.25
A	4375	0.00/4.33	55	0.14/0.79	54	0.08/0.95			470	0.00/1.62
B	3312	0.00/4.52			50	0.04/0.68			470	0.00/1.5
V	4260	0.07/8.56							470	0.11/2.84

[a] A cutoff so that S < 1 was used.

These solute descriptors can be combined in an LFER (eq 1).

$$\mathrm{LogSP} = c + eE + sS + aA + bB + vV \qquad (1)$$

The dependent variable, log SP, is a solute property in a given system. For example, it might be log P for a set of solutes in a given water−solvent partition system. The coefficients in the equations are found by the method of multiple linear regression. Since the development of the five descriptors, a descriptor database (Abraham's database) has been constructed from experimental data; the maximum and minimum range of these descriptors in the database are shown in Table 1. The descriptors represent the solute influence on various solute−solvent phase interactions. Hence the regression coefficients c, e, s, a, b, and v correspond to the complimentary effect of the phases on these interactions. The coefficients can be regarded as system constants which characterize and contain chemical information of the phase in question. The system constants can be interpreted as follows. The e-coefficient shows the tendency of the phase to interact with solutes through π and n-electron pairs. Usually the e-coefficient is positive, but for a phase which contains fluorine atoms, it can be negative. The s-coefficient represents 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 basic phases), 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 v-coefficient is a measure of the hydrophobicity of the phase. The coefficients in the solvation parameter equation are therefore not just fitting constants but must obey general chemical principles. An example to illustrate the chemical information contained in the system constants is partition of solutes between two phases. In this case, the system constants will reflect differences in properties of the two phases and hence can take positive or negative values. The important water/octanol system is characterized by the following equation.

$$\log P_{oct} = 0.088 + 0.562E - 1.054S + 0.034A - 3.460B + 3.814V \quad (2)$$

$$N = 613 \ R = 0.9974 \ SD = 0.116 \ F = 23161.6$$

Thus octanol (actually, wet octanol) is revealed to be more able to interact with π- and n-electron pairs than is water (positive e-coefficient) but is less dipolar/polarizable than water, hence the negative s-coefficient. Octanol has almost the same hydrogen-bond basicity as water (almost zero a-coefficient) but is a weak hydrogen-bond acid (negative b-coefficient). The large v-coefficient means that octanol is able to interact with solutes by dispersion forces and/or that the energy required to create a given sized cavity in octanol is relatively low. Octanol would be regarded as a hydrophobic phase.

Any application of the general solvation equation (eq 1) depends on the availability of the solute descriptors, and the need to calculate descriptors for new compounds will always be of primary importance. As explained earlier, the descriptor V can be calculated quite simply for any structure from the molecular formula and the number of rings in the molecule, using the algorithm of Abraham for the number of bonds in the molecule.[13] The E descriptor can also be calculated from the refractive index at 20 °C, using either the observed refractive index for a liquid or a calculated refractive index for the solid. This descriptor can also be estimated by the addition of fragment values (substructures).

The remaining three descriptors S, A, and B have to be obtained from experimental measurements of physicochemical properties. Traditionally, the solute descriptors are derived from experimental measurements such as water−solvent and gas-solvent partitions and chromatographic methods including GC, GLC, and HPLC. Although this approach clearly provides the best descriptors for most kinds of molecules, it certainly has its limitations. First the fact that one must physically obtain a sample of the compound, imposes many difficulties especially when equations are required for screening purposes. Second, certain techniques used for measuring these parameters might not be applicable in certain cases. For example, use of UV spectroscopy as an analytical tool requires a chromophore in the molecule. Third some of the experimental methods are laborious and time-consuming and this limits their applicability especially in high throughput setups. Because of these difficulties in obtaining the Abraham physicochemical descriptors, attempts have been made to escape the reliance on experimental data for the determination of new S, A, and B values. Such attempts include the work of Sevcik[18,19] and co-workers who have reported an additive scheme for the estimation of the descriptor logL[16] and the neural network approach to estimate the S parameter. The former approach adds contributions to logL[16] from a given set of fragments, the contributions being derived from multivariate regression analysis (MLRA). The latter approach 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.

Platts et al.[20] have introduced an additive model (UNIX method) for the estimation of Abraham's molecular descriptors E, S, A, B, V, and logL.[16] This model was developed from a set of 81 atom and functional group fragments and intramolecular interactions for which an evaluation of their contribution to each descriptor was carried out through a process of multiple linear regression. They proceeded to apply this group contribution model on sets of molecules for which partitioning data was predicted using Abraham's solvation equations.[21] The method gives good results for predicting the molecular descriptors in question and partitioning data for a number of compounds, but as with all group contribution methods it retains the basic disadvantage of being unable to resolve molecular details such as isomeric tautomeric and conformational effects and it is hard to apply on large complicated compounds with diverse functional groups. A commercial package has been developed recently following Platts work with the addition of new fragments.

Dearden and Ghafourian[22,23] have obtained hydrogen bonding parameters comparable to the Abraham A and B values using alternative ways of calculation. Based on the postulation that hydrogen bonding is mainly electrostatic in nature, they used electrostatic interactions to model the hydrogen bonding ability of molecules. Atomic charges and LUMO energies (Lowest Unoccupied Molecular Orbitals) calculated using various semiempirical methods such as AM1, PM3 and MNDO, MNDO electrostatic-potential-derived atomic charges, were correlated with Abraham's hydrogen bond acidity and basicity, and relationships were found to calculate hydrogen-bonding descriptors for QSAR correlations.

Platts et al.[24−26] have recently carried out calculations on the structure and properties of rather small sets of compounds using ab initio and DFT calculations. The calculated properties for these molecules were assessed for their ability to correlate and predict experimentally derived values of S, A, and B from Abraham's database. Some comparisons with Platts work will be attempted in this paper.

The work of Politzer and Murray[27] make use of a methodology which involves a General Interaction Properties Function (GIPF) which is applicable to properties involving both covalent and noncovalent interactions. GIPF is a collective term that has been also applied to correlate properties of molecules with a series of computed parameters which involve surface area, average local ionization energy, indicators of long range attraction for nucleophiles and electrophiles, local polarity, and the variability of the surface electrostatic potential. By means of the GIPF approach quantities computed for an isolated molecule can be used to correlate and predict solution and liquid phase properties. Although the method of Politzer and Murray is a well-known one, no attempt has been made to date to correlate this set of descriptors with those of Abraham so we only mention them here for literature completeness.

Finally the procedure of Famini and Wilson et al.[28,29] in general design is quite close to that of Abraham, except that the descriptors are calculated using semiempirical quantum mechanical calculations with MNDO or AM1. Six calculated descriptors are used linearly to correlate properties, and the linear correlation equations are termed theoretical linear solvation energy relationships, TLSERs. It was shown that the TLSER descriptors could be used in exactly the same way as the experimental descriptors of Abraham, so that the five experimental descriptors and the six calculated descriptors must encode nearly the same information. We shall not pursue this comparison, however, because a more powerful calculation procedure has recently been developed by Klamt et al.[30−35] Computational methods for calculating the five Abraham descriptors so far have been applied on limited ranges of descriptors and rather small numbers of compounds.

The aim of the present work is 3-fold. First, to compare the information content of the five Abraham descriptors and the five COSMOments. Second, to investigate whether any or all of the Abraham descriptors can be obtained from those of Klamt's and vice versa, and third to compare methods for the calculation of the Abraham descriptors, where S, A, and B are considered. These are the *ab initio* methods of Platts et al.,[24−26] the work of Sevcic,[19] Dearden,[22,23] and the method developed in this work under the second aim.

COSMO AND COSMO-RS

COSMO-RS is a model combining quantum theory, dielectric continuum models, surface interactions, and statistical thermodynamics. Since a full derivation of the theory of COSMO-RS is beyond the scope of this article, a short summary of the essentials is given here. More details can be found in refs 30−35.

COSMO-RS considers a liquid system as an ensemble of molecules of different kinds, including solvent and solute. For each kind of molecule X a density functional (DFT) calculation with the dielectric continuum solvation model COSMO[30] is performed in order to get the total energy $E^X{}_{COSMO}$ and the polarization (or screening) charge density (SCD) (sigma) σ on its molecular surface. σ is very good local descriptor for molecular surface polarity. For the purpose of an efficient statistical thermodynamics calculation the liquid ensemble of molecules now is considered as an ensemble of pairwise interacting molecular surfaces. The most important parts of the specific interaction between molecular surfaces, i.e., electrostatics (es) and hydrogen bonding (hb), are expressed by the SCDs σ and σ' of the contacting surface pieces:

$$E_{es}(\sigma,\sigma') = \frac{\alpha'}{2}(\sigma + \sigma')^2 \qquad (3)$$

and

$$E_{hb}(\sigma,\sigma') = c_{hb} \min\{0,\sigma\sigma' + {\sigma_{hb}}^2\} \qquad (4)$$

The three parameters α', c_{hb}, and σ_{hb} have been adjusted to a large number of thermodynamic data. Since all relevant interactions depend on σ, the distribution functions (histograms) $p^X(\sigma)$ are required for the statistical thermodynamics. These "σ-profiles" are easily derived from the COSMO output. Note, that the σ-profiles provide a vivid picture of the molecular polarity (see Figure 1 and a discussion given in refs 30 and 32). Furthermore we need the σ-profile $p_S(\sigma)$ of the ensemble S, which is calculated as a sum of the molecular σ-profiles weighted by mol-fractions.

Now the chemical potentials of the compounds in the solvent are calculated by a novel, exact, and very efficient statistical thermodynamics procedure. The first step is the iterative solution of the equation

$$\mu_S(\sigma) = -\frac{RT}{a_{eff}}\ln\left\{\int d\sigma' p_S(\sigma')\exp\left(\frac{a_{eff}}{RT}(\mu_S(\sigma') - E(\sigma,\sigma'))\right)\right\} \qquad (5)$$

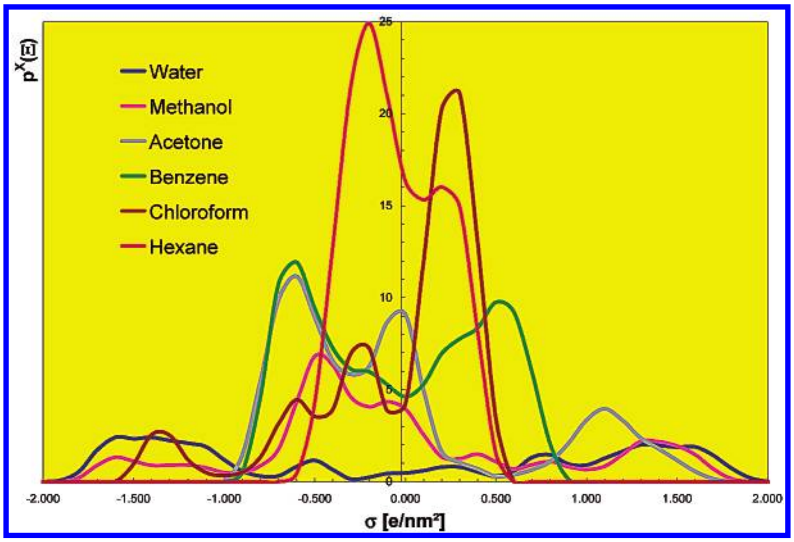

Figure 1. *σ*-profiles of different solvents.

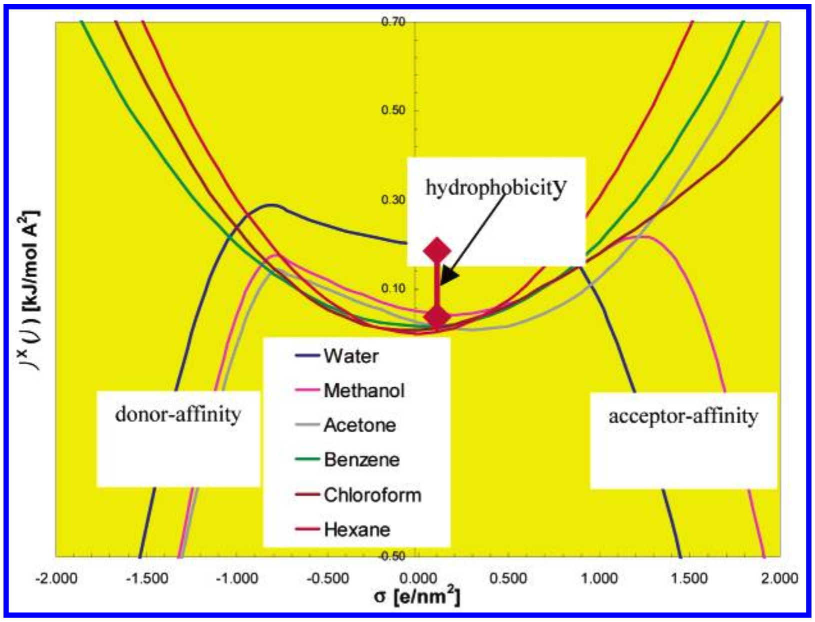

Figure 2. *σ*-potentials of solvents.

a_{eff} denotes the size of an effectively independent piece of molecular surface area. This implicit equation can be solved by iteration within milliseconds on a PC. It yields the function $\mu_S(\sigma)$, called σ-potential, which tells how much the solvent S likes surface of polarity σ. This is a characteristic function for each solvent. Examples are given in Figure 2.

These σ-potentials describe the solvent behavior regarding electrostatics, HB-affinity, and hydrophobicity. In a second step, the σ-potential is integrated over the surface of each compound X, yielding the chemical potential of X in S:

$$\mu_S{}^X = \int p^X(\sigma)\mu_S(\sigma)d\sigma + \mu^X_{combS} \qquad (6)$$

In this equation the surface integral is evaluated as an σ-integral, making use of the σ-profile of the solute X. The combinatorial contribution $\mu^X{}_{comb,S}$ to μ takes into account size and shape effects of solute and solvent. Usually it is small compared to the first term in eq 6 which results from the surface interactions. It is sufficient to consider it as a solvent specific constant, here.

As a result of this series of relatively simple steps, starting from a quantum chemical calculation for each compound an expression is found for the chemical potential of an almost arbitrary chemical compound X in an almost arbitrary solvent S, which may be a pure compound or a mixture. This allows the calculation of any partition coefficient as well as solubility. Based on density functional COSMO calculations, the few parameters required in COSMO-RS, have been fitted to a large set of experimental data,[33] covering 215 diverse chemical compounds and the properties ΔG_{hydr}, $logP_{vapor}$, and the aqueous partition coefficients with octanol, hexane, benzene, and ether. Note, that the properties ΔG_{hydr} and $logP_{vapor}$ involve the gas-phase, which requires a small addendum to the steps given above that is not of interest here. However, since $logS_{aq}$ is the difference of $\Delta G_{hydr}/RT$ and lnP_{vapor}, aqueous solubility was implicitly taken into account in the parametrization of COSMO-RS. The initial COSMO-RS parametrization yielded a rms-error of 0.3 log-units for the diverse partition and solubility properties of small and medium sized molecules. In recent parametrizations the error has been reduced to about 0.23 log-units.

EXTENSION OF COSMO-RS TO COMPLEX SOLUTIONS

COSMO-RS is a reliable method for the a priori prediction of thermophysical data and phase equilibria of pure liquids and liquid mixtures of well defined composition. But there exist several thermodynamic equilibria of industrial importance, which involve one or more phases, which are either chemically less defined, or which are disordered, but not really liquid, or both. Examples of such systems are physiological phases like blood, brain, or special tissue, structurally sophisticated polymers, and solid adsorbents, like activated carbon. In such phases no surface composition function $p_S(\sigma)$ is available. Hence the σ-potential $\mu_S(\sigma)$ of the phase S and the chemical potentials $\mu_S X$ of solutes X in these phases cannot be directly calculated by COSMO-RS. But an indirect treatment of such phases by COSMO-RS is enabled by the following extension.

Consideration of a large number of different solvents led to the finding (see Figure 2) that σ-potentials can be described very well by a Taylor-like expansion of the form

$$\mu_S(\sigma) \cong \sum_{i=-2}^{m} c_S{}^i f_i(\sigma) \tag{7}$$

with

$$f_i(\sigma) = \sigma^i \ for\ i \geq 0 \tag{8}$$

and

$$f_{-2/-1}(\sigma) = f_{acc/don}(\sigma) \cong \begin{cases} 0 \ if \pm\sigma < \sigma_{hb} \\ \mp\sigma + \sigma_{hb} \ if \pm\sigma > \sigma_{hb} \end{cases} \tag{9}$$

The highest order of the polynomial contributions (eq 8) required for a sufficient description of σ-potentials typically is $m = 3$. The hydrogen bonding contributions expressed by eq 9 are necessary to describe the acceptor and donor behavior of the solvent. As can be seen in Figure 2, this behavior corresponds to a linear descent in the σ-potentials starting from some threshold $\sigma_{\rm hb}$. The functions $f_{acc}(\sigma)$ and $f_{don}(\sigma)$ are well capable of describing just these features of the σ-potentials. Using this Taylor expansion, we may characterize each solvent (at fixed temperature, usually room temperature) by the set of σ-coefficients c^i_S. Obviously any difference between the σ-potentials of two solvents is of the same kind of expansion, with coefficients $c^i_{S,S'}$ being just the difference of the coefficients of the two solvents. Partition coefficients are connected with the pseudochemical potentials by the equation

$$kT \ln K^X_{S,S'} = [\mu^X_{S'} - \mu^X_S] \tag{10}$$

Using eq 5 for $\mu_s(\sigma)$, we thus find that any partition coefficient between two solvents S and S' should be expressible in the form

$$\ln K^X_{S,S'} = \frac{1}{kT}[c_{S,S'} + \int p^X(\sigma)(\mu_{S'}(\sigma) - \mu_S(\sigma))d\sigma] \cong \tilde{c}_{S,S'} + \int p^X(\sigma)\sum_{i=-2}^{m} \tilde{c}^i_{S,S'} f_i(\sigma)d\sigma = \tilde{c}_{S,S'} + \sum_{i=-2}^{m} \tilde{c}^i_{S,S'} M^X_i \tag{11}$$

where the combinatorial contributions have been subsumed in $\tilde{c}_{S,S'}$ and the σ-moments M_i^X of the solute X are defined by

$$M_i^X = \int p^X(\sigma) f_i(\sigma) d\sigma \tag{12}$$

Equation 10 implies that any logarithmic partition coefficient can be represented as a linear combination of σ-moments. As a consequence, the set of σ-moments M_i^X, $i = 0,2,3$, complemented by the hydrogen bond moments $M_{acc}{}^X$ ($=M_{-2}{}^X$) and $M_{don}{}^X$ ($=M_{-1}{}^X$), should be a very good and almost complete set of molecular descriptors for a linear regression analysis of any partition problem, i.e., for linear free energy regression (LFER). Note that the first moment $M_1{}^X$ usually is of no importance, because it is just the negative of the total charge of the molecule. Hence, for neutral compounds $M_1{}^X$ trivially vanishes. By definition of the σ-profiles, the zeroth moment $M_0{}^X$ (CSA) is identical with the molecular surface. The second moment (sig2) is an excellent measure of the overall electrostatic polarity of the solute, and the third moment (sig3) is a measure of the asymmetry of the sigma profile. The hydrogen bond moments (Hbacc3 and Hbdon3) are quantitative measures of the acceptor and donor capacities of the compound X, respectively.

METHODOLOGY

We have assembled 470 compounds for which we have experimental descriptors in our existing database. For these compounds we calculated the five COSMOments which represent polarity/polarizability (sig2, sig3) and hydrogen bond acidity (Hbdon3) and basicity (Hbacc3) as well as a surface area descriptor (CSA). A detailed statistical analysis was carried out in order to determine how much of the chemical information is enclosed in each set of descriptors and how big is the overlapping information space. A training set of 35 compounds was selected in such a way as to cover a descriptor space as big as possible. Our descriptors were correlated with the five COSMOments using Excel and JMP[36] statistical tools. The resulting equations have been applied to the remaining 435 compounds comprising our test set and the predictability of the model was assessed. Equations were also obtained from the total number of compounds Tables 5 and 6. Finally comparison of our method and the method of other workers has been carried out.

RESULTS AND DISCUSSION

Stability of the Chemical Dimensions Underlying Each Set of Descriptors. To investigate the stability of the chemical dimensions underlying each set of descriptors, a Principal Component Analysis (PCA) was carried out separately for each set. In a PCA, the percentage of variance accounted for gives an indication of the stability of a component, while examination of the loadings can indicate its interpretation. The summaries of the principal components analyses (PCA) carried out on the Abraham and Klamt descriptor sets for the 470 compounds are given in Tables 3 and 4. Because the descriptors have quite different units, they were standardized prior to this analysis by subtracting off their means and then dividing by their standard deviations, so each descriptor then had zero mean and unit variance.

Table 2. Training Set of 35 Compounds Used

name	sig2	Sig3	Hb don3	HB acc3	area [CSA]	E	S	A	B	Vx
propionic acid	69.765	−2.102	3.719	1.937	110.823	0.233	0.65	0.6	0.45	0.6057
ethanol	53.537	23.480	1.993	4.473	85.890	0.246	0.42	0.37	0.48	0.4491
methanol	53.585	20.284	2.141	4.227	66.362	0.278	0.44	0.43	0.47	0.3082
cyclohexane	5.713	0.400	0.000	0.000	126.003	0.305	0.1	0	0	0.8454
methane	4.762	−0.231	0.000	0.000	55.697	0	0	0	0	0.2495
cyclohexene	16.188	6.310	0.000	0.000	124.438	0.395	0.2	0	0.1	0.8024
piperazine	68.257	76.204	0.114	8.165	126.174	0.57	0.83	0.11	1.14	0.7632
o-methylphenol	62.278	−21.047	3.470	0.550	148.665	0.84	0.86	0.52	0.3	0.916
phenol	65.252	−25.693	3.929	0.679	127.586	0.805	0.89	0.6	0.3	0.7751
tolune	28.583	1.514	0.000	0.000	135.789	0.601	0.52	0	0.14	0.8573
imidazole	85.311	13.854	3.627	5.420	101.308	0.71	0.85	0.42	0.78	0.5363
p-bromophenol	67.615	−35.937	4.024	0.477	154.735	1.08	1.17	0.67	0.2	0.9501
tetrafluoromethane	5.908	−3.175	0.000	0.000	86.812	−0.55	−0.25	0	0	0.3203
m-cyanophenol	86.733	−24.643	4.493	1.127	155.499	0.93	1.55	0.84	0.25	0.9298
p-nitrophenol	93.699	−30.965	5.070	0.683	158.111	1.07	1.72	0.82	0.26	0.9493
1,4-diethylbenzene	29.551	4.839	0.000	0.000	186.217	0.645	0.5	0	0.18	1.28
nonane	10.579	0.960	0.000	0.000	204.845	0	0	0	0	1.3767
2-hexanone	47.927	36.706	0.000	2.896	153.393	0.136	0.68	0	0.51	0.9697
formic acid	72.256	−24.292	4.987	1.228	71.326	0.3	0.79	0.72	0.34	0.3239
heptan-3-ol	46.812	20.749	1.207	3.384	174.010	0.178	0.36	0.33	0.56	1.1536
m-dihydroxybenzene	100.956	−50.624	7.821	1.312	136.232	0.98	1.11	1.09	0.52	0.8338
1-naphthol	71.152	−34.026	4.158	0.272	170.153	1.52	1.05	0.6	0.37	1.1441
dibenzofuran	46.690	−5.535	0.000	0.000	191.203	1.407	1.02	0	0.17	1.2743
1,3-dihydroxybenzene	102.752	−55.119	8.183	1.309	136.464	0.98	1.11	1.09	0.52	0.8338
3,5-dichlorophenol	65.322	−52.151	4.979	0.211	165.486	1.02	1	0.91	0	1.0199
orthene	117.083	38.598	2.946	5.490	191.737	0.505	2.02	0.36	1.19	1.2732
gallic acid	153.290	−93.554	14.062	2.095	173.024	1.29	1.45	1.62	0.85	1.1078
4-hydroxybenzonitrile	87.572	−28.045	4.797	1.111	151.633	0.94	1.63	0.80	0.29	0.9298
phenazone	103.651	71.013	0.000	8.539	211.844	1.32	1.5	0	1.48	1.5502
succinic acid	120.847	−8.431	5.669	3.695	133.936	0.37	1.36	0.85	0.7	0.821
2-butene-cis	16.853	4.618	0.000	0.000	111.610	0.142	0.08	0	0.05	0.6292
ethylenediamine	75.950	81.492	0.074	8.637	105.772	0.462	0.17	0.04	1.29	0.59
dimethylamine	38.598	42.945	0.061	4.471	93.207	0.189	0.3	0.08	0.66	0.4902
chloropentafluoroethane	5.116	−1.799	0.000	0.000	128.851	−0.36	−0.12	0	0	0.6013
1-propanol-2,2-dimethyl	47.889	18.340	1.466	3.203	130.435	0.22	0.36	0.37	0.53	0.8718

Table 3. Summary of Principal Component Analysis: Abraham Descriptors for 470 Compounds

statistics used for autoscaling	E	S	A	B	V
mean	0.5356	0.6825	0.1526	0.3573	0.9463
std dev	0.5198	0.4537	0.2734	0.2818	0.3678
summary statistics	Comp1	Comp2	Comp3	Comp4	Comp5
standard deviation	1.53	1.06	0.91	0.74	0.40
percentage of variance	47	22	17	11	3
loadings (eigen-vectors)	Comp1	Comp2	Comp3	Comp4	Comp5
E	0.5169	0.3026	0.4467	0.3053	−0.5903
S	0.6018	−0.0776	0.0899	0.3214	0.7214
A	0.3251	−0.6431	0.3233	−0.6030	−0.1121
B	0.3851	−0.3013	−0.7904	0.1664	−0.3293
V	0.3415	0.6309	−0.2511	−0.6421	0.1003

Table 4. Summary of Principal Component Analysis: Klamt Descriptors for 470 Compounds

statistics used for autoscaling	Sig2	Sig3	Hbdon3	Hbacc3	CSA
mean	46.59	9.574	0.8126	1.555	150.3
std dev	26.01	24.08	1.808	1.922	44.86
summary statistics	Comp1	Comp2	Comp3	Comp4	Comp5
standard deviation	1.42	1.31	1.00	0.40	0.32
percentage of variance	40	34	20	3	2
loadings (eigen-vectors)	Comp1	Comp2	Comp3	Comp4	Comp5
Sig2	0.6607	−0.1404	0.0278	0.7287	−0.1094
Sig3	0.1487	0.7268	−0.0009	0.1047	0.6623
Hbdon3	0.4715	−0.5215	−0.1529	−0.4416	0.5361
Hbacc3	0.5416	−0.4243	−0.1880	−0.4790	−0.5117
CSA	0.1605	0.0047	0.9698	−0.1832	−0.0109

Table 5. Correlation Matrix of the Remaining Four Descriptors (Orthogonalized to Size)

E	1							
S	0.68	1						
A	0.24	0.45	1					
B	0.06	0.47	0.26	1				
Sig2	0.34	0.78	0.71	0.73	1			
Sig3	−0.39	−0.10	−0.49	0.58	0.02	1		
Hbdon3	0.24	0.42	0.92	0.29	0.71	−0.48	1	
Hbacc3	−0.10	0.29	0.25	0.84	0.58	0.65	0.17	1
	E	S	A	B	Sig2	Sig3	Hbdon3	Hbacc3

This procedure, often called “autoscaling”, is equivalent to carrying out the PCA on the correlation matrix rather than on the covariance matrix. For each descriptor set, the means and standard deviations used for autoscaling are given, followed by the standard deviations of the principal components themselves (i.e. the square roots of their eigenvalues) with the variance accounted for by each component as a percentage of the total. Finally the loadings are given. These loadings apply to the autoscaled descriptors and are equivalently the eigenvectors of the correlation matrix in each case.

Chemical Content Overlap of the Two Sets of Descriptors. While it is informative to look at the regressions of each of the Abraham descriptors in turn on the Klamt descriptors, and vice versa, when considering the information held in individual variables, a different methodology is required to compare the two sets each considered as a whole. A natural approach to this question is through canonical correlation analysis. Here, the association between two sets of variables is measured, in the first instance, by the largest correlation that can be found between two new variables: one being a linear combination of the variables in the first set and the other a linear combination of the variables in the

Table 6. Correlation Equations and Statistics on 35 Compounds

descriptor	C	Sig2	Sig3	Hbdon3	Hbacc3	CSA	N	R^2	SD	F
S	−0.263	0.029	−0.007	−0.229	−0.083	0.000	35	0.916	0.180	63.0
	0.122	0.003	0.006	0.065	0.058	0.001				
A	0.120	0.003	−0.005	0.063	0.027	−0.001	35	0.955	0.095	125.4
	0.065	0.002	0.003	0.035	0.031	0.001				
B	−0.179	0.002	0.008	0.073	0.056	0.001	35	0.956	0.088	126.6
	0.060	0.001	0.003	0.032	0.028	0.001				
E	−0.274	0.017	−0.035	−0.373	0.219	0.002	35	0.707	0.288	14.0
	0.196	0.005	0.010	0.105	0.092	0.002				
V	−0.276	0.000	0.001	0.003	−0.001	0.008	35	0.975	0.055	221.7
	0.038	0.001	0.002	0.020	0.018	0.000				

second set. After this first pair of new variables has been constructed, further “maximum correlation” pairs can be derived sequentially, under the constraint that each new variable is uncorrelated with all the previously constructed variables (being correlated only with its “partner” from the other space). Thus, each descriptor-space is decomposed into a new set of uncorrelated variables (or axes), in a manner analogous to principal components analysis (PCA), but with the aim of providing the “best” description of the association between two spaces, rather than (as with PCA) the best description of variance. The sequence of values for the correlation between the pairs of axes indicates the closeness of the two spaces over successive dimensions, and attempts can sometimes be made to give a physical interpretation to the axes by reference to the coefficients attached to the underlying variables.

Canonical correlation analysis applied to the full set of Abraham and COSMOments descriptors produced a correlation for the first pair of derived variables of 0.99. For both these new variables, the overwhelmingly dominant component was the size descriptor in each case (the McGowan characteristic volume V for the Abraham set and the surface area descriptor CSA for the Klamt set). Given that the simple pairwise correlation between these descriptors was also 0.99 to two places of decimals, it is reasonable to say that the two spaces share a common “size” dimension. The possibility that a power relationship (of the form $y = x^r$) might be a better representation of the relationship between V and CSA (since one is a volume and the other an area) was investigated, but this was found to be no better in practice than a simple linear relationship with nonzero intercept. Therefore, the common size dimension was estimated by the average of V and CSA, and all the remaining descriptors orthogonalized to this, and then autoscaled for balance. In this way, the problem was simplified to the comparison of 4- (as opposed to 5-) dimensional spaces, while retaining all the chemical information. The correlation matrix of the remaining descriptors (orthogonalized to size) is given in Table 5.

It is worth noting that the above is actually very similar to the corresponding portion of the correlation matrix of the original data, though the interpretation of the individual pairwise correlations themselves is not straightforward. This is because a high correlation between two variables may be interpreted as due to a common correlation with a third. While such effects can be examined to some extent by looking at the partial correlations (i.e. after orthogonalizing to selected “third variables”), they cannot be resolved purely statistically, as they depend on a scientific view of the causality behind the associations.

Canonical correlation analysis applied now to the reduced (and orthogonalized to size) set of descriptors produced correlations for the pairs of derived variables of 0.98, 0.94, 0.76, and 0.17, respectively. Incidentally, and reassuringly, these agree to 2 places of decimals with those for dimensions 2 to 5 in the canonical correlation analysis of the original data. This implies excellent correspondence between the spaces over 2 of their 4 dimensions, with fair agreement on a third. There was no immediately obvious interpretation of the first three pairs of axes with reference to the coefficients attached to the underlying variables. However, it is interesting that **E** only really makes a notable contribution to the fourth axis of the Abraham set. This is consistent with the finding that **E** has clearly the worst regression (judged by its R^2) on the Klamt set (both when considering the original variables, and the reduced sets orthogonalized to size). Indeed, it would appear that any additional information present in the Abraham set over the Klamt set is held largely in **E**. Notably, if we re-run the canonical correlation analysis on the orthogonalized variables, but now omitting **E**, the correlations are virtually unchanged at 0.98, 0.91, and 0.76 (only 3 can now be calculated).

To summarize, there is considerable overlap between the set of five Abraham descriptors and the set of five COSMOments. There is essentially a shared size axis and close commonality over 3 “chemical” dimensions orthogonal to that. No obvious interpretation of this “chemical commonality” in terms of the original descriptors was found, but it does seem likely that, for the Abraham set, **E** is the main repository of any additional information present. Incidentally though, approaching the problem from the other end, this does not mean that **E** contains no information unique to itself within the Abraham set but shared with the Klamt set. The reason for saying this is that there remained a clearly statistically significant regression of **E** on the Klamt set, even after orthogonalization of both to the other Abraham descriptors.

Linear Relationships between the Two Sets of Descriptors. A set of equations was constructed using a carefully selected training set of 35 compounds, Table 6. These 35 compounds (Table 2) were selected in such a way to cover both the numerical spread and the chemical diversity of molecules with known S, A, and B values. Within this training set the relationships, obtained using multiple linear regression, and relate each one of the Abraham descriptors to the five of Klamt’s, are quite reasonable with correlation factors ranging from 0.707 to 0.975 with excess molar refraction descriptor (E) showing the smallest correlation coefficient. The important experimental descriptors S, A, and B, however, are predicted within this training set showing

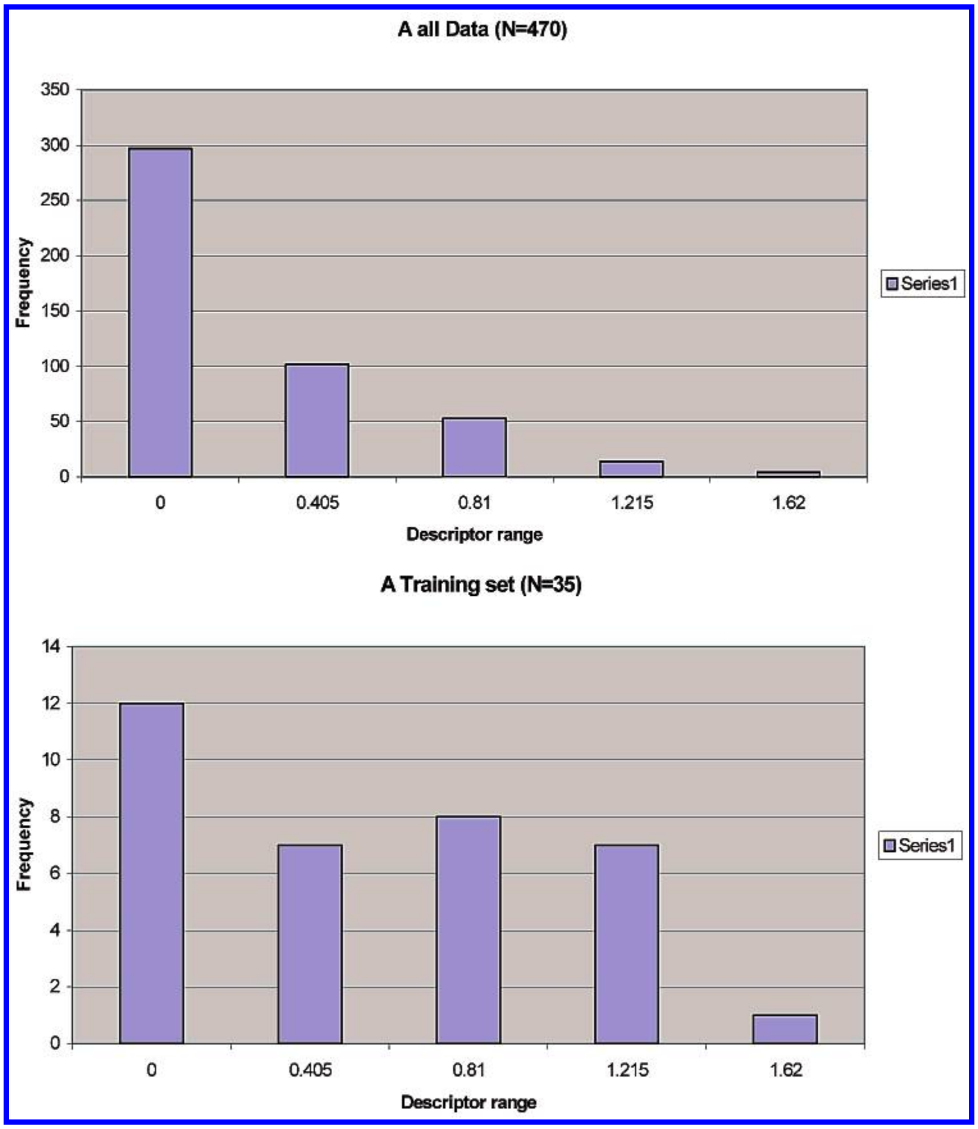

Figure 3. Histograms showing the distribution of A descriptor in the whole set (N = 470) and training set (N = 35).

standard deviations of 0.180, 0.095, and 0.088, respectively, for the three descriptors. The predictability of the three models was demonstrated by applying the models to calculate the descriptors of the remaining 435-compound test set. The three descriptors S, A, and B were calculated with standard deviations of 0.226, 0.089, and 0.113, respectively, which is very encouraging considering the small basis set used to train the model.

Before we analyze the contribution of each of the five COSMOments to each of the five Abraham descriptors, it is of primary importance to check how representative are the different data sets we have used. Of course, it is not feasible to ascertain if one or another is a representative set, over all the possible compounds that might comprise a data set, but it is useful to check one set against another, and to note whether the descriptors in the sets cover a reasonable range. In the process of demonstrating the predictive capabilities of our model, it became clear that a comparison between the various correlation equations obtained from the whole set of 470 compounds and the equations obtained from smaller training set chosen was important. This importance stems from the fact that the whole data set of 470 compounds was not as well balanced as far as the distribution of descriptors in each range. As an example, we give in Figure 3 histograms of the distribution of the **A**-descriptor, for both the 470-compound data set and the 35-compound training set. The later data set contains far more compounds with large **A**-descriptors; such a disparity could bias the correlations.

A summary of the five-term correlations of the five descriptors E, S, A, B, and V along with the statistical data are given in Table 7. Reasonable correlations have been obtained for all five descriptors. A problem however arises because of the big differences in the ranges of the five COSMOment descriptors. The range of the CSA descriptor, for example, is some 20-fold larger than that of Hbdon3 and Hbacc3 (with Sig2 and Sig3 somewhere between). As a result the regression coefficients for CSA are correspondingly smaller even for terms that contribute equally to the regressions. Therefore in Table 7 all coefficients are given in two significant figures instead of three decimal places so that its always possible to tell if a coefficient is big or not in comparison to its standard error. Also given in Table 7 are the t-statistics which have been used in stepwise backward elimination of parameters in the model. A cutoff point of t ≤ 3 was chosen. Hydrogen bond acidity (A) in Table 7 gives the best fit when compared to S and B and E correlations with a squared correlation value of 0.928 and a standard

Table 7. Correlation Equations and Statistics on 470 Compounds

descriptor	C	Sig2	Sig3	Hbdon3	Hbacc3	CSA	N	R^2	SD	F
S	−0.228	0.02285	−0.00588	−0.157	−0.037	0.00058	470	0.780	0.214	328.2
std. err.	0.037	0.00073	0.00088	0.011	0.011	0.00024				
t-statistic	−6.2	31.1	−6.7	−13.9	−3.5	2.4				
A	0.042	0.00084	−0.00639	0.0777	0.0688	−0.00025	470	0.928	0.074	1200.1
std. err.	0.013	0.00025	0.00030	0.0039	0.0037	0.00008				
t-statistic	3.0	3.3	−21.1	19.9	18.9	−3.0				
B	−0.062	0.00578	0.00648	0.0243	0.0196	0.00025	470	0.880	0.098	680.2
std. err.	0.017	0.00034	0.00040	0.0052	0.0049	0.00011				
t-statistic	−3.7	17.2	16.1	4.7	4.0	2.3				
E	−0.396	0.0162	−0.0156	−0.201	0.011	0.00316	470	0.504	0.368	94.3
std. err.	0.063	0.0013	0.0015	0.019	0.018	0.00041				
t-statistic	−6.3	12.8	−10.3	−10.3	0.6	7.7				
V	−0.2622	−0.00087	0.00035	0.0089	0.0040	0.0082	470	0.978	0.055	4083.8
std. err.	0.0094	0.00019	0.00023	0.0029	0.0027	0.00006				
t-statistic	−27.8	−4.6	1.5	3.1	1.4	133.8				

deviation of 0.074 and an F-statistic of 1200.1. From this equation the Abraham parameter for overall hydrogen bond acidity correlates mainly with the COSMOments representing donor and acceptor capacity. To check how significant the contributions of Sig2, Sig3, and area (CSA) COSMOments are in the correlation, the regressions were executed without these parameters. Sig3 was found to contribute strongly to the regression ($t = -21.1$) as the fitted 4-term equation without it has an SD of 0.103. Omitting both Sig2 and CSA gives as almost as good a fit as the five-term equation with an SD of 0.075.

$$A = 0.030 - 0.006\text{Sig3} + 0.085\text{Hbdon3} + 0.074\text{Hbacc3} \quad (13)$$

Here and elsewhere N is the number of data used in the

$$N = 470 \ R^2 = 0.926 \ SD = 0.075 \ F = 1941.5$$

correlation, R^2 is the overall correlation coefficient, SD is the standard deviation for the equation, F is the Fischer's F-statistic, and t is the t-statistic.

Hydrogen bond basicity correlates well with four of the COSMOments as shown by eq 14, and although the coefficients of Sig2 and Sig3 are rather small they do contribute strongly to the regression ($t = 17.2$ and 16.1, respectively). Leaving out each one in turn gives fitted 4-term equations with SDs of 0.125 and 0.122, respectively. Indeed and perhaps surprisingly (especially in view of the high pairwise correlation of B with Hbacc3 shown in Table 5), Hbdon3 and Hbacc3 jointly contribute less to the regression than do Sig2 and Sig3, in the sense that a fitted 2-term equation with only Sig2 and Sig3 in it has an SD of 0.103. CSA descriptor can certainly be omitted with little loss ($t = 2.3$). Eliminating the insignificant correlating parameters eq 14 is obtained for the B-descriptor.

$$B = -0.033 + 0.006\text{Sig2} + 0.007\text{Sig3} + 0.022\text{Hbdon3} + 0.017\text{Hbacc3} \quad (14)$$

Correlation of S polarity/polarizability descriptor gives a

$$N = 470 \ R^2 = 0.879 \ SD = 0.099 \ F = 841.2$$

reasonable relationship with Klamt's COSMOments, although the correlation is not as good as those for descriptors A and B. The polarity/polarizability descriptor correlates with all COSMOments apart from the area descriptor which gives a nearly zero coefficient. A fitted 4-term equation gives an SD of 0.215.

$$S = -0.161 + 0.023\text{Sig2} - 0.006\text{Sig3} - 0.162\text{Hbdon3} - 0.042\text{Hbacc3} \quad (15)$$

$$N = 470 \ R^2 = 0.777 \ SD = 0.215 \ F = 404.5$$

Molar refraction (E) descriptor does not correlate well at all with the five COSMOments. This is evident from the low correlation factor of 0.504 and a comparatively low F-statistic of 118.0. The descriptor is predicted rather poorly showing a rather large standard deviation of 0.368.

$$E = -0.394 + 0.016\text{Sig2} - 0.015\text{Sig3} - 0.198\text{Hbdon3} + 0.003\text{CSA} \quad (16)$$

$$N = 470 \ R^2 = 0.504 \ SD = 0.368 \ F = 118.0$$

The volume descriptor used by Abraham (McGowan's volume), not surprisingly, correlates very well with the five COSMOments. The overwhelmingly important term in this equation is (obviously) CSA to the extent that a fitted 1-term equation with CSA only has an SD value of 0.056.

$$V = -0.262 - 0.001\text{Sig2} + 0.001\text{Sig3} + 0.01\text{Hbdon3} + 0.008\text{CSA} \quad (17)$$

$$N = 470 \ R^2 = 0.978 \ SD = 0.055 \ F = 5092.2$$

As regards the characterization of the Abraham descriptors in terms of the theoredically calculated COSMOments, the V descriptor is well correlated to the susface area descriptor CSA. The S descriptor includes a number of interactions such as dipole/dipole and dipole/induced dipole and not surprisingly is correlated to a number of COSMOments as shown by eq 15. It might be expected that A should be well related to Hbdon3 and that B should be well related to Hbacc3. In the event, the correlation between the sum of the hydrogen bonding descriptors (eq 18) concludes that the combined descriptor A+B contains similar amounts of information to the combined descriptor Hbdon3 + Hbacc3.

$$(A+B) = 0.168 + 0.144(\text{Hbdon3} + \text{Hbacc3}) \quad (18)$$

$$N = 470 \ R^2 = 0.883 \ SD = 0.150 \ F = 3541.4$$

Table 8. Equations Obtained for the Five COSMOments from the Abraham Descriptors for 470 Compounds

descriptor	C	E	S	A	B	V	N	R^2	SD	F
Sig2	8.438	−6.004	28.365	38.687	37.034	3.040	470	0.930	6.941	1224.2
std. err.	0.956	1.020	1.330	1.323	1.443	0.989				
t-statistic	8.8	−5.9	21.3	29.2	25.7	3.1				
Sig3	1.001	−15.768	3.052	−56.000	64.375	0.502	470	0.862	9.002	578.5
std. err.	1.240	1.323	1.725	1.716	1.872	1.282				
t-statistic	0.8	−11.9	1.8	−32.6	34.4	0.4				
Hbdon3	−0.375	0.195	−0.257	6.079	0.474	0.171	470	0.853	0.697	538.7
std. err.	0.096	0.102	0.133	0.133	0.145	0.099				
t-statistic	−3.9	1.9	−1.9	45.8	3.3	1.7				
Hbacc3	0.154	−0.519	−0.303	0.635	5.906	−0.340	470	0.726	1.012	245.6
std. err.	0.139	0.149	0.194	0.193	0.210	0.144				
t-statistic	1.1	−3.5	−1.6	3.3	28.1	−2.4				
CSA	34.969	−4.846	7.587	−0.168	−5.812	121.38	470	0.978	6.625	4209.4
std. err.	0.913	0.974	1.269	1.263	1.377	0.944				
t-statistic	38.3	−5.0	6.0	−0.1	−4.2	128.6				

Table 9. Comparison of This Work with Previous Theoretical Predictive Methods of Abraham Descriptors

descriptor	method	N	SD	R^2	no. of descriptors used	training/test set	reference, author
A	MLR	39	0.144	0.636	2	training set	24, Platts et al.
A	MLR	15			2	test set	24, Platts et al.
A	MLR	55	0.107	0.880	2	training set	23 Dearden et al.
A	MLR	35	0.180	0.916	5	training set	this work
A	MLR	435	0.089		5	test set	this work
B	MLR	38			3	training set	25, Platts et al.
B	MLR	35	0.088	0.956	5	training set	this work
B	MLR	435	0.113		5	test set	this work
S	MLR	58	0.219	0.764	3, 4	training set	26, Platts et al.
S	PLS	58	0.176	0.840	4	training set	26, Platts et al.
S	MLR	32	0.178	0.787	6	test set	26, Platts et al.
S	PLS	32	0.175	0.796	2	test set	26, Platts et al.
S	MLR	8	0.175		4	test set	26, Platts et al.
S	MLR	67		0.940	17	training	19, Sevcic et al.
S	MLR	266		0.585	17	test set	19, Sevcic et al.
S	NN	69		0.908	7	training	19, Sevcic et al.
S	NN	264		0.537	7	test set	19, Sevcic et al.
S	MLR	35	0.180	0.916	5	training set	this work
S	MLR	435	0.226		5	test set	this work

[a] Abbreviations: MLR = multiple linear regression; PLS = partial least squares; NN = neural network.

Finally from the above correlations it can be concluded that the five COSMOments do not include some of the chemical information incorporated in descriptor E.

The regression exercise was repeated; only this time the reverse was done by regressing each one of the COSMOments against the five Abraham descriptors. Scale differences between the two sets of descriptors are not such a big problem here, and thus the regressions obtained, shown in Table 8, are easier to interpret because of larger coefficients. One significant point has to be made however from comparing the two sets of regressions in Tables 7 and 8. The regressions predicting A and B are not so clear-cut as far as the contributions of Hbdon3 and Hbacc3 are concerned. This, in comparison with the regressions predicting Hbdon3 and Hbacc3 in which case Hbdon3 has the strongest contribution coming from A and similarly Hbacc3 from B. This imbalance seems to arise because of the influence of Sig2 and Sig3 in the former case. Both of these COSMOments contribute strongly to the regression for B, while Sig3 contributes strongly to the equation for A. Since Sig2 and Sig3 are also correlated with Hbdon3 and Hbacc3, their presence influences the regression coefficients attached to Hbdon3 and Hbacc3. Interpreting this chemically is hard and it certainly depends on the actual chemical content of Sig2 and Sig3. Incidentally, the simple pairwise correlations between A, B, and Hbdon3, Hbacc3 are exactly as one would expect (Table 5).

Comparison with Other Methods. Finally, it is useful to compare the predictive power of our equations to give the three Abraham descriptors S, A, and B, with other methods reported in the literature. Such methods include the work carried out by Platts,[24−26] Sevcic,[19] and Dearden.[22,23] For this purpose, we have constructed Table 9 which tabulates in detail the parameters and statistics for various methods. The extent to which these methods can predict these descriptors can be assessed by comparing the standard deviations obtained from calculating descriptors for various test sets. We have tried to include, where available, the number of compounds used in each study to train the model used and also to test it by means of test sets. Platts in his work carries out calculations on the structure and properties of compounds using ab initio and DFT calculations. The calculated properties for these molecules are assessed for their ability to correlate and predict experimentally derived values of S, A, and B from Abraham's database. The approach of Sevcic 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. This work deals only with polarity/polarizability descriptor S. The predictability of the model is reasonable, but it has to be noted that a cutoff point of 1.0 for S has been imposed on the data.

Finally, Dearden and Ghafourian have calculated atomic charges and LUMO energies using semiempirical methods and correlated these with Abraham's hydrogen bond acidity descriptor. Their work used a training set of 55 compounds and the correlations reported are quite reasonable.

Our method deals with all three experimental descriptors S, A, and B. In all cases our model is trained using a small number of compounds (35) but tested on a reasonably large test set (435). The predictability of the model is as good as any other method so far. The deviations obtained are 0.226 for S 0.089 for A and 0.113 for B, close to the experimental error. For the S descriptor, the method of Platts gives rise to a lower standard deviation on his calculated test sets, but it has to be noted that these are rather small sets of compounds. For descriptor A, it is encouraging that all methods lead to small errors in their calculations. It seems that hydrogen bond acidity is the best calculated property with errors in the region of 0.09. As far as B is concerned there is not much room for comparison with other methods because of the lack of test sets in other calculations.

CONCLUSIONS

The five Abraham experimental descriptors and the five COSMOments of the 470-compound set exhibit a large overlap as far as chemical information content. This information however is distributed differently in the two sets and direct comparison of descriptors is not necessarily beneficial, though it does appear that any additional information present in the Abraham set over the Klamt set is mainly incorporated in the Abraham excess molar refraction descriptor E.

Problems of interpretation in the analyses described arise because of the high degree of correlation within both parameter sets. This is reflected in the relatively small eigen values associated with the final axes of principal components analyses carried out on the auto-scaled descriptors. Here, the last PC from the Abraham set and the last two PCs from the Klamt set each accounted for less than 5% of the total variance. Thus, on this basis, the Abraham descriptors provide a more stable description of the space that they span. However, how much of this difference depends on this particular data set is another matter.

The regression equations constructed can be useful in predicting the experimental Abraham descriptors from theoretically calculated COSMOments. It is difficult to interpret the coefficients of these equations chemically, however, because the information is not distributed in the same way in the two sets of descriptors. The predictability of these equations is tested successfully using a reasonably large set of data. Our model is shown to compare effectively with recent attempts[24−26] to calculate the Abraham descriptors from a theoretical basis. Computational methods, however, used for the prediction of these parameters are generally slow methods, something which is evident from the small number of compounds used in each of the studies we have included in our discussion. Our method could be useful when looking at specific compounds which are difficult to profile using experimental means. Although we use equations with $N = 35$ (Table 4) for direct comparison, we suggest that in general it is better to use the equations with the largest number of data points, that is the equations with $N = 470$ (Tables 7 and 8).

The statistical analysis of the two almost complete sets of LFER descriptors, the first being more heuristic and resulting from a long experience in LFER and the second resulting from a rather theoretical approach, and the analysis of the interrelations given in this paper, provides further evidence for the fact, that the solvent space is about five-dimensional with respect to partition behavior of solutes.

Supporting Information Available: Table with listings for molecule name, CAS registry number, FW, alternative name, sig2, sig3, Hbdon3, HBacc3, CSA, E, S, A, B, and V. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Sutter, J. M.; Jurs, P. C. Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.

(2) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of gas chromatographic retention indices of alkylbenzenes. *Anal. Chim. Acta* **1997**, *342*, 113−122.

(3) Mitchell, B. E.; Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.

(4) McClelland, H. E.; Jurs, P. C. Quantitative Structure−Property Relationships for the Prediction of Vapor Pressures of Organic Compounds from Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2000**, *40*(4), 967−975.

(5) 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**, 279−287.

(6) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T. QSPR studies on vapor pressure, aqueous solubility and the prediction of water−air partition coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.

(7) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(8) Katritzky, A. R.; Tatham, D. B.; Maran, U. Correlation of the solubilities of Gases and Vapors in Methanol and Ethanol with their Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 358−363.

(9) Kamlet, M. J.; Doherty, R. M.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. Solubility: a new look. *Chemtech* **1986**, *16*, 566−576 .

(10) Abraham, M. H.; Doherty, R. M.; Kamlet, M. J.; Taft, R. W. A new look at acids and bases. *Chemical. Brit.* **1986**, *22*, 551−554.

(11) 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* **1986**, 1451−1460.

(12) Abraham, M. H.; Whiting, G. S.; Doherty, R. M.; Shuely, W. J. Hydrogen bonding. Part 14. The characterisation of some N-substituted amines as solvents: comparison with gas−liquid chromatography stationary phases. *J. Chem. Soc., Perkin Trans. 2* **1986**, 1851−1857.

(13) Abraham, M. H.; Whiting, G. S.; Doherty, R. M.; Shuely, W. J. Hydrogen bonding. Part16. A new solute solvation parameter, π^{H}_{2} from gas chromatographic data. *J. Chromatogr.* **1991**, *587*, 213−228.

(14) Abraham, M. H. Hydrogen bonding. Part 31. Construction of a solute effective or summation hydrogen-bond basicity. *J. Phys. Org. Chem.* **1993**, *6*, 660−684.

(15) Abraham, M. H. Application of solvation equations to chemical and biochemical processes. *Pure Appl. Chem.* **1993**, *65*, 2503−2512.

(16) Abraham, M. H. Scales of hydrogen-bonding: their construction and application to physicochemical and biochemical processes. *Chem. Soc. Revs.* **1993**, *22*, 73−83.

(17) 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.

(18) Havelec, P.; Sevcik, J. G. K. Extended Additivity Model of Parameter log(L16). *J. Phys. Chem. Ref. Data* **1996**, *25*, 1483−1439.

(19) Svozil, D.; Sevcik, J. G. K. Neural Network Prediction of the Solvatochromic Polarity/Polarisability Parameter π_2H. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 338−342.

(20) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835−845.

(21) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors using a Group Contribution Approach. 2. Prediction of Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71−80.

(22) Dearden, J. C.; Ghafourian, T. Hydrogen Bonding Parameters for QSAR: Comparison of Indicator Variables, Hydrogen Bond Counts, Molecular Orbital and Other Parameters. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 231−235.

(23) Dearden, J. C.; Ghafourian, T. The Use of Atomic Charges and Orbital Energies as Hydrogen-bonding -donor Parameters for QSAR Studies: Comparison of MNDO, AM1 and PM3 Methods. *J. Pharm. Pharmacol.* **2000**, *52*, 603−610.

(24) Platts, A. J. Theoretical Prediction of Hydrogen Bond Donor Capacity. *Phys. Chem. Chem. Phys.* **2000**, *2*(5), 973−980.

(25) Platts, A. J. Theoretical Prediction of Hydrogen Bond Basicity. *Phys. Chem. Chem. Phys.* **2000**, *2*(14), 3115−3120.

(26) Lamarche, O.; Platts, A. J.; Hersey, A. Theoretical Prediction of the polarity/poarizability parameter π_2H. *Phys. Chem. Chem. Phys.* **2001**, *3*(14), 2747−2753.

(27) Politzer, P.; Murray, J. S. *Quantitative Treatments of Solute/Solvent Interactions*; Elsevier Science B. V.: 1994.

(28) Famini, G. R.; Penski, C. A.; Wilson, L. Y. Using theoretical descriptors in quantitative structure activity relationships: some physicochemical properties. *J. Phys. Org. Chem.* **1992**, *3*, 395−408.

(29) Famini, G. R.; Loumbev, V. P.; Frykman, E. K.; Wilson, L. Y. Using theoretical descriptors in correlation analysis of adenosine activity. *Quantum Struct.-Act. Relat.* **1998**, *17*, 558−564.

(30) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

(31) Klamt, A. *J. Phys. Chem.* **1995**, *99*, 2224.

(32) Klamt, A. COSMO and COSMO-RS. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, L., Eds.; Wiley: New York, 1998; Vol. 2, pp 604−615.

(33) Klamt, A.; Jonas, V.; Buerger, T.; Lohrenz, J. C. W. *J. Phys. Chem.* **1998**, *102*, 5074.

(34) Klamt, A.; Eckert, F. *Fluid Phase Equilibria* **2000**, *172*, 43.

(35) Klamt, A.; Eckert, F.; Hornig, M. *J. Comput.-Aid. Mol. Design* **2001**, *15*, 355.

(36) JMP statistical package V. 3.2.5; SAS Institute Inc.: 1989−1999; http://www. JMPdiscovery.com.

CI025530O