Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1037 (2004) 29–47
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Review

Determination of sets of solute descriptors from chromatographic measurements

Michael H. Abraham*, Adam Ibrahim, Andreas M. Zissimos

Department of Chemistry, University College London, 20 Gordon Street, London WC1H OAJ, UK

Abstract

The use of gas–liquid chromatographic (GLC) retention data to obtain sets of solute descriptors is outlined, with reference to the schemes of Laffort and of Weckwerth. The method of Snyder and Dolan to obtain a set of solute descriptors from reverse phase high performance chromatographic (RP-HPLC) measurements is described. The work of Abraham on the construction of solvation parameters, or descriptors, from water–solvent partitions, GLC retention data and RP-HPLC data is considered in some detail. A comparison is made between the schemes of Laffort, Weckwerth and Abraham, and it is shown that the latter two yield exactly the same fits for a test data set of gas–methanol partition coefficients, although the distribution of chemical information amongst the terms in the multiple linear regressions is not quite the same. A comparison between the above 'experimental' descriptors and theoretical descriptors is made, and it is shown that the experimental Abraham and the theoretical Klamt descriptors encode almost the same chemical information. It is concluded that for processes that entail transfer of a solute from one phase to another, only a small number of solute descriptors, no more than five or six, is needed to provide a reasonably accurate analysis of the process.

Keywords: Reviews; Solute descriptors; Solute–solvent interactions; Solvation equation; Partition coefficients

Contents

* Corresponding author. Fax: +44-2076797463.
E-mail address: m.h.abraham@ucl.ac.uk (M.H. Abraham).

0021-9673/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2003.12.004

1. Introduction

Because of the ease of use, the availability of commercial equipment, and the reproducibility of measurements, it is not surprising that chromatographic methods have long been used to obtain properties or 'descriptors' that characterize compounds. Both gas–liquid chromatography (GLC) and high-performance liquid chromatography (HPLC) can operate at very low compound concentrations and so, with few exceptions, can yield descriptors for compounds as the simple monomeric species.

The assignment of more than one 'descriptor' to a compound was first based on the solubility parameter theory of Hildebrand [1]. A number of multi-component systems were developed [2,3], that of Hansen [4–7] being especially comprehensive. Hansen characterized compounds in terms of partial solubility parameters: a dispersion solubility parameter, δ_d, a polar solubility parameter, δ_p, and a hydrogen bond solubility parameter, δ_h. The Hansen system has found wide applicability in polymer chemistry, as described by Barton [8], and was later extended by Karger et al. [9,10].

Cramer [11,12] selected six particular physical properties, viz. hydration energy, water–octanol partition coefficient, boiling point, molar refraction, volume and vaporization enthalpy and through factor analysis obtained six characteristic descriptors for some 500 compounds, denoted as *B*, *C*, *D*, *E* and *F*. These six descriptors were later used to correlate a number of biological properties with some success [13,14]. Unfortunately, the physical properties selected are not ideal, because hydration energy and partition coefficient refer to compounds as single monomeric species, whereas boiling point and vaporization enthalpy refer to the bulk liquids. Nevertheless, the work of Cramer showed clearly that it was possible to derive useful compound descriptors from physical properties.

2. Descriptors from GLC data

The earliest work on descriptors of solutes from GLC data was that of Rohrschneider [15,16] who used solute factors to calculate Kovats retention indices; McReynolds later extended the method [17]. Weiner and Howery [18,19] used factor analysis on the GLC data of Rohrschneider and McReynolds to obtain eight rather abstract solute factors. Somewhat later, Karger et al. [9,10] applied their partial solubility parameters to chromatographic retention, thus demonstrating that non-chromatographic data could be used to describe chromatographic retention. The most convincing work on the calculation of solute descriptors solely from GLC data was that of Laffort and Patte [20]. These workers first used the GLC retention data of McReynolds [22] on 25 stationary phases and through factor analysis obtained five solute descriptors for 75 compounds [20]. Note that the numerical values of the solute descriptors [20] are not the same as those in the later paper of Laffort and co-workers [21]. In this paper, Patte et al. [21] generated their own GLC retention data on five stationary phases and obtained the five solute factors for 240 compounds. The five stationary phases used are listed in Table 1, and the five solute factors are shown in Table 2. Laffort and co-workers [21,23] used the solute factors in Table 2 to correlate a number of physicochemical and biochemical properties. In general the solute factors have not been widely used, although Voelkel and Janas [24] characterized a number of GLC stationary phases in this way.

Table 1
The five stationary phases used by Laffort et al. [20,21]

No.	Phase
Z	Zonyl E7
C	Carbowax 1000
T	Tricyanoethoxypropane
P	Polyphenyl ether (6 rings)
D	Diethylene glycol succinate

Li et al. [25] used the retention data of Laffort and co-workers [21] on the five phases shown in Table 1, together with their own retention data for 53 compounds on eight capillary columns [26] and their own unpublished retention data on six basic phases to set out scales of solute dipolarity/polarizability, ${\pi_2}^{Ca}$, and solute hydrogen bond acidity, ${\alpha_2}^{Ca}$. Although Li et al. [25] used the corresponding scales of Abraham (see Section 4) as starting points, the final values of ${\pi_2}^{Ca}$ and ${\alpha_2}^{Ca}$ differed appreciably from the Abraham scales. A rather different method was used by Li et al. [27] to construct a scale of solute hydrogen bond basicity, denoted as ${\beta_2}^{C}$. Retention data, as $\log k$, were determined on a fluorinated benzyl alcohol and the corresponding fluorinated benzyl methyl ether, and ${\beta_2}^{C}$ defined as follows:

$$ {\beta_2}^{C} = \frac{\log k^{ALCOHOL} - \log k^{ETHER} + 0.089 + 0.23\delta_2}{2.15} \quad (1) $$

In Eq. (1), δ_2 is a polarizability correction factor taken as 0 (aliphatic solutes), 0.5 (polyhalogenated solutes) and 1.0 (aromatic solutes).

In subsequent papers, Li et al. [28,29] used the above solute descriptors in an equation for the characterization of stationary phases:

$$ \log k = c + l \log L^{16} + s{\pi_2}^{Ca} + d\delta_2 + a{\alpha_2}^{Ca} + b{\beta_2}^{C} \quad (2) $$

Table 2
The five solute factors of Laffort and co-workers [20,21]

Factor	Interpretation [20,21]
α	Apolar factor, proportional to volume
ω	Orientation factor, proportional to dipole moment for simple molecules
ε	Electron factor, related to dispersion interactions
π	Hydrogen bond acidity
β	Hydrogen bond basicity

Good correlations were obtained, although a number of strong bases were outliers and were omitted from the correlations [29]. The Carr descriptors (Li et al.) have not been widely applied, however, and the recent work of Weckwerth et al. [30] describes the construction of new scales altogether.

Weckwerth et al. [30] set out to obtain solute descriptors on the lines of those used in Eq. (2) and of those of Abraham, but which would correspond to chemically distinct properties. The key equation relating chromatographic retention data to the solute descriptors is Eq. (3), where we have used the superscript 'v' to avoid confusion with other descriptors.

$$\log k = c + vV^{v} + pP^{v} + \mathrm{d}D^{v} + aA^{v} + bB^{v} \quad (3)$$

The descriptors are V^{v} the solute volume, P^{v} the solute polarizability, D^{v} the solute dipolarity, A^{v} the solute hydrogen bond acidity and B^{v} the solute hydrogen bond basicity. Weckwerth et al. [30] used a data set of 53 compounds on seven stationary phases at various temperatures [31]. They assigned values of $D^{v} = A^{v} = 0$ for cyclohexane, $D^{v} = 1$, $A^{v} = 0$ for benzonitrile, and $A^{v} = 1$ for phenol. The term in bB^{v} was redundant because the stationary phases had no hydrogen bond basicity. The calculation of the volume descriptor was initiated by use of McGowan's volume, V; see Section 4.3. For the alkanes $V^{v} = 6.56 + V$.

The result of the calculations is a set of descriptors V^{v}, P^{v}, D^{v} and A^{v} for 53 compounds. The cavity formation descriptor V^{v} for homologous series is of interest because the CH_2 increment seems to vary from series to series, e.g. from 12.98 (benzene, toluene, ethylbenzene, and propylbenzene) to 16.82 (methanol, ethanol, and propan-1-ol). The descriptors are claimed to be chemically distinct solute parameters. Whether or not Weckwerth et al. [30] have achieved this cannot be ascertained. Actually, this makes little difference to the usefulness of Eq. (3) in the correlation and prediction of GLC retention data, for which it is eminently suitable. However, the present method cannot be used to correlate other processes such as water–solvent partitions and HPLC retention data, because the important B^{v} descriptor cannot be obtained from GLC retention data on non-acidic stationary phases. Weckwerth et al. [30] attempted to overcome this problem by using Abraham's B parameter in the correlation of a number of processes. However, one might just as well use the complete equation of Abraham (see Section 4), especially since the Weckwerth descriptors are available for only 53 compounds whereas the Abraham descriptors are known for several thousand compounds.

Since the Laffort descriptors and the Weckwerth descriptors have been obtained solely from GLC retention data, they should both be able to describe other GLC data better than descriptors obtained more generally. It will be of interest to see if one set leads to better correlations than the other set, although the Weckwerth descriptors are preferred on the grounds that they are much easier to interpret.

3. Descriptors from HPLC data

The main physicochemical use of HPLC data has been in the determination of water–octanol partition coefficients, $P_{o/w}$. Numerous workers have described HPLC systems, that, once calibrated, can be used to obtain $\log P_{o/w}$ values just from retention factors, as $\log k$. Dross et al. [32] have surveyed some of the literature, but there are numerous other papers on this subject [33–36]. The use of HPLC retention data to construct scales of solute parameters has received surprisingly little attention. Roses et al. [37] have developed a one-parameter system that shows considerable promise in the prediction of retention factors, as $\log k$ values, but the only attempt to develop a multi-parameter system is that of Snyder and co-workers [38–40].

Wilson et al. [38] start with retention factors, as $\log k$ values, for 67 varied solutes on ten different RP-HPLC stationary phases, all with 50% acetonitrile as the mobile phase. They then reduced the 67×10 data matrix to a 67×5 matrix, where all the solute information is compressed into five descriptors. Wilson et al. [38] also examined large data sets of 86 solutes on five other columns, 61 of the solutes being different to the original 67 solutes, and so were able to extend the number of characterized solutes considerably.

The final equation for $\log k$ is:

$$\log k = \log k_{ref} + H\eta' + S\sigma' + A\beta' + B\alpha' + C\kappa' \quad (4)$$

Wilson et al. [38] write Eq. (4) slightly differently, but we use the form of Eq. (4) because it shows the relationship to Abraham's equation, see Section 4. The independent variables η', σ', β' and α' are descriptors of the solute molecules in the neutral form. The descriptor κ', the ion-exchange parameter, refers to neutral compounds (that is non-acidic and non-basic compounds) as such, and to basic compounds in their protonated form. The coefficients H, S, A, B and C are corresponding properties of the stationary phase. The solute descriptors were interpreted as shown in Table 3.

In a second paper, Wilson et al. [39] investigated the retention of the 67 solutes as a function of temperature and mobile phase composition, and in the final paper of the series [40] a detailed analysis of the interpretation of the solute descriptors was given. It was pointed out that the hydrophobicity descriptor, η', increased with increase in carbon number for a homologous series, as shown in Table 4. However, there appears to be not any strong connection with volume, as can be seen in Table 4 from the volumes of naphthalene

Table 3
The RP-HPLC solute descriptors of Wilson et al. (Snyder) [38]

Descriptor	Interpretation [38]
η'	Hydrophobicity
σ'	Steric parameter
β'	Basicity
α'	Acidity (partly)
κ'	Cation-exchange parameter

Table 4
Some values of the hydrophobicity descriptor, η', of Wilson et al. (Snyder) [40] and the McGowan volume, V

Solute	η'	V
Benzene	−0.434	0.716
Toluene	−0.213	0.857
Ethylbenzene	0.000	0.998
Propylbenzene	0.240	1.139
Butylbenzene	0.480	1.280
Naphthalene	−0.053	1.085
Anthracene	0.353	1.454

and anthracene, so that the actual solute factors that determine η' are not very clear.

The σ' descriptor is calculated relative to ethylbenzene ($\sigma' = 0.000$) and so can be positive or negative. Wilson et al. [40] were careful to point out that steric selectivity might well be different from shape selectivity.

The β' descriptor was characterised [40] as arising from hydrogen bonding between solutes and non-ionized silanols in the stationary phase. (Rather oddly, the role of the 50% water in the mobile phase seems to be of no importance). Because of restricted access to silanols, β' is very dependent on steric effects of groups near to the basic site in the solute. Thus β' is 0.89 (*N*,*N*′-dimethylformamide) but only 0.22 (*N*,*N*′-dibutylformamide). However, this cannot be the whole story; the nitro group in 1-nitropropane and nitrobenzene is not sterically hindered, and yet their β' values are less than those of benzene or toluene, see Table 5 where a number of other hydrogen bond basicity values are listed [29,41,42].

Interpretation of the α' descriptor presented some difficulty [40], and it was suggested that more than one type of solute-phase interaction was involved, one of which was probably the solute acting as a hydrogen bond acid towards basic sites in the stationary phase. Finally, κ' is an interesting descriptor that relates to ionic interaction between charged solutes, specifically protonated bases, and ionized silanols in the stationary phase.

Whatever the interpretation of the new RP-HPLC descriptors, they should more accurately describe RP-HPLC retention data than descriptors derived from several sources of data, as pointed out by Wilson et al. [38]. This can only be ascertained by using the determined descriptors to analyze new data sets, the latter then being 'test' sets. This was done in the fourth paper of the series by Gilroy et al. [43]. These workers examined retention data for a particular subset of 16 compounds on no less than 92 RP-HPLC systems, in order to apply Eq. (4). It might be argued that for a multiple linear regression equation with five independent variables, a data set of 16 is far too small. In the event, Gilroy et al. [43] found that the original descriptors required revision in order accurately to represent the log *k* data. They achieved this by an iterative procedure in which the original descriptors were used to obtain coefficients in Eq. (4), the latter were then used to re-calculate the descriptors; these were used to obtain revised coefficients which in turn yielded re-calculated descriptors. Eventually a best fit of descriptors and coefficients were obtained. Not surprisingly, the final equation fits of the log *k* data were very good. Gilroy et al. [43] suggested that the iterative procedure resulted in 'minor revision' of the descriptors. Values of the initial and revised descriptors are in Table 6. Bearing in mind the scale of the descriptors, many of the changes seem to be much too large to be classed as 'minor'. For the original 67 compounds [38] the range of descriptors is 2.385 (η'), 1.988 (σ'), 1.113 (β'), 3.679 (α') and 1.658 (κ'). But for 5,5-diphenylhydantoin, the original and revised descriptors differ by 1.258 (σ') and by 1.016 (α'), substantial portions of the entire range for the 67 compounds.

Of course, solute descriptors based on experimental data will always be subject to change. It remains to be seen if the system of Snyder and Dolan will yield a set of descriptors that can be used in general to analyze RP-HPLC data, or whether variations in experimental procedures, such as mobile phase, stationary phase, buffers, etc will require a number of different sets of solute descriptors. There is clearly a trade off between sets of descriptors that are of general applicability but which fit retention data less well, and sets that are of limited applicability but which fit retention data in certain specific systems much better. Subsequent to the original three papers, Wang and Carr [44] examined retention data on 22 solutes in order to derive global linear solvation energy relationships. However, they chose to work with the Abraham descriptors rather than with Eq. (4).

Table 5
Some values of solute hydrogen bond strength

Solute	β'^{a}	β_2^{Cb}	B^{c}	β_2^{Hd}
Benzene	0.013	0.10	0.14	0.15
Toluene	0.004	0.11	0.14	0.14
1-Nitropropane	0.005	0.18	0.31	0.25
Nitrobenzene	−0.009	0.21	0.28	0.34

[a] Refs. [38,40].
[b] The hydrogen bond basicity descriptor of Li et al. (Carr) [29].
[c] The hydrogen bond basicity descriptor of Abraham [41].
[d] The 1:1 hydrogen bond basicity descriptor of Abraham et al. [42].

4. The system of Abraham

4.1. Introduction

We deal with this system separately, because it uses data both from GLC and HPLC, as well as data on water–solvent partition coefficients. A starting point is the solvatochromic solvent parameters developed by Kamlet and co-workers [45–50], that were used as solvent parameters in a general equation [51] for the effect of solvents on a given solute,

$$Y = Y_0 + d\delta + s\pi_1^* + a\alpha_1 + b\beta_1 + d(\delta_H)^2 \quad (5)$$

In Eq. (4), Y is a property of a given solute in a series of solvents and δ, π_1^*, α_1, β_1, and $(\delta_H)^2$ are the independent

Table 6
Revised and original descriptors of Snyder and co-workers [38–40,43]

Solute	η'	σ'	β'	α'	κ'
Acetophenone					
Revised	−0.744	0.133	0.059	−0.152	−0.009
Original	−0.748	0.186	0.039	−0.047	−0.009
Benzonitrile	−0.703	0.317	0.003	0.080	−0.030
	−0.715	0.245	0.016	−0.020	−0.026
Anisole	−0.467	0.062	0.006	−0.156	−0.009
	−0.473	0.042	0.001	−0.052	−0.019
Toluene	−0.205	−0.095	0.011	−0.214	0.005
	−0.206	−0.133	0.004	−0.014	−0.008
4-Nitrophenol	−0.968	0.040	0.009	0.098	−0.021
	−0.956	0.057	−0.034	0.217	−0.017
5,5-Diphenylhydantoin	−0.940	0.026	0.003	0.568	0.007
	−0.881	1.284	−0.046	−0.448	0.029
cis-Chalcone	−0.048	0.821	−0.030	0.466	−0.045
	−0.052	0.817	−0.024	0.066	−0.021
trans-Chalcone	0.029	0.918	−0.021	−0.292	−0.017
	0.032	0.918	−0.030	0.179	−0.042
N,*N*-Dimethylacetamide	−1.903	0.001	0.994	−0.012	0.001
	−1.921	0.000	1.000	0.000	0.000
N,*N*-Diethylacetamide	−1.390	0.214	0.369	−0.215	0.047
	−1.341	0.402	0.409	0.097	0.065
4-*n*-Butylbenzoic acid	−0.266	−0.223	0.013	0.838	0.045
	−0.272	−0.280	0.015	1.024	0.044
Mefenamic acid	0.049	0.333	−0.049	1.123	−0.008
	0.038	0.262	−0.039	0.917	−0.006
Nortriptyline	−1.163	−0.018	−0.024	0.289	0.845
	−1.169	0.059	−0.036	0.381	0.833
Amitriptyline	−1.094	0.163	−0.041	0.300	0.817
	−1.096	0.049	−0.030	0.321	0.834

variables, that is descriptors of the solvents, as follows: δ is an empirical polarizability correction term, π_1^* the solvent polarizability/dipolarity, α_1 the solvent hydrogen bond acidity, β_1 the solvent hydrogen bond basicity and $(\delta_H)^2$ the solvent Hildebrand cohesive energy density. This equation became known as the solvatochromic equation and is still one of the most widely used equations for the interpretation of solvent effects. The same workers then reasoned [52] that a similar equation could be used for the study of solute effects, that is for a series of solutes in a given solvent. The independent variables would then be solute descriptors, and the equation could be written as,

$$\mathrm{SP} = c + \mathrm{d}\delta + s\pi^* + a\alpha + b\beta + vV \quad (6)$$

In Eq. (6), SP is a property of a series of solutes (for example $\log P_{o/w}$) and the descriptors now refer to properties of the solutes, V being the solute volume. Eq. (6) was successfully [52] applied to several physicochemical and biochemical processes. One important drawback of the method was that the solvent parameters π_1^* and β_1 had to be used as surrogates for the (then) unobtainable corresponding true solute parameters, π^* and β. Although this was reasonable for compounds that were unassociated as solvents, it was not valid for compounds that were associated as solvents. In addition, parameters for solids and gases could not be obtained in this way. There was also a difficulty over the solute hydrogen bond acidity parameter, and a new descriptor, α_m, had to be invented; the subscript 'm' refers to monomeric species in the case of compounds such as water and alcohols. Eq. (6) was successful enough to indicate that the general principles were correct and so Abraham began the task of obtaining true solute parameters that could be used in a similar-type equation to Eq. (6). The method of Abraham is now some 15 years old [53], and it is 10 years since a comprehensive review was published [41], and so we give a rather detailed account.

4.2. General principles

The first step was to consider processes in which the only, or main, step was the transfer of a solute from one phase to another. Partitions between solvent phases, expressed in

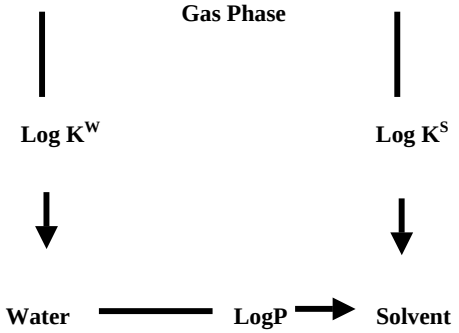

Fig. 1. Transfer of a solute from the gas phase to water and a solvent, and from water to the solvent.

terms of a partition coefficient, P, could be regarded as the resultant of two gas–solvent partitions, defined in terms of an equilibrium constant, K, as shown in Fig. 1:

$$K = \frac{\text{conc. of solute in a solvent}}{\text{conc. of solute in the gas phase}} \quad (7)$$

Then the factors that influence the partition between two solvent phases can more easily be expressed in terms of the factors that influence partition between the gas phase and a solvent, because we have that:

$$\log P = \log K^{S} - \log K^{W} \quad (8)$$

Here, K^S is an equilibrium constant for partition of a solute from the gas phase into a given solvent, and K^W the equilibrium constant for partition of the same solute from the gas phase to water. There a number of methods of separating out such factors, one of the most popular being the cavity theory of solution [54] which is an integral part of Pieriotti's scaled particle theory of solution [55]. In this theory, see Fig. 1, the solution of a gaseous solute is composed of three terms:

(a) A cavity of suitable size to accommodate the solute is created in the solvent. This step involves the endoergic breaking of solvent–solvent interactions; these will be proportional to the size of the cavity and hence to the size of the solute.
(b) The solvent molecules round the cavity are reorganized into their equilibrium position for interaction with the solute. The Gibbs free energy of reorganization is negligible. However, the enthalpy and entropy of reorganization may be large.
(c) The solute is inserted into the reorganized cavity, and various solute–solvent interactions are set up. These interactions are exoergic and aid the processes of solution.

Considerable simplification is effected if the solvent phase is constant, and only the solute changes. Then solvent properties need not be considered at all, and only relevant properties or 'descriptors' of the solute need to be devised. In step (a) either the solute volume, V, or the L descriptor (see later) was taken as the solute 'size' parameter. In step (c) there will be a number of solute–solvent interactions, all of which, in principle, should be related to given solute properties. In practice, it is not possible to separate out exactly the various interactions, especially those due to dipole and induced dipole effects, and the solute properties or descriptors finally used are given in Table 7; both the old complicated notation and the new notation are shown.

Table 7
Notation of the Abraham descriptors

Descriptor	Old symbol	New symbol
Excess molar refraction	R_2	E
Dipolarity/polarizability	π_2^H	S
Overall hydrogen bond acidity	$\Sigma\alpha_2^H$	A
Overall hydrogen bond basicity	$\Sigma\beta_2^H$	B
McGowan volume	Vx	V
Gas–hexadecane partition coefficient	$\log L^{16}$	L

The descriptors shown in Table 7 were combined into two linear equations, Eqs. (9) and (10). The former was designed to deal with transfers from the gas phase to a condensed phase, and the latter for transfers from one condensed phase to another.

$$\text{SP} = c + eE + sS + aA + bB + lL \quad (9)$$

$$\text{SP} = c + eE + sS + aA + bB + vV \quad (10)$$

In Eq. (9) the dependent variable, SP, can be $\log K^S$ as shown in Fig. 1, or can be GLC retention data for a series of solutes such as $\log t$(rel) or $\log V_g$ or I, where t(rel) is the relative retention time, V_g the retention volume and I the Kovats retention index. In Eq. (10), SP can be $\log P$ or $\log k$ where k is the HPLC retention factor, etc. The coefficients c, e, s, a, b, and l or v, can be found by standard procedures for multiple linear regression analysis.

4.3. The E, V and L descriptors

The definition of E is straightforward [56]. It is the molar refraction of the compound calculated using McGowan's volume, MR_X, less the molar refraction of an alkane with the same McGowan volume. The molar refraction itself is defined as,

$$\text{MR}_X = 10\left[\frac{(\eta^2 - 1)}{(\eta^2 + 2)}\right]V \quad (11)$$

where η is the refractive index of the compound as a pure liquid at 20 °C, and V is in units of $(\text{cm}^3\,\text{mol}^{-1})/100$. MR_X thus has units of $(\text{cm}^3\,\text{mol}^{-1})/10$. For compounds that are solid at 20 °C a refractive index for the liquid at 20 °C can be calculated by the ACD software [57]. It is interesting that molar refraction is one of the few properties that is the same for gaseous solutes as for liquid solutes, even for associated liquids such as water. $(\text{MR}_X)_{\text{alkane}}$ is given by,

$$(\text{MR}_X)_{\text{alkane}} = 2.83195V - 0.52553 \quad (12)$$

and so E is computed as,

$$E = (\mathrm{MR}_X) - 2.83195V + 0.52553 \quad (13)$$

The units of E are the same as those of MR_X, that is $(\mathrm{cm}^3\,\mathrm{mol}^{-1})/10$. A computer program, VR, is available from the authors to calculate E. It requires as input the solute molecular formula and the number of rings in the molecule (to calculate V) and the refractive index at 20 °C. E has been found to be very nearly an additive property, and so another method of obtaining E is through addition of fragment values.

The McGowan volume can simply be calculated from atomic fragments and the number of bonds in a molecule, all bonds being counted as one, no matter whether single, double or triple. It is not necessary actually to count the number of bonds, B_n, in complicated molecules because the algorithm of Abraham [41] can be used, Eq. (14), where N_a is the total number of atoms and R_g the number of rings. The VR program, above, calculates V using as input only the compound molecular formula, and the number of rings in the compound; N_a is of course obtained from the molecular formula.

$$B_n = N_a - 1 + R_g \quad (14)$$

The excess molar refraction, E, is derived from the refractive index function, and hence gives a measure of the polarizable electrons in a molecule. So E can be taken as an indication of the solute–solvent interaction that arises through the presence of polarizable electrons in the solute. The solute volume, V, was set up as a measure of the cavity effect, that is the endoergic effect of disrupting solvent–solvent bonds. However, solute volume is always well correlated with molar refraction and with polarizability [58], and so will include not only the endoergic cavity effect but also exoergic solute–solvent effects that arise through solute polarizability.

The L descriptor is defined [41] as the logarithm of the equilibrium constant in Eq. (6), where the solvent is hexadecane at 25 °C; $L = \log K$ (gas to hexadecane). Abraham et al. [59] showed that L could be obtained by direct measurement of retention volumes by GLC, using a hexadecane stationary phase thermostatted at 25 °C. Since then, Abraham and co-workers [60–62] have shown that GLC data, SP, for solutes on a non-polar stationary phase, usually at 100 °C or higher, can be correlated through the simple equation,

$$\mathrm{SP} = c + eE + lL \quad (15)$$

Once the phase has been calibrated with solutes of known SP, E and L values, then other L values can be obtained for further solutes for which SP and E are known. In this way, L values for hundreds of solutes with L values up to 7.71 (decylbenzene) were obtained. The determination of L through equations like Eq. (15) is a very simple direct experimental method. Again, although L may be regarded as another 'size' parameter, it will also include exoergic solute–solvent effects, just as the V descriptor does.

Li et al. [63] have also shown that L values can be obtained from GLC retention data on rather non-polar columns at temperatures between 60 and 120 °C, along the above lines. Mutelet and Rogalski [64] used conventional packed columns with hexatriacontane and pentacontane stationary phases at 100 and 150 °C to obtain L values up to 10.7, and confirmed a number of large L values up to 10.7 determined by Abraham and co-workers [60–62] previously. A very novel approach was also reported by Mutelet and Rogalski [64] who used a temperature gradient method with a DB-1 capillary column operated between 40 and 320 °C. Reduced retention times were related to L values for standard compounds through a fitting equation,

$$L = \exp\left[\frac{f(t_R)}{T}\right] \quad (16)$$

Because DB-1 is slightly polar, different fitting equations are needed for different series of solutes, which may be non-polar or moderately polar. Once this is done, the method can be used to obtain extraordinarily large L values, up to 18.7, very quickly indeed.

4.4. The S, A and B descriptors

Eq. (14) can be used to correlate GLC retention data for solutes on non-polar phases such as squalane or OV-101 at 100 °C. However, if the stationary phase is polar, then Eq. (14) will apply only to non-polar solutes such as alkanes. For polar compounds that have no hydrogen bond acidity, one additional descriptor is needed [60–62] that refers to the dipolarity/polarizability of the solute, that is descriptor S. Thus for relative retention times, t(rel), of solutes on di-*n*-propyl tetrachlorophthalate at 90 °C, Abraham and Whiting [61] found,

$$\log t(\mathrm{rel}) = -3.433 + 1.640S + 0.618L \quad (17)$$

A large number of equations of this type, or equations with the additional E descriptor, were used to obtain S descriptors, mostly for reasonably volatile solutes [60–62].

GLC phases that are polar are invariably also hydrogen bond bases, for example the dialkyl phthalates. Then if the solutes that are studied on such stationary phases are not hydrogen bond acids, equations such as Eq. (18), or equations in E, S and L, will suffice to correlate retention data, SP. However, if the data set includes hydrogen bond acids, then any correlation equation will require also the A descriptor [56,60–62], as shown in Eq. (18).

$$\mathrm{SP} = c + eE + sS + aA + lL \quad (18)$$

Then in order to obtain A values for compounds, not only must the stationary phase be calibrated, but values of SP, E, S and L must be known. The actual scale of A values was constructed by using 1:1 hydrogen bond acidities [65] as provisional values in order to set up a scale, but the final obtained values constitute a new scale. This GLC stepwise procedure was the original method of obtaining A (and also

Table 8
Hydrogen bond basicities for 'variable basicity' solutes

Solute	B	B^0
Aniline	0.41	0.50
m-Toluidine	0.45	0.55
N,*N*-Dimethylaniline	0.41	0.47
m-Chloroaniline	0.30	0.36
m-Methoxyaniline	0.59	0.70
Pyridine	0.52	0.47
4-Methylpyridine	0.54	0.43
Quinoline	0.54	0.51
Isoquinoline	0.54	0.47
Indole	0.22	0.31
Imidazole	0.78	0.50
Pyrazole	0.45	0.34
Purine	1.08	0.78
Dimethylsulfoxide	0.97	0.76
Diphenylsulfoxide	0.96	0.88
Triphenylphosphine oxide	1.50	1.32

E, *S* and *L* values). As will be outlined later, much more elegant methods are now available.

There are very few GLC stationary phases that have been examined that are significant hydrogen bond acids, and almost no such phases are commercially available. This means that the above GLC step by step method cannot be extended to the determination of *B* values. Abraham [66] therefore turned to the use of water–solvent partition data, as log *P* values, in order to establish a scale of hydrogen bond basicity. The equation used was the full Eq. (10), where SP is now log *P*.

Once again, equations could be established for solutes that were not hydrogen bond bases, for several water–solvent partitions. Then provisional values of *B* were assigned to solutes, using the 1:1 hydrogen bond basicity scale [42] as an approximation. By a process of iteration, a second set of equations was established, new solute basicities were calculated, a third set of equations was established, etc. Finally, a self-consistent set of equations and *B* values was obtained.

During the course of this work, it became apparent that for certain solutes in certain water–solvent systems, the solute hydrogen bond basicity was not constant. This phenomenon had been observed previously by Leahy et al. [67] who identified some solutes containing the S=O and P=O groups as possessing variable basicity. It seems that not all solutes containing these groups behave in an anomalous way and in Table 8 are listed anomalous solutes. Note that sulfones, sulfonamides, sulfonates and phosphates do not show variable basicity. In order to deal with this problem, Abraham [66] identified other types of variable basicity solutes such as anilines and pyridines, and assigned an alternative hydrogen bond basicity, B^0, to variable basicity solutes in water–solvent systems where the organic layer contains considerable quantities of water, see Table 9. It is important to note that these variable basicity solutes behave quite normally in other water–solvent partitions, and in all gas-solvent partitions, where the usual *B* descriptor can be used.

Table 9
Water–solvent partition systems for which the alternative B^0 descriptor is appropriate

Water–wet alcohols
Water–wet diethyl ether[a]
Water–wet diisopropyl ether
Water–wet ethyl acetate[b]
Water–wet butyl acetate

[a] Dibutyl ether is a borderline case.
[b] Not the esters olive oil and PGDP.

4.5. Descriptors via water–solvent partitions

The above methods for the determination of descriptors were worked out before equations for many water–solvent partitions had been developed. Now that a large number of such systems have been characterized through Eq. (10), it is possible to determine descriptors through log *P* values for a given solute in a number of water systems. As shown in Section 4.3, the descriptors *E* and *V* can readily be obtained, and so only the descriptors *S*, *A* and *B* in Eq. (10) need to be determined.

In principle, if log *P* values are known for a compound in three water–solvent systems which have been characterized by Eq. (10), then we have three equations and three unknowns (*S*, *A* and *B*) and so the unknowns can be evaluated. In practice, this procedure will only work satisfactorily if the coefficients in the three equations are substantially different. Two stratagems can be applied.

Firstly, log *P* values in a large number of water–solvent systems may be known or may be determined. Then the *S*, *A* and *B* values that lead to best reproduction of the log *P* values can be evaluated. This is illustrated in Table 10 for the case of ephedrine, for which log *P* values in a large number of systems have been determined and recorded in the MedChem data base [68]. The Medicinal Chemistry database, organized by Leo, is a most valuable source of experimental data that can be used in the determination of descriptors. From Table 10, with $E = 0.916$ and $V = 1.4385$, the 10 log *P* values could be reproduced with SD = 0.13 using $S = 0.79$, $A = 0.27$ and $B = 1.18$ units. It is often the

Table 10
Observed [68] and calculated values of log *P* for ephedrine[a]

Water–solvent system	Observed	Calculated
Octanol	1.12	1.18
Pentanol	1.37	1.41
Hexanol	1.22	1.17
Decanol	0.94	0.88
Dichloromethane	0.62	0.70
Trichloromethane	1.12	1.27
Heptane	−0.77	−0.64
Cyclohexane	−0.48	−0.69
Benzene	0.45	0.27
Diethyl ether	0.30	0.39

[a] With $E = 0.916$, $V = 1.4385$ and assigned values of $S = 0.79$ $A = 0.27$ and $B = 1.18$.

Table 11
Calculation of descriptors for the variable basicity solute, methyl phenyl sulfoxide[a]

Water–solvent system	Observed	Calculated
Octanol	0.55	0.55[b]
Trichloromethane	1.41	1.32
Hexane	−1.49	−1.48
Cyclohexane	−1.29	−1.28
Dibutyl ether	−0.86	−0.91
PGDP	−0.41	−0.37
Gas phase	–	7.37

[a] With $E = 1.104$, $S = 1.73$, $A = 0.00$, $B = 0.88$, $V = 1.0795$.
[b] With $B^0 = 0.71$.

case that one or more $\log P$ values are out of line, and values in the water isobutanol (observed 1.18, calculated 1.44) and water-toluene (observed 0.40, calculated 0.09) systems were left out.

For compounds with variable hydrogen bond basicity, the position is more complicated, and $\log P$ values in several water–solvent systems are required for any rigorous analysis. As an example we take methyl phenyl sulfoxide, for which $\log P$ values are available in seven water–solvent systems [68]. Knowing $E = 1.104$ and $V = 1.0795$, only two descriptors, S and B, need to be determined because $A = 0$. With $S = 1.73$ and $B = 0.88$, $\log P$ values in six of the systems are fitted very satisfactory, with SD = 0.06 log units, see Table 11. However, for one of the solvent systems that appears in Table 11, viz. the water–octanol system, the calculated value for $\log P$, −0.04, is far away from the observed value of 0.55 [68]. The latter value can be taken as reliable because it is a designated 'starred' value in the MedChem data base [68], and because it is exactly the same as the calculated $C\log P$ value, also 0.55 [69]. If we then take $E = 1.104$, $V = 1.0795$, $A = 0$, and $S = 1.73$ as found in Table 11, the alternative B^0 value of 0.71 is required to reproduce the 0.55 value for $\log P_{o/w}$.

The second method is to use a small number of systems, but to select them carefully so that the corresponding equations are as different as possible. Many years ago, Taylor et al. [70] argued that four water–solvent systems, if carefully chosen, were sufficient to encapsulate the information contained in the varied water–solvent systems available. They suggested that octanol, an alkane, chloroform and an ester (propylene glycol dipelagronate, PGDP) was a suitable quartet of solvents. Zissimos et al. [71] chose a slightly different quartet of solvents: octanol, cyclohexane, chloroform and toluene, partly because PGDP is not commercially available. They then examined a number of mathematical procedures to extract the three required descriptors for a compound which had $\log P$ values measured in the four water–solvent systems. These procedures were:

(a) The use of the 'Solver' facility in the Microsoft Excel spreadsheet.
(b) An in-house program similar to Solver, denoted as 'Descfit'.
(c) The use of multiple linear regression equations for each of the three descriptors.
(d) The use of three simultaneous equations, TripleX.

The Solver and Descfit methods are very similar and can work with any number of equations, not just four. Both of them calculate $\log P$ values using combinations of values for S, A and B, and the combination that leads to the smallest standard deviation between observed and calculated $\log P$ values is taken. The multiple regression method is quite different, and involves setting up separate equations for S, A and B in terms of the four $\log P$ values. Zissimos et al. [71] chose 47 varied compounds for which either the four $\log P$ values were known, or for which they were measured. The equations for required descriptors were found to be:

$$S = 0.049 - 0.092\log P_{o/w} + 0.229\log P_{chl} - 0.713\log P_{cyc} + 0.625\log P_{tol} + 0.355E - 0.188V \quad (19)$$
$$N = 47, \quad R^2 = 0.916, \quad SE = 0.152, \quad F = 73.0$$

$$A = 0.108 + 0.261\log P_{o/w} - 0.155\log P_{chl} - 0.248\log P_{cyc} + 0.171\log P_{tol} - 0.049E - 0.097V \quad (20)$$
$$N = 47, \quad R^2 = 0.964, \quad SE = 0.058, \quad F = 177.2$$

$$B = -0.089 - 0.033\log P_{o/w} + 0.338\log P_{chl} + 0.178\log P_{cyc} - 0.587\log P_{tol} + 0.137E + 0.595V \quad (21)$$
$$N = 47, \quad R^2 = 0.881, \quad SE = 0.137, \quad F = 49.2$$

In the final method, TripleX, the five parameter Eq. (10) is reduced to a three-parameter equation, because the terms eE and vV are known,

$$(\log P - eE - vV) = sS + aA + bB \quad (22)$$

Four such equations are constructed, in $\log P_{o/w}$, $\log P_{chl}$, $\log P_{cyc}$ and $\log P_{tol}$. Then any three of these equations can be solved for the three unknowns S, A and B. There will be four combinations of the three simultaneous equations, and the four solutions can be averaged. The TripleX method can be applied to any number of water–solvent systems, although some type of computer program is needed when the number of combinations becomes very large.

Zissimos et al. [71] found that the method of multiple regression (c) was significantly worse than the other three. In terms of ease of use, methods (a) and (b) were preferred. For the 47 training compounds the standard deviations on observed and calculated values using methods (i) or (ii) were: S (0.16), A (0.07) and B (0.16). For a test set of 13 drug compounds $\log P_{cyc}$ was predicted with SD = 0.06 and 29 varied $\log P$ values were predicted with SD = 0.48 log units. Some of the results of Zissimos et al. [71] are collected in

Table 12
Determination of *S*, *A* and *B* using log *P* values in four water–solvent systems

Solute	Solver/Descfit			Preferred values[a]				
	S	*A*	*B*	*E*	*V*	*S*	*A*	*B*
Benzoic acid	0.94	0.68	0.35	0.730	0.9317	0.90	0.59	0.48
Phenol	0.91	0.60	0.30	0.805	0.7751	0.89	0.60	0.30
p-Toluidine	1.02	0.06	0.56[a]	0.923	0.9571	0.95	0.23	0.52[b]
Aniline	0.98	0.19	0.52[a]	0.955	0.8162	0.96	0.26	0.50[b]
Resorcinol	1.32	1.02	0.50	0.980	0.8338	1.11	1.09	0.52
Benzene	0.51	0.00	0.15	0.610	0.7164	0.52	0.00	0.14
Toluene	0.45	0.00	0.16	0.601	0.8573	0.52	0.00	0.14
Salicyclic acid	0.65	0.72	0.42	0.890	0.9900	0.84	0.71	0.38
Phenylacetic acid	1.07	0.58	0.59	0.730	1.0700	1.01	0.59	0.61
1-Naphthol	1.10	0.63	0.37	1.520	1.1441	1.05	0.60	0.37
Ibuprofen	0.45	0.57	0.85	0.860	1.7771	0.59	0.59	0.81
Lidocaine	1.34	0.02	1.38	1.010	2.0589	1.50	0.12	1.21
Procaine	1.58	0.42	1.23	1.135	1.9767	1.68	0.44	1.23
2-Chlorophenol	0.38	0.38	0.41	0.853	0.8975	0.88	0.32	0.31
4-Nitrophenol	1.65	0.94	0.22	1.070	0.9493	1.72	0.82	0.26

[a] These are the values for the B^0 descriptor. The *B* descriptor was calculated as 0.52 for *p*-toluidine and 0.45 for aniline.
[b] The values for *p*-toluidine and aniline are for the descriptor B^0. The *B* descriptor is 0.45 for *p*-toluidine and 0.41 for aniline.

Table 12. Results for the Solver and Descfit methods are so similar that we give the average values. For comparison, we give in Table 12 the preferred descriptor values, as determined from all the available information. The two sets of descriptors are in reasonable agreement, with the exception of 2-chlorophenol. For the 15 solutes in Table 12, the SD between the Solver/Descfit values, from the four log *P* measurements, and the preferred values are 0.17 (*S*), 0.07 (*A*) and 0.07 (*B*). We shall carry out a more detailed analysis later.

If the descriptors *S*, *A* and *B* are required for use in equations of the type as Eq. (10), then determination through log *P* measurements is probably the most convenient method. Either a well-chosen restricted set of water–solvent systems can be used, as in Table 12, or a large set of systems can be used, as shown in Table 10. In either case, the MedChem collection of log *P* values is an indispensable aid to the determination of descriptors.

4.6. Descriptors via HPLC measurements

As for log *P* data, HPLC retention data are best analyzed through Eq. (10). Thus *E* and *V* are calculated first, and *S*, *A* and *B* remain to be determined. The same principles obtain as for descriptors via log *P* values. If a small number of HPLC systems are used, then the corresponding calibration equations must be as far apart as possible. Plass et al. [72] obtained descriptors for a number of tripeptides using RP-HPLC gradient elution data on five systems. Zissimos et al. [73] carried out a much larger analysis, on the same lines to that on descriptors via log *P* measurements. They first characterized seven HPLC systems operated in the reverse phase mode with fast gradient elution. These seven systems were chosen to be as different as possible, that is as orthogonal as possible, using non-linear mapping [74–76]. The systems are listed in Table 13. In order to standardize the gradient elution procedure, the experimental log k' values were converted into the chromatographic hydrophobicity index, CHI, values as described by Valko and co-workers [74–76]. CHI values for 80 compounds, of known descriptors, were determined on all seven systems.

The 80 compounds were divided into a training, or fitting set, of 40 compounds, and a test set of 40 compounds. Then using Eq. (9), with SP = CHI, the seven systems were characterized using the 40 compound training set. Five mathematical methods were then used in order to predict descriptors for the 40 compound test set. These methods were the same as those used previously, see a–d above, with the addition of a modified regression method. The unmodified regression equations were as follows, with the systems designated as in Table 13. In the equations, under 'System 1' for example would be entered the CHI value for a given solute measured in system 1.

$$\begin{aligned} S = {} & 0.673 - 0.013(\text{System 1}) + 0.008(\text{System 2}) \\ & - 0.050(\text{System 3}) + 0.015(\text{System 4}) \\ & + 0.023(\text{System 5}) - 0.019(\text{System 6}) \\ & + 0.013(\text{System 7}) + 0.273E + 1.398V \end{aligned} \quad (23)$$

$$N = 40, \quad r^2 = 0.950, \quad \text{SE} = 0.212, \quad F = 63.03$$

$$\begin{aligned} A = {} & 1.499 + 0.010(\text{System 1}) - 0.001(\text{System 2}) \\ & + 0.002(\text{System 3}) - 0.011(\text{System 4}) \\ & - 0.025(\text{System 5}) + 0.026(\text{System 6}) \\ & - 0.007(\text{System 7}) + 0.149\,\boldsymbol{E} - 0.369V \end{aligned} \quad (24)$$

$$N = 40, \quad r^2 = 0.896, \quad \text{SE} = 0.136, \quad F = 28.58$$

Table 13
The RP-HPLC, gradient elution systems studied by Zissimos et al. [73]

No.	Stationary phase	Mobile phase
1	Luna $C_{18}(2)$ 50 × 4.6 mm (Phenomenex)	Aq. acetonitrile
2	Luna $C_{18}(2)$ 50 × 4.6 mm (Phenomenex)	Aq. methanol
3	Luna $C_{18}(2)$ 50 × 4.6 mm (Phenomenex)	Aq. trifluoroethanol
4	Perfluorooctyl silica 50 × 4.6 mm (ES Industries)	Aq. trifluoroethanol
5	PLRP-S-100 50 × 4.6 mm (Polymer Labs.)	Aq. acetonitrile
6	Develosil CN 50 × 4.6 mm (Phenomenex)	Aq. methanol
7	Develosil CN 50 × 4.6 mm (Phenomenex)	Aq. acetonitrile

$$B = 0.103 + 0.001(\text{System 1}) - 0.007(\text{System 2}) + 0.008(\text{System 3}) + 0.002(\text{System 4}) - 0.009(\text{System 5}) - 0.020(\text{System 6}) + 0.012(\text{System 7}) + 0.09E + 0.788V \quad (25)$$

$$N = 40, \quad r^2 = 0.963, \quad \text{SE} = 0.108, \quad F = 86.46$$

The various equations, set up with the 40 training set of compounds, were then used to predict descriptors for the 40 compound test set that had not been used to construct the equations. Results of predictions of the teat set using Solver or Descfit were identical, as observed for the log *P* calculations, and use of the more soundly based modified regression instead of the original unmodified regression actually made very little difference. Details are in Table 14, where the standard deviations are between the predicted descriptors and the 'preferred' descriptors.

There is very little difference between the Solver/Descfit method and the regression method; both are somewhat better than the TripleX program. Details for the same 15 compounds that are in Table 12 are given in Table 15; we shall discuss these values later. For correlations with RP-HPLC data, the B^0 descriptor is prefered to *B*. Hence for 'variable basicity' solutes, a system of equations that are all for RP-HPLC data will yield the B^0 descriptor.

4.7. Estimate of the L descriptor from log P data

In Section 4.3, it was shown how the *L* descriptor could be obtained experimentally through determination of GLC retention data. In some cases, it may not be practical to carry out such determinations, but it is possible to obtain an estimate of *L* through log *P* values. The use of log *P* values cannot lead directly to the determination of the *L* descriptor, because the defining Eq. (10) for log *P* uses *V* as the size descriptor and not *L*. However, if the five descriptors *E*, *S*, *A*, *B* and *V* have been determined via log *P* values, then the water-hexadecane partition coefficient, P_{16}, and the gas–water partition coefficient, K^W, can be calculated through the well known equations, [77,78]

$$\log K^W = -0.994 + 0.577E + 2.549S + 3.813A + 4.841B - 0.869V \quad (26)$$

$$\log P_{16} = 0.087 + 0.667E - 1.617S - 3.587A - 4.869B + 4.433V \quad (27)$$

Now $\log K^W$ and $\log P_{16}$ are related to *L* through Eq. (28), cf. Eq. (8), as can be seen from Fig. 1.

$$\log P_{16} = L - \log K^W \quad (28)$$

Hence knowing $\log K^W$ and $\log P_{16}$ it is possible to estimate *L*.

Table 14
Standard deviations of the predicted descriptors for a 40-compound test set by the HPLC method [73]

Method	*S*	*A*	*B*
Solver, Descfit	0.29	0.15	0.15
Regressions	0.30	0.15	0.12
TripleX	0.33	0.21	0.15

4.8. Descriptors from solubilities

Abraham, Acree and co-workers [79–85] have shown that solubility data for a given solute in water and solvents can also be used to generate descriptors. The partition coefficient

Table 15
Determination of *S*, *A* and *B* using CHI values in seven RP-HPLC gradient elution systems

Solute	Solver/Descfit[a]		
	S	*A*	*B*
Benzoic acid	0.73	0.66	0.47
Phenol	1.05	0.49	0.39
p-Toluidine	1.03	0.02	0.60
Aniline	1.20	0.00	0.64
Resorcinol	0.77	1.06	0.69
Benzene	0.45	−0.30	0.29
Toluene	0.41	−0.09	0.14
Salicyclic acid	0.93	0.84	0.32
Phenylacetic acid	0.86	0.59	0.64
1-Naphthol	1.39	0.73	0.17
Ibuprofen	0.63	0.60	0.84
Lidocaine	0.99	0.29	1.27
Procaine	1.89	0.57	1.10
2-Chlorophenol	1.00	0.69	0.17
4-Nitrophenol	1.78	0.68	0.24

[a] The values for *p*-toluidine and aniline in the last column are for the descriptor B^0.

of a solute between water and a solvent can be obtained from the ratio of solubilities in water, S_W, and the solvent, S_S, provided that a number of conditions are met:

(a) The solute in equilibrium with the two saturated solutions must be the same species. This means that neither hydrate nor solvate formation should occur.
(b) The secondary medium activity coefficient should be near unity for the solute in the two saturated solutions. In practice, this means that the solubility should not be too high.
(c) For ionizable solutes such as strong proton acids or bases, the solubilities must refer to the unionized species.

Granted that conditions (a–c) are met, the partition coefficient is given by,

$$P = \frac{S_S}{S_W} \quad \text{or} \quad \log P = \log S_S - \log S_W \tag{29}$$

Of course, P will refer to the particular state of the solvent, which nearly always will be the dry solvent, and not the wet solvent as found in practical partitions. It is therefore very important to distinguish equations on the lines of Eq. (10) that refer to 'hypothetical' partition between water and a dry solvent, and those that refer to 'practical' partitions, where the solvent is saturated with water (and the water is saturated with the solvent).

Once a series of $\log P$ values have been obtained through Eq. (29), then exactly the same procedure as used for 'practical' $\log P$ values is set up. A number of dry solvents have been characterized via Eq. (10) and also via Eq. (11); the characteristic coefficients are in Tables 16 and 17.

The analysis of solubilities has been carried out for a variety of solutes, such as *trans*-stilbene [79], Buckminsterfullerene [80], diuron [81] and other pesticides [82], polyaromatic hydrocarbons [83] such as fluorene [84], benzil [85], and ferrocene [86]. *trans*-Stilbene [79] is a straightforward example, and in Table 18 are given solubilities, as $\log S_S$, and the corresponding $\log P$ values based on $\log S_W = -5.80$ where S_W is in mol dm^{-3}. The value of V is easily calculated as 1.5630 and E estimated as 1.45 from known values for benzene and styrene. With the descriptors given in Table 19, obtained through the Solver method, the various $\log P$ values were calculated as shown in Table 18. A value of $\log K_W$ was also available, and so a whole series of $\log S_W$ values could be calculated, and fitted to the set of descriptors used in Eq. (9); in this way a value of 7.525 for L was obtained.

The solubility method is therefore a very powerful way of obtaining descriptors. It complements the GLC method because it is applicable to very involatile compounds. The use of HPLC data is probably more convenient than the

Table 16
Characteristic coefficients in Eq. (10) for partitions between water and dry solvents

Solvent	c	e	s	a	b	v
Methanol/dry	0.329	0.299	−0.671	0.080	−3.389	3.512
Ethanol/dry	0.208	0.409	−0.959	0.186	−3.645	3.928
Propan-1-ol/dry	0.148	0.436	−1.098	0.389	−3.893	4.036
Butan-1-ol/dry	0.152	0.438	−1.177	0.096	−3.919	4.122
Pentan-1-ol/dry	0.080	0.521	−1.294	0.208	−3.908	4.208
Hexan-1-ol/dry	0.044	0.470	−1.153	0.083	−4.057	4.249
Heptan-1-ol/dry	−0.026	0.491	−1.258	0.035	−4.155	4.415
Octan-1-ol/dry	−0.034	0.490	−1.048	−0.028	−4.229	4.219
Decan-1-ol/dry	−0.062	0.754	−1.461	0.063	−4.053	4.293
Ethyleneglycol/dry	−0.269	0.586	−0.522	0.712	−2.492	2.708
TFE/dry	0.395	−0.094	−0.594	−1.280	−1.274	3.088
Propanone/dry	0.335	0.349	−0.231	−0.411	−4.793	3.963
Acetonitrile/dry	0.413	0.077	0.326	−1.566	−4.391	3.364
Hexane[a]	0.361	0.579	−1.723	−3.599	−4.764	4.344
Heptane[a]	0.325	0.670	−2.061	−3.317	−4.733	4.543
Octane[a]	0.223	0.642	−1.647	−3.480	−5.067	4.526
Nonane[a]	0.240	0.619	−1.713	−3.532	−4.921	4.482
Decane[a]	0.160	0.585	−1.734	−3.435	−5.078	4.582
2,2,4-Trimethylpentane[a]	0.318	0.555	−1.737	−3.677	−4.864	4.417
Hexadecane[a]	0.087	0.667	−1.617	−3.587	−4.869	4.433
Cyclohexane[a]	0.159	0.784	−1.678	−3.740	−4.929	4.577
Tetrachloromethane[a]	0.260	0.573	−1.254	−3.558	−4.588	4.589
Trichloromethane[a]	0.327	0.157	−0.391	−3.191	−3.437	4.191
Toluene[a]	0.143	0.527	−0.720	−3.010	−4.824	4.545
Benzene[a]	0.142	0.464	−0.588	−3.099	−4.625	4.491
Chlorobenzene[a]	0.040	0.246	−0.462	−3.038	−4.769	4.640
Carbon disulfide[a]	0.047	0.686	−0.943	−3.603	−5.818	4.921
Octan-1-ol/wet	0.088	0.562	−1.054	0.034	−3.460	3.814
Gas–water	−0.994	0.577	2.549	3.813	4.841	−0.869

[a] For these solvents, practical and hypothetical partitions are regarded as the same.

Table 17
Characteristic coefficients in Eq. (9) for partitions between the gas phase and dry solvents

Solvent	*c*	*e*	*s*	*a*	*b*	*l*
Methanol/dry	−0.004	−0.215	1.173	3.701	1.432	0.769
Ethanol/dry	0.012	−0.206	0.789	3.635	1.311	0.853
Propan-1-ol/dry	−0.028	−0.185	0.648	4.022	1.043	0.869
Butan-1-ol/dry	−0.039	−0.276	0.539	3.781	0.995	0.934
Pentan-1-ol/dry	−0.042	−0.277	0.526	3.779	0.983	0.932
Hexan-1-ol/dry	−0.035	−0.298	0.626	3.726	0.729	0.936
Heptan-1-ol/dry	−0.062	−0.168	0.429	3.541	1.181	0.927
Octan-1-ol/dry	−0.120	−0.203	0.560	3.560	0.702	0.939
Decan-1-ol/dry	−0.136	−0.068	0.325	3.674	0.767	0.947
Ethylene glycol/dry	−0.898	0.217	1.427	4.474	2.687	0.568
TFE/dry	−0.092	−0.547	1.339	2.213	3.807	0.645
Propanone/dry	0.154	−0.277	1.522	3.258	0.078	0.863
Acetonitrile/dry	−0.007	−0.595	2.461	2.085	0.418	0.738
Hexane[a]	0.292	−0.169	0.000	0.000	0.000	0.979
Heptane[a]	0.275	−0.162	0.000	0.000	0.000	0.983
Octane[a]	0.215	−0.049	0.000	0.000	0.000	0.967
Nonane[a]	0.200	−0.145	0.000	0.000	0.000	0.980
Decane[a]	0.156	−0.143	0.000	0.000	0.000	0.989
2,2,4-Trimethylpentane[a]	0.264	−0.230	0.000	0.000	0.000	0.975
Hexadecane[a]	0.000	0.000	0.000	0.000	0.000	1.000
Cyclohexane[a]	0.163	−0.110	0.000	0.000	0.000	1.013
Tetrachloromethane[a]	0.282	−0.303	0.460	0.000	0.000	1.047
Trichloromethane[a]	0.168	−0.595	1.256	0.280	1.370	0.981
Toluene[a]	0.121	−0.222	0.938	0.467	0.099	1.012
Benzene[a]	0.107	−0.313	1.053	0.457	0.169	1.020
Chlorobenzene[a]	0.053	−0.553	1.254	0.364	0.000	1.041
Carbon disulfide[a]	0.101	0.251	0.177	0.027	0.095	1.068
Octan-1-ol/wet	−0.198	0.002	0.709	3.519	1.429	0.858
Gas–water	−1.271	0.822	2.743	3.904	4.814	−0.213

[a] For these solvents, practical and hypothetical partitions are regarded as the same.

use of solubility data, however. The HPLC method requires much smaller sample sizes, and has one other important advantage. In the determination of log *k* or CHI values, the purity of the sample is not normally an issue. However, in the determination of solubilities, small amounts of impurities may lead to considerable experimental errors.

Table 18
Log *P* values for *trans*-stilbene obtained from solubilities, and log *P* values calculated from the descriptors in Table 19 [79]

Solvent	log S_S	log *P*	log *P* (calc)
Methanol/dry	−1.32	4.48	4.40
Ethanol/dry	−1.27	4.53	4.70
Octanol/dry	−1.10	4.70	4.76
TFE/dry	−2.04	3.76	3.70
Acetonitrile/dry	−0.74	5.06	5.09
Hexane	−1.14	4.66	4.58
Heptane	−1.13	4.67	4.65
Octane	−1.12	4.68	4.79
Nonane	−1.11	4.69	4.69
Decane	−1.11	4.69	4.64
2,2,4-Trimethyl-pentane	−1.32	4.48	4.54
Hexadecane	−1.13	4.67	4.65
Cyclohexane	−0.90	4.90	5.10
Tetrachloromethane	−0.40	5.40	5.38
Toluene	−0.26	5.54	5.62
Benzene	−0.18	5.62	5.65
Chlorobenzene	−0.16	5.64	5.58

4.9. A general method for the determination of descriptors

There is no reason why the determination of descriptors should be restricted to GLC data or to log *P* data or to HPLC data. Now that numerous GLC, log *P* and HPLC systems have been characterized through Eqs. (9) and (10), any combination of data can be used to determine *S*, *A* and *B*, or *S*, *A*, *B* and *L*. For example, in Tables 12 and 15 are given results obtained by using log *P* data in four systems or RP-HPLC data in seven systems. But we could combine all the data and analyze the eleven systems together. The most convenient methods are Solver or Descfit, which again lead to almost identical descriptors. The method is as before: *E* and *V* are known or can be calculated, and values of *S*, *A* and *B* are calculated that best reproduce the eleven experimental log *P* and CHI values. However, care has to be

Table 19
Descriptors for *trans*-stilbene [79]

Descriptor	Value
E	1.45
S	1.04
A	0.00
B	0.34
V	1.5630
L	7.525

Table 20
Determination of *S*, *A* and *B* using log *P* values in four systems and CHI values in seven RP-HPLC gradient elution systems

Solute	Solver/Descfit		
	S	*A*	*B*
Benzoic acid	0.92	0.63	0.40
Phenol	0.99	0.53	0.35
p-Toluidine	0.83	0.13	0.60[a]
Aniline	0.78	0.27	0.64[a]
Resorcinol	1.18	0.97	0.57
Benzene	0.35	−0.11	0.26
Toluene	0.30	0.00	0.18
Salicyclic acid	0.86	0.77	0.35
Phenylacetic acid	1.00	0.51	0.65
1-Naphthol	1.18	0.73	0.27
Ibuprofen	0.59	0.56	0.84
Lidocaine	1.25	0.14	1.30
Procaine	1.80	0.47	1.15
2-Chlorophenol	0.88	0.48	0.24
4-Nitrophenol	1.38	0.98	0.22

[a] These are for the B^0 descriptor. Values for the *B* descriptor are 0.55 (*p*-toluidine) and 0.44 (aniline).

taken because for any given solute, the average value of log *P* is very much less than the average value of CHI. If log *P* and CHI values are used as such, then Solver/Descfit will minimize the differences in calculated and observed CHI values with little regard to the log *P* values. Some method of weighting therefore has to be used. We use the average values of the dependent variable as a guide to a weighting scheme. In the present case, the CHI values are about 20 times the log *P* values. We therefore divide all the coefficients in the CHI equations and all the CHI values by 20 in order that the CHI equations do not have an undue proportional weight in the minimization procedure.

As an example, we use the compounds in Tables 12 and 15 and calculate *S*, *A* and *B* on this basis. Results are in Table 20. Note that the total set of eleven equations includes equations in B^0 and in *B*; it is possible to determine both of these basicities for variable basicity solutes. In Tables 12, 15 and 20 we have listed Descfit calculated values of *S*, *A* and *B* for the same data set of fifteen compounds obtained from four log *P* values, seven CHI values, and a combined set of the total eleven systems. Although the Descfit method is not quite as good as the regression method for the RP-HPLC systems, it is useful to compare the SD values of calculated and preferred descriptors from all three sets. These are listed in Table 21, together with SD values for the larger data sets. Rather surprisingly, results from the combined log *P* and HPLC systems for the 15 common solutes are almost the same as for the results for the log *P* systems alone. From Table 21 it seems as though the log *P* method is somewhat better than the RP-HPLC method in the determination of solute descriptors, as regards SD values.

Such a procedure as used for the log *P* plus HPLC systems, can greatly be extended, especially if GLC data are available. Then in principle, the four descriptors *S*, *A*, *B* and *L* can

Table 21
Standard deviations between calculated and preferred values of descriptors for series of solutes

Data set	SD (*S*)	SD (*A*)	SD (*B*)
log *P*, 15 compounds[a]	0.17	0.07	0.07
HPLC, 15 compounds[a]	0.23	0.18	0.11
Combined systems, 15 compounds[a]	0.16	0.09	0.08
log *P*, 47 compounds	0.15	0.07	0.16
HPLC, 40 compounds (training)	0.24	0.13	0.13
HPLC, 40 compounds (test)	0.29	0.15	0.15
HPLC, regression method[b]	0.24	0.13	0.11

[a] The same 15 compounds.
[b] Average of the training set and test set results.

be calculated simultaneously. Occasionally, the gas–water partition coefficient, K^W, is available. This is a very valuable piece of information in its own right, but also because it leads to values of gas-solvent partition coefficients, K^S, through Eq. (8). Once log K^S values are known, the *L* descriptor becomes available through Eq. (9).

An example of a calculation in which log *P* values, RP-HPLC data and GLC data were all used in the determination of descriptors is the solute *n*-octanoic acid [87] (Table 22). In this case, *E* could be obtained from the experimental liquid refractive index and the calculated value of *V*. Log *P* values in six 'practical' water–solvent systems were available [68], and a log K^W value of 4.09 was estimated by trial-and error. This enabled the corresponding six log K^S values to be deduced, via Eq. (8). In addition, log *k* values were available in two RP-HPLC systems for which coefficients in Eq. (10) were known, and Kovats retention indices were known in eighteen GLC systems, see ref. [87]. Because the absolute values of the Kovats indices are so much larger than those of log *P* or log *k*, it was again essential to weight the GLC data by dividing all the system coefficients and all the indices by a factor of 1000. With gas-solvent partition coefficients available as well as water–solvent partition coefficients and the GLC and RP-HPLC data, no less than 34 systems could be considered. Knowing that $E = 0.15$, and $V = 1.3102$, the remaining descriptors were determined as: $S = 0.65$, $A = 0.62$, $B = 0.45$, and $L = 4.680$ which reproduced the 34 data with an SD of 0.098 log units.

As well as numerous individual solutes, several large sets of solutes have been examined in this way, using chromatographic data from GLC and RP-HPLC as well as water–solvent partition data. These sets include 35 N-nitrosodialkylamines [88], and all the 75 polychloronaphthalenes [89].

5. Comparison of descriptors from GLC data

There are two comprehensive sets of solute descriptors obtained solely from GLC retention data, those of Laffort and co-workers [21,23] and those of Weckwerth et al. [30].

Table 22
Calculation of descriptors for *n*-octanoic acid [87][a]

System	SP	SP_{calc}	SP_{obs}
Water–octanol	$\log P$	2.948	3.050
Water–trichloromethane	$\log P$	2.062	2.170
Water–hexane	$\log P$	0.644	0.660
Water–heptane	$\log P$	0.852	0.630
Water–hexadecane	$\log P$	0.529	0.560
Water–benzene	$\log P$	1.711	1.670
Gas–water	$\log K_W$	4.153	4.090
RP-HPLC 50% methanol	$\log k$	1.177	1.288
RP-HPLC 75% methanol	$\log k$	0.125	0.177
Gas–octanol	$\log K$	7.103	7.140
Gas–trichloromethane	$\log K$	6.209	6.260
Gas–hexane	$\log K$	4.848	4.750
Gas–heptane	$\log K$	4.851	4.720
Gas–hexadecane	$\log K$	4.680	4.650
Gas–benzene	$\log K$	5.877	5.760
Gas–water	$\log K$	4.225	4.090
GLC-PA	*I*/1000	1.117	1.162
GLC-PB	*I*/1000	1.238	1.266
GLC-PC	*I*/1000	1.290	1.312
GLC-PD	*I*/1000	1.281	1.305
GLC-PE	*I*/1000	1.419	1.445
GLC-PF	*I*/1000	1.664	1.674
GLC-PG	*I*/1000	2.090	2.106
GLC-WS	*I*/1000	2.210	2.072
GLC-FFAP	*I*/1000	2.016	2.042
GLC-MLT	*I*/1000	1.145	1.172
GLC-DON	*I*/1000	1.861	1.966
GLC-GTA	*I*/1000	1.122	1.194
GLC-GTB	*I*/1000	1.108	1.165
GLC-CARB	*I*/1000	2.104	2.298
GLC-DEGS	*I*/1000	2.271	2.179
GLC-POLY	*I*/1000	1.015	1.265
GLC-TCEP	*I*/1000	2.520	2.447
GLC-Zonyl	*I*/1000	1.406	1.471

[a] With $E = 0.15$, $S = 0.65$, $A = 0.62$, $B = 0.45$, $V = 1.3102$, $L = 4.680$.

In order to compare their efficacy in fitting data, it is imperative that a set of data, not used in the construction of the solute descriptors, should be used. Unfortunately, the Laffort set includes only 240 solutes, and the Weckwerth set includes even fewer, 52 solutes only. It proved very difficult to identify a new set of GLC data that included enough of the Laffort and Weckwerth data sets to be statistically significant. We therefore resorted to recent published data [90] on the gas–wet ether partition coefficient, K^{ETHER}. This is a gas solvent system and so should be amenable to analysis through descriptors that are obtained from similar processes, namely gas-stationary phase. The gas–wet ether system has also an advantage that solute hydrogen bond basicity is not important [90], and so there is no need to augment the Vitha descriptor set to include a basicity parameter.

The Abraham equation, Eq. (8), with $SP = \log K^{ETHER}$, when applied to values for 114 solutes resulted in Eq. (30). All the required data is given in ref. [90].

$$\log K^{ETHER} = 0.206 - 0.169E + 0.873S + 3.402A + 0.882L$$
$$N = 114, \quad R^2 = 0.981, \quad SD = 0.262, \quad F = 1474.0 \tag{30}$$

The Weckwerth descriptors were available for only 23 compounds out of the 114 and result in Eq. (31). For comparison, the Abraham equation for the same set of compounds is given as Eq. (32)

$$\log K^{ETHER} = 0.88 - 0.03V^v + 0.41P^v + 2.39D^v + 3.45A^v$$
$$N = 23, \quad R^2 = 0.959, \quad SD = 0.242, \quad F = 82.3 \tag{31}$$

$$\log K^{ETHER} = 0.25 - 0.70E + 1.02S + 3.44A + 0.90L$$
$$N = 23, \quad R^2 = 0.959, \quad SD = 0.242, \quad F = 82.3 \tag{32}$$

Comparison of the two Abraham equations shows that the 23 data set is not a very representative subset of the 114 data set, but that does not prevent a useful analysis of Eqs. (31) and (32). First of all, the statistics are (amazingly) exactly the same. Thus for this very limited set of solutes, the Weckwerth descriptors and the Abraham descriptors both perform reasonably well. What is of more interest is how the various solute–solvent interactions are distributed amongst the four terms. The coefficients themselves are not enough, the entire terms have to be calculated for particular solutes. This is done with the Weckwerth descriptors for the solutes benzene, ethyl acetate and methanol, see Table 23. The largest interaction, on this scheme, is due to solute polarizability.

On the system of Abraham, the *lL* term includes both a cavity term and general dispersion interactions. The only way to separate these effects is through a separate calculation of the cavity term, as was carried out by Abraham et al. [78] for transfer from the gas phase to water. We adopt the same stratagem and use scaled particle theory, SPT, as outlined by Pierotti [55]. In order to derive the required interaction parameter, ε/k, and hard sphere diameter, σ, for ether solvent we calculated the values of $\log K^{ETHER}$ for a number of non-polar solutes for which only ε/k and σ were needed (and were available). Details are in Table 25.

Table 23
Solute–solvent interactions for solution of gaseous solutes into diethyl ether in terms of $\log K$, after Weckwerth et al. [30], see Eq. (31)

Solute	Cavity	Polarizability	Dipolarity	Acidity	Total	Total observed
Benzene	−2.35	4.02	0.43	0.00	2.98	3.08
Ethyl acetate	−2.25	3.44	0.93	0.00	3.00	3.06
Methanol	−0.88	1.29	0.60	1.07	2.96	2.89

Table 24
Determination of ε/k and σ for solvent diethyl ether; interaction and cavity effects are in terms of $\log K$ values

Solute	σ	ε/k	Cavity	Interaction	Total	$\log K^{\mathrm{ETHER}}$
Argon	3.40	122	−2.16	2.00	−0.16	−0.22
Methane	3.82	157	−2.55	2.61	0.06	0.02
Radon	4.36	290	−3.10	4.21	1.11	1.11
Ethane	4.40	236	−3.14	3.84	0.70	0.85

Cavity and interaction terms calculated with $\varepsilon/k = 500$ and $\sigma = 5.42$ for diethyl ether.

Once ε/k and σ were obtained for diethyl ether, then the calculations given in Table 24 could be carried out. Comparison of Table 25 with Table 23 shows that there is remarkable agreement between the two systems as regards the general size of the various effects On both the Weckwerth and Abraham systems, the largest solute–solvent interactions for the three listed solutes are general interactions, regarded as dispersion interactions by Abraham et al. and as polarizability interactions by Weckwerth et al. [30]. The only real difference is that on Abraham's system cavity effects, as calculated by SPT, are more negative than on Weckwerth's system, and consequently general interactions are more positive. Indeed, cavity effects on Weckwerth's system are only about half of those calculated by SPT. Why this is so is not clear. It maybe that Weckwerth et al. have not fully separated out cavity effects and polarizability effects; further examples are needed to come to any definite conclusion.

The Laffort descriptors can also be used to analyze the $\log K^{\mathrm{ETHER}}$ data. Equations were constructed using exactly the same data set for both the Laffort and the Abraham descriptors:

$$\log K^{\mathrm{ETHER}} = 1.19 + 0.81\alpha + 0.30\omega + 0.86\varepsilon + 1.07\pi + 1.44\beta \quad (33)$$
$$N = 34, \quad R^2 = 0.969, \quad \mathrm{SD} = 0.251, \quad F = 141.6$$

$$\log K^{\mathrm{ETHER}} = 0.31 - 0.18E + 0.90S + 3.33A + 0.89L \quad (34)$$
$$N = 34, \quad R^2 = 0.992, \quad \mathrm{SD} = 0.122, \quad F = 756.8$$

The statistics for the Laffort equation are considerably worse than those for the Abraham equation. Further work is needed fully to compare statistics for the Laffort and Weckwerth systems, but using the Abraham equations as a yardstick, in this one instance the Weckwerth system performs better than the Laffort system. The latter is not so easy to interpret. In particular, it seems strange that there is a substantial term in solute hydrogen bond basicity in the Laffort equation. It would be of considerable value if Weckwerth descriptors were available for a much larger set of solutes, in order to pursue these various comparisons.

6. Comparison with theoretical calculations

There have been several studies that seek to determine the number of 'indicators' or 'factors' that are necessary to account for GLC retention data of a selection of solutes on a selection of stationary phases. Of course, for solutes that are very similar, or for stationary phases that are very similar, only few factors are needed. However, a number of studies have dealt with a range of solutes and stationary phases; all have concluded that very few factors are needed. For a range of solutes with sets of non-acidic stationary phases, only three major factors seem to be necessary [91–93]. Possibly, if acidic stationary phases were included, the number of factors might increase to four, but this is still a small number. Other work also suggests that only a small number of factors is required. Lucic et al. [94] used the software program CODESSA to calculate 296 descriptors for each of 152 compounds, and then used these descriptors to analyze retention data on a given stationary phase. They suggest that a multiple linear regression (MLR) equation with only seven descriptors was a reasonable model.

Katritzky and Tatham [95] also used data of Abraham et al. [96,97] on gas to methanol [96] and gas to ethanol [97] partitions for analysis by the CODESSA method. These gas to solvent partitions are quite analogous to gas to stationary phase partition in GLC. Katritzky and Tatham [95] started with 550 calculated descriptors for each compound, but the final MLR equations, Eqs. (35) and (36) included only four descriptors. We have abbreviated the symbols for the descriptors somewhat, but a list is in Table 26.

$$\log K(\mathrm{MeOH}) = -1.113 + 0.0532P + 29.163\mathrm{HDCA} + 0.4195\mu + 0.8871\mathrm{IC} \quad (35)$$
$$N = 87, \quad R^2 = 0.9446$$

$$\log K(\mathrm{EtOH}) = 2.1110 + 0.0523P + 0.6732\mu + 0.0497\mathrm{HASA} + 0.2081\mathrm{HOMO} \quad (36)$$
$$N = 61, \quad R^2 = 0.9686$$

Table 25
Solute–solvent interactions for solution of gaseous solutes into diethyl ether in terms of $\log K$, after Abraham et al.

Solute	Cavity[a]	Dispersion[b]	Dip/Pol[c]	Acidity	Total	Total observed
Benzene	−4.23	6.43	0.45	0.00	2.86	3.08
Ethyl acetate	−4.41	6.59	0.54	0.00	2.93	3.06
Methanol	−2.80	3.61	0.38	1.46	2.86	2.89

[a] Calculated by SPT.
[b] Obtained as $(eE + lL)$ − cavity term.
[c] Dipolarity/polarizability.

Table 26
Descriptors used by Katritzky and Tatham [95]

Symbol	Descriptor
P	α-Polarizability
μ	Solute dipole moment
HDCA	A hydrogen bond donor descriptor
IC	A topological zeroth-order average information content
HASA	Hydrogen acceptor surface area
HOMO	Energy of the highest occupied molecular orbital

For comparison, the equations of Abraham et al. [96,97] are given as Eqs. (37) and (38). There are fewer solutes in the Katritzky equations, because descriptors for the rare gases could not be calculated by the CODESSA program.

$$\log K(\text{MeOH}) = -0.004 - 0.215E + 1.173S + 3.701A + 1.432B + 0.769L \quad (37)$$
$$N = 93, \quad R^2 = 0.9952, \quad \text{SD} = 0.13, \quad F = 3681$$

$$\log K(\text{EtOH}) = 0.012 - 0.206E + 0.789S + 3.636A + 1.311B + 0.853L \quad (38)$$
$$N = 68, \quad R^2 = 0.9966, \quad \text{SD} = 0.14, \quad F = 3534$$

What is interesting is that, once again, only a small number of solute descriptors is required in order to obtain reasonably good fits of data for gas to solvent phase transfers, in agreement with previous work on GLC retention data [91–93].

Although many calculational methods produce hundreds of solute descriptors, those of Famini et al. [98], and of Klamt et al. [99] are quite different in that only a small number of descriptors are calculated. The procedure of Famini and Wilson leads to six descriptors for each solute. These are closely related in principle to the descriptors used by Abraham et al. and by Weckwerth et al. in that they include volume, two basicity descriptors, two acidity descriptors, and a dipolarity/polarizability descriptor. The Famini and Wilson theoretical descriptors have been used to correlate several physicochemical properties through MLR equations.

The Klamt method, known as COSMO-RS, yields five descriptors as shown in Table 27. There is an almost exact match between the Klamt descriptors and those of Abraham et al. and of Vitha et al., and so it appeared interesting to see if the information content of these sets of descriptors is the same. There are not enough characterized solutes in the Weckwerth set to carry out a rigorous analysis, but Zissimos et al. [100] have compared the solute descriptors of Abraham et al and of Klamt.

Table 27
The solute descriptors of Klamt et al. [99]

Symbol	Descriptor
Sig2	Polarity/polarizability
Sig3	Polarity/polarizability
Hbdon3	Hydrogen bond acidity
Hbacc3	Hydrogen bond basicity
CSA	Surface area

Zissimos et al. [100] obtained five Abraham experimental descriptors, E, S, A, B and V, and calculated the five COSMO-RS descriptors for a varied solute set of 470 compounds. They showed that the information content of the two sets of five descriptors is very nearly the same, with the E descriptor incorporating a small amount of extra information. However, the chemical information is distributed differently amongst the descriptors, so there is no 1:1 correspondence between the descriptors (except between the two 'size' descriptors V and CSA). There were a number of inter correlations that might be useful in calculating descriptors of one set from descriptors in the other set. The best correlations were:

$$A = 0.042 + 0.00084\text{Sig2} - 0.00639\text{Sig3} + 0.0777\text{Hbdon3} + 0.0688\text{Hbacc3} - 0.00025\text{CSA} \quad (39)$$
$$N = 470, \quad R^2 = 0.928, \quad \text{SD} = 0.074, \quad F = 1200$$

$$\text{Sig2} = 8.438 - 6.004E + 28.365S + 38.687A + 37.034B + 3.040V \quad (40)$$
$$N = 470, \quad R^2 = 0.930, \quad \text{SD} = 6.941, \quad F = 1224$$

However, the most important finding was that the five experimental descriptors of Abraham and the five theoretical Klamt descriptors encode almost the same chemical information. This work, together with considerable evidence from experimental descriptors, from theoretical descriptors and from data analysis, suggests that only a small number of descriptors, probably no more than five, are needed to represent a very large number of physicochemical processes, including GLC and HPLC. These processes are all what may be referred to as 'transport' processes, in that they have as the only, or main, step, the transfer of a solute from one phase to another.

7. Conclusions

In general, processes in which a solute is transferred from one phase to another are selective, in that they respond to changes in solute properties, but they are not very specific. Thus 3- and 4-hydroxybenzyl alcohol have quite similar (but not identical) CHI values, as do 3-fluoro- and 4-fluorobenzoic acid [73]. The partition coefficients for transfer of conformational isomers between solvents or between the gas phase and solvents are almost the same, unless the conformational isomers have different internal hydrogen bonding or have different dipole moments. Thus for the transfer of the conformational isomers of fluoro-, chloro- and bromo-cyclohexane, there is almost no difference at all in partition coefficients between solvents [101,102], as pointed out by Eliel and Martin [102].

The number of types of solute–solvent interaction that control such general transfers is limited. It is therefore not surprising that all the investigations on the determination of sets of descriptors from chromatographic data have resulted in not more than five descriptors in each set. These include descriptors from GLC data by Laffort et al. and by Weckwerth et al., descriptors from RP-HPLC data by Snyder and Gilroy et al., and descriptors from GLC, RP-HPLC and other processes by Abraham et al. All these sets of descriptors, as well as the Famini and Wilson and the Klamt theoretical descriptors, encode similar chemical information, that is on London (dispersion) interactions, dipolarity/polarizability interactions, hydrogen bond acid and base interactions, and some descriptor of solute size. This information appears to be adequate to account for solute-solvent phase interactions in a very wide range of physicochemical processes, including GLC and HPLC. One exception appears to be a number of RP-HPLC systems where the actual shape of a solute is important, and some other descriptor is needed to account for this. However, to date there seems to be no 'shape' descriptor that is of general applicability.

No matter what is the exact nature of the solute descriptors used, there will always be a 'trade-off' between specificity and generality. The sets of descriptors of Laffort and of Weckwerth have been obtained from specific training sets of GLC stationary phases. It might therefore be expected that for the particular solutes in the training set, there will be very good fits of retention data for further GLC phases that are close to the ones used in the training set. As systems become less and less similar to the training set systems, so will the fits of data from these systems become less and less good. The problem is very severe with the solute descriptors of Snyder and Dolan obtained from a specific training set of RP-HPLC systems. It is already clear that for systems outside the training set, altered or amended descriptors have to be used. On the other hand, solute descriptors obtained from a wide variety of systems, such as the descriptors of Abraham, will be more general, in that they will apply to a much wider variety of other systems. However, the fits of data in these systems will not be as good as those obtained for 'specific' sets of solute descriptors when these specific sets are applied to systems near to their training sets. The two methods of obtaining sets of solute descriptors, the specific and general methods, are therefor complimentary.

Acknowledgements

We are grateful to the EPSRC for a postdoctoral fellowship (to A.M.Z.) and a postgraduate studentship (to A.I.). We thank our colleagues at GSK (Chris Bevan, Klara Valko and Chau Du), at Sirius Analytical Inc. (John Comer, Karl Box and Kin Tan) and at Reytek Ltd. (Derek Reynolds) for their help and advice. We are very grateful to William E. Acree Jr., for his collaboration on the determination of descriptors by solubility measurements, and to Colin Poole for his help and advice generally, and particularly on gas chromatography.

References

[1] J.H. Hildebrand, Chem. Rev. 44 (1949) 37.
[2] J.D. Crowley, G.S. Teague, J.W. Lowe, J. Paint Technol. 38 (1966) 269.
[3] R.C. Nelson, V.F. Figurelli, J.G. Walsham, G.D. Edwards, J. Paint Technol. 42 (1970) 644.
[4] C.M. Hansen, J. Paint Technol. 39 (1967) 104.
[5] C.M. Hansen, J. Paint Technol. 39 (1967) 505.
[6] C.M. Hansen, K. Skaarup, J. Paint Technol. 39 (1967) 511.
[7] C.M. Hansen, I&EC Prod. Res. Dev. 8 (1969) 2.
[8] A.M. Barton, Pure Appl. Chem. 57 (1985) 905.
[9] B.L. Karger, L.R. Snyder, C. Eon, J. Chromatogr. 125 (1976) 71.
[10] B.L. Karger, L.R. Snyder, C. Eon, Anal. Chem. 50 (1978) 2126.
[11] R.D. Cramer III, J. Am. Chem. Soc. 102 (1980) 1837.
[12] R.D. Cramer III, J. Am. Chem. Soc. 102 (1980) 1848.
[13] R.D. Cramer III, Quant. Struct.-Act. Relat. 2 (1983) 7.
[14] R.D. Cramer III, Quant. Struct.-Act. Relat. 2 (1983) 13.
[15] L. Rohrschneider, J. Chromatogr. 17 (1965) 1.
[16] L. Rohrschneider, J. Chromatogr. 22 (1966) 6.
[17] W.O. McReynolds, J. Chromatogr. Sci. 8 (1970) 685.
[18] P.H. Weiner, D.G. Howery, Can. J. Chem. 50 (1972) 448.
[19] P.H. Weiner, D.G. Howery, Anal. Chem. 44 (1972) 1189.
[20] P. Laffort, F. Patte, J. Chromatogr. 126 (1976) 625.
[21] F. Patte, M. Etcheto, P. Laffort, Anal. Chem. 54 (1982) 2239.
[22] W.O. McReynolds, personal communication to Laffort and Patte, see ref. [20].
[23] P. Laffort, F. Patte, J. Chromatogr. 406 (1987) 51.
[24] A. Voelkel, J. Janas, J. Chromatogr. 623 (1992) 83.
[25] J. Li, Y. Zhang, A.J. Dallas, P.W. Carr, J. Chromatogr. 550 (1991) 101.
[26] J. Li, A.J. Dallas, P.W. Carr, J. Chromatogr. 519 (1990) 103.
[27] J. Li, Y. Zhang, H. Ouyang, P.W. Carr, J. Am. Chem. Soc. 114 (1992) 9813.
[28] J. Li, Y. Zhang, P.W. Carr, Anal. Chem. 64 (1992) 210.
[29] J. Li, Y. Zhang, P.W. Carr, Anal. Chem. 65 (1993) 1969.
[30] J.D. Weckwerth, M.F. Vitha, P.W. Carr, Fluid Phase Equilib. 183–184 (2001) 143.
[31] Ref. 17 of Weckwerth et al [30] is to Li et al., our ref. [26].
[32] K. Dross, R.F. Rekker, G. de Vries, R. Mannhold, Quant. Struct.-Act. Relat. 18 (1999) 549.
[33] G. Mitchell, E.D. Clarke, S.M. Ridley, D.T. Greenhow, K.J. Killen, S.K. Vohra, P. Wardman, Pest. Sci. 44 (1995) 49.
[34] K. Valko, C.M. Du, C. Bevan, D.P. Reynolds, M.H. Abraham, Curr. Med. Chem. 8 (2001) 1137.
[35] C.M. Du, K. Valko, C. Bevan, D. Reynolds, M.H. Abraham, J. Liq. Chromatogr. Rel. Technol. 24 (2001) 635.
[36] F. Lombardo, M.Y. Shalaeva, K.A. Tupper, F. Gao, J. Med. Chem. 44 (2001) 2490.
[37] M. Roses, E. Bosch, Anal. Chim. Acta 274 (1993) 147;
E. Bosch, P. Bou, M. Roses, Anal. Chim. Acta 299 (1994) 219;
J.R. Torres-Lapasio, M.C. Garcia-Alvarez-Coque, M. Roses, E. Bosch, J. Chromatogr. A 955 (2002) 19.
[38] N.S. Wilson, M.D. Nelson, J. W Dolan, L.R. Snyder, R.G. Wolcott, P.W. Carr, J. Chromatogr. A 961 (2002) 171.
[39] N.S. Wilson, M.D. Nelson, J. W Dolan, L.R. Snyder, P.W. Carr, J. Chromatogr. A 961 (2002) 195.
[40] N.S. Wilson, J. W Dolan, L.R. Snyder, P.W. Carr, L.C. Sander, J. Chromatogr. A 961 (2002) 217.
[41] M.H. Abraham. Chem. Soc. Rev. (1993) 73.

[42] M.H. Abraham, P.L. Grellier, D.V. Prior, J.J. Morris, P.J. Taylor, J. Chem. Soc., Perkin Trans. 2 (1990) 521.
[43] J.J. Gilroy, J.W. Dolan, L.R. Snyder, J. Chromatogr. A 1000 (2003) 757.
[44] A. Wang, P.W. Carr, J. Chromatogr. A 968 (2002) 2.
[45] R.W. Taft, M.J. Kamlet, J. Am. Chem. Soc. 98 (1976) 377.
[46] R.W. Taft, M.J. Kamlet, J. Am. Chem. Soc. 98 (1976) 2886.
[47] T. Yokoyama, R.W. Taft, M.J. Kamlet, J. Am. Chem. Soc. 98 (1976) 3233.
[48] M.J. Kamlet, J.L. Abboud, R.W. Taft, J. Am. Chem. Soc. 99 (1977) 6027.
[49] M.J. Kamlet, J.-L.M. Abboud, M.H. Abraham, R.W. Taft, J. Org. Chem. 48 (1983) 2877.
[50] M.H. Abraham, P.L. Grellier, J.-L.M. Abboud, R.M. Doherty, R.W. Taft, Can. J. Chem. 66 (1988) 2673.
[51] R.W. Taft, M.H. Abraham, R.M. Doherty, M.J. Kamlet, J. Am. Chem. Soc. 107 (1985) 3105.
[52] M.J. Kamlet, R.M. Doherty, J.-L.M. Abboud, M.H. Abraham, R.W. Taft, Chemtech 16 (1986) 566.
[53] M.H. Abraham, P.L. Grellier, I. Hamerton, R.A. McGill, D.V. Prior, G.S. Whiting, Faraday Discuss. Chem. Soc. 85 (1988) 107.
[54] M.H. Abraham, J. Liszi, J. Chem. Soc., Faraday Trans. 174 (1978) 1604.
[55] R.A. Pieriotti, Chem. Rev. 76 (1976) 717.
[56] M.H. Abraham, G.S. Whiting, R.M. Doherty, W.J. Shuely, J. Chem. Soc., Perkin Trans. 2 (1990) 1451.
[57] ChemSketch, Advanced Chemistry Development Inc., Toronto, Canada.
[58] J.C. Dearden, S.J.A. Bradburne, M.H. Abraham, in: C. Silipo, A. Vittoria (Eds.), QSAR: Rational Approaches to the Design of Bioactive Compounds, Elsevier, Amsterdam, 1991.
[59] M.H. Abraham, P.L. Grellier, R.A. McGill, J. Chem. Soc., Perkin Trans. 2 (1987) 797.
[60] M.H. Abraham, G.S. Whiting, R.M. Doherty, W.J. Shuely, J. Chromatogr. 587 (1991) 213.
[61] M.H. Abraham, G.S. Whiting, J. Chromatogr. 594 (1992) 229.
[62] M.H. Abraham, J. Chromatogr. 644 (1993) 95.
[63] Q. Li, C.F. Poole, W. Kiridena, W.W. Koziol, Analyst 125 (2000) 2180.
[64] F. Mutelet, M. Rogalski, J. Chromatogr. A 923 (2001) 153.
[65] M.H. Abraham, P.L. Grellier, D.V. Prior, P.P. Duce, J.J. Morris, P.J. Taylor, J. Chem. Soc., Perkin Trans. 2 (1989) 699.
[66] M.H. Abraham, J. Phys. Org. Chem. 6 (1993) 659.
[67] D.E. Leahy, J.J. Morris, P.J. Taylor, A.R. Wait, J. Chem. Soc., Perkin Trans. 2 (1992) 705.
[68] The MedChem Data Base 2002, BioByte Corp. and Pomona College, Daylight Chemical Information Systems, 27401 Los Altos, CA.
[69] A.J. Leo, ClogP for Windows, Version 4.0, BioByte Corp., Claremont, CA, 2001.
[70] D.E. Leahy, J.J. Morris, P.J. Taylor, A.R. Wait, J. Chem. Soc., Perkin Trans. 2 (1992) 723.
[71] A.M. Zissimos, M.H. Abraham, M.C. Barker, K.J. Box, K.Y. Tam, J. Chem Soc., Perkin Trans. 2 (2002) 470.
[72] M. Plass, K. Valko, M.H. Abraham, J. Chromatogr. A 803 (1998) 51.
[73] A.M. Zissimos, M.H. Abraham, C.M. Du, K. Valko, C. Bevan, D. Reynolds, J. Wood, K.Y. Tam, J. Chem Soc., Perkin Trans. 2 (2002) 2001.
[74] K. Valko, S. Espinosa, C.M. Du, E. Bosch, M. Roses, C. Bevan, M.H. Abraham, J. Chromatogr. A 933 (2001) 73.
[75] C.M. Du, K. Valko, C. Bevan, D. Reynolds, M.H. Abraham, J. Chromatogr. Sci. 38 (2000) 503.
[76] K. Valko, M. Plass, C. Bevan, D. Reynolds, M.H. Abraham, J. Chromatogr. A 797 (1998) 41.
[77] M.H. Abraham, H.S. Chadha, G.S. Whiting, R.C. Mitchell, J. Pharm. Sci. 83 (1994) 1085.
[78] M.H. Abraham, J. Andonian-Haftvan, G.S. Whiting, A. Leo, R.W. Taft, J. Chem. Soc., Perkin Trans. 2 (1994) 1777.
[79] M.H. Abraham, C.E. Green, W.E. Acree Jr., C.E. Hernandez, L.E. Roy, J. Chem. Soc., Perkin Trans. 2 (1998) 2677.
[80] M.H. Abraham, C.E. Green, W.E. Acree Jr., J. Chem. Soc., Perkin Trans. 2 (2000) 281.
[81] K.M. De Fina, T.L. Sharp, M.A. Spurgin, I. Chuca, W.E. Acree Jr., C.E. Green, M.H. Abraham, Can. J. Chem. 78 (2000) 184.
[82] C.E. Green, M.H. Abraham, W.E. Acree Jr., K.M. De Fina, T.L. Sharp, Pest. Manag. Sci. 56 (2000) 1043.
[83] W.E. Acree Jr., M.H. Abraham, Fluid Phase Equlib. 201 (2002) 245.
[84] C.A. Monarrez, W.E. Acree Jr., M.H. Abraham, Phys. Chem. Liq. 40 (2002) 581.
[85] W.E. Acree Jr., M.H. Abraham, J. Sol. Chem. 31 (2002) 293.
[86] M.H. Abraham, N. Benjelloun-Dakhama, J.M.R. Gola, W.E. Acree Jr., W.S. Cain, J.E. Cometto-Muniz, N. J. Chem. 24 (2000) 825.
[87] M.H. Abraham, J. Environ. Monit. 5 (2003) 747.
[88] M.H. Abraham, J.M. Al-Hussaini, J. Environ. Monit. 4 (2002) 743.
[89] M.H. Abraham, J.M. Al-Hussaini, J. Environ. Monit. 3 (2001) 377.
[90] M.H. Abraham, A.M. Zissimos, W.E. Acree Jr., N. J. Chem. 27 (2003) 1041.
[91] S. Wolde, K. Andersson, J. Chromatogr. 80 (1973) 43.
[92] M. Chastrette, J. Chromatogr. Sci. 14 (1976) 357.
[93] K. Heberger, Chemometr. Int. Lab. Syst. 47 (1999) 41.
[94] B. Lucic, N. Trinajstic, S. Sild, M. Karelson, A.R. Katritzky, J. Chem. Inf. Comput. Sci. 39 (1999) 610.
[95] A.R. Katritzky, D.B. Tatham, J. Chem. Inf. Comput. Sci. 41 (2002) 358.
[96] M.H. Abraham, G.S. Whiting, P.W. Carr, H. Ouyang, J. Chem. Soc., Perkin Trans. 2 (1998) 1385.
[97] M.H. Abraham, G.S. Whiting, W.J. Schuely, R.M. Doherty, Can. J. Chem. 76 (1998) 703.
[98] G. Famini, C.A. Penski, L.Y. Wilson, J. Phys. Org. Chem. 3 (1992) 395.
[99] A. Klamt, V. Jonas, T. Burger, J.C.W. Lohrenz, J. Phys. Chem. A 102 (1998) 5074.
[100] A. Zissimos, M.H. Abraham, A. Klamt, F. Eckert, J. Wood, J. Chem. Inf. Comput. Sci. 42 (2002) 1320.
[101] J. Reisse, J.C. Celotti, A. Chiudoglu, Tetrahedron Lett. (1965) 397.
[102] E.L. Eliel, R.J.L. Martin, J. Am. Chem. Soc. 90 (1968) 689.