Correlation and Estimation of Aqueous Solubilities of Polycyclic Aromatic Hydrocarbons

Ronald J. Baker, William E. Acree, Jr., and Chun-che Tsai*

Department of Chemistry, Kent State University, Kent, Ohio 44242, U.S.A.

List of Abbreviations and Symbols

MP	Melting point temperature
MW	Molecular weight
N	Total number of atoms in the molecule (used to calculate molecular redundancy index)
n_i	Number of atoms in i th atom set (used to calculate molecular redundancy index)
PAH	Polycyclic aromatic hydrocarbon
R	Molecular redundancy index
r_d	Summation of interatomic distances in a molecule
r_{ij}	Distance between i th and j th atoms
ΔS_f	Entropy of fusion
TSA	Total molecular surface area
V_w	Molecular van der Waals volume
X	Mole fraction solubility

Abstract

Quantitative structure-activity relationships have been studied for the aqueous solubilities of 25 polycyclic hydrocarbons (PAH) at 25°C. Multiple regression correlations are reported between aqueous solubility (X) and molecular size descriptor (molecular surface area, molecular volume or molecular weight), melting point temperature (MP), and a newly-defined molecular symmetry and structure descriptor ($R \cdot r_d$). A generalized correlation takes the form of

$$\log X = a_0 + a_1 (\text{Molecular size descriptor}) + a_2 (\text{MP-25}) + a_3 R \cdot r_d$$

in which a_i refers to the i th regression coefficient, R is a molecular redundancy index and r_d is the summation of interatomic distances in a molecule. These correlations enable the aqueous solubility to be estimated from melting point temperatures and structural information. The 3-parameter expressions best predict the experimental solubilities and are superior to the 2-parameter expressions ($a_3 = 0$) suggested in the past. For the majority of compounds considered, agreement between observed and estimated solubility is quite good. The correlations may also be useful in screening experimental solubility data for which serious discrepancies exist in the literature.

1 Introduction

In recent years several workers have addressed the problem of correlating and predicting the solubilities of polycyclic aromatic hydrocarbons (PAH's) and other non-electrolytes in aqueous media [1–8]. The development of a model which provides a semi-quantitative means of predicting the solubility of a given PAH is of particular importance considering the well

documented carcinogenic activity of this class of compounds, and especially in the light of the frequent danger imposed by the pollution of natural waterways by petroleum hydrocarbons. Additionally, a broad spectrum of antitumor, antiviral and interferon inducing agents contain polycyclic aromatic chromophores as major constituents. As it is widely accepted that aqueous solubility is an important consideration in seeking to maximize drug efficacy, it is anticipated that these investigations might form the basis for fruitful research aimed at the development of a general theory to predict the aqueous solubilities and, ultimately, the therapeutic properties of more pharmacologically important classes of compounds.

Earlier correlations involving this family of chemical compounds attempted to utilize empirically derived quantities such as the partial molar volumes (9) and octanol/water partition coefficients [3, 4]. Recently, correlations have also been noted as a linear relationship between the logarithm of aqueous solubility and the number of atoms in the carbon skeleton [1]. On a more fundamental level, the ability to correlate and/or accurately predict the aqueous solubility of a given solute will require a more detailed understanding of the exact nature of solute-solvent interactions at the molecular level than these correlations would be expected to evince. Precise theoretical treatment of the collisional and electrostatic interactions extant in the solution phase which govern the aqueous solubility of a given compound, however, proves to be enormously tedious and far too complex to be feasible. A far simpler approach to this problem, which has already found widespread acceptance and application in the correlation and prediction of the properties of certain drug molecules, would be to examine the relationships between aqueous solubility and intrinsic molecular properties which are readily determined via computational or experimental processes, using multiregression analysis. In these initial investigations it, nonetheless, remains prudent to employ appropriate chemical intuition in the selection of suitable macroscopic and microscopic properties to be employed.

For non-polar solutes the major factor governing aqueous solubility is the amount of energy required to create a solvent cavity wherein a solute molecule or, possibly, an aggregation of solute molecules would reside. This quantity would reasonably vary according to the size of the molecule (i.e. the physical dimensions of the cavity which would be required) and has been historically considered to be linearly related to the molecular surface area of the solute. Several reports in recent years have demonstrated the efficacy of surface area as a molecular descriptor in correlating and predicting aqueous solubility [2–5, 7]. This technique suffers, however, from an inability to successfully distinguish between structural isomers (i.e. anthracene and phenanthrene, chrysene and triphenylene), which quite often have very similar calculated surface areas, but exhibit markedly dissimilar solubilities (cf. Table 1).

Another important factor influencing the solubility of solids at the molecular level is the energy required to overcome the crystalline lattice forces inherent in the solute. Employing

* To receive all correspondence.

Table 1. Parameters Used in the Derivation of Correlation Equations.

COMPOUND	$\log x^b$	MW	MP ⁱ (°C)	ΔS_f^m (cal/mol-K)	R ^o	TSA (nm ²)	V _w ³ (nm ³)	r _d (nm)	Structural Reference
BENZENE C ₆ H ₆	-3.395	78.12	5.5		0.7211	1.050	0.0800	17.69	[12]
NAPHTHALENE C ₁₀ H ₈	-5.342	128.19	80.6	12.98	0.4530	1.503	0.1233	52.39	[13]
FLUORENE C ₁₃ H ₁₀	-6.670 ^c	166.23	116.5	11.60	0.2115	1.935	0.1601	101.58	[14]
ANTHRACENE C ₁₄ H ₁₀	-8.103	178.24	216.3	13.99	0.3999	1.978	0.1663	115.49	[15]
PHENANTHRENE C ₁₄ H ₁₀	-6.907	178.24	101.0	10.78	0.2181	2.003	0.1705	113.33	[16]
PYRENE C ₁₆ H ₁₀	-7.878	202.26	156.0	9.64	0.3764	2.137	0.1855	135.84	[17]
FLUORANTHENE C ₁₆ H ₁₀	-7.643 ^d	202.26	107.8 ^j	11.84	0.1964	2.185	0.1872	139.32	[18]
CHRYSENE C ₁₈ H ₁₂	-9.849	228.30	255.5	11.76	0.2032	2.381	0.2088	204.11	[19]
TRIPHENYLENE C ₁₈ H ₁₂	-8.471	228.30	199.0	12.65	0.5268	2.397	0.2118	197.44	[20]
BENZ[<i>a</i>]ANTHRACENE C ₁₈ H ₁₂	-9.027	228.30	162.0	11.76	0.0000	2.424	0.2111	206.96	[21]
BENZ[<i>a</i>]PYRENE C ₂₀ H ₁₂	-9.549	252.32	177.0	9.11	0.0000	2.537	0.2269	236.84	[22]
PERYLENE C ₂₀ H ₁₂	-10.549	252.32	278.0	13.70	0.3750	2.564	0.2291	232.66	[23]
BENZ[<i>a,h</i>]PERYLENE C ₂₂ H ₁₂	-10.762	276.35	277.0 ^d	7.48	0.1966	2.693	0.2462	268.80	
DIBENZ[<i>a,h</i>]ANTHRACENE C ₂₂ H ₁₄	-10.489	278.36	269.5	13.69	0.1934	2.835	0.2509	337.72	[24]
CORONENE C ₂₄ H ₁₂	-11.068 ^d	300.36	439.0	6.45 ^d	0.6289	2.845	0.2625	309.58	[25]
ACENAPHTHENE ^a C ₁₂ H ₁₀	-6.244 ^e	154.21	96.2	14.04 ⁿ	0.2446	1.780	0.1460	85.09	[26]
BENZ[<i>a</i>]FLUORENE ^a C ₁₇ H ₁₂	-8.570 ^{e,f}	216.29	189.5		0.0000	2.370	0.2022	187.32	[27]
BENZ[<i>b</i>]FLUORENE ^a C ₁₇ H ₁₂	-8.839 ^{e,f}	216.29	208.5 ^k		0.0000	2.306	0.1977	188.15	
NAPHTHACENE ^a C ₁₈ H ₁₂	-10.432	228.30	357.0		0.3940	2.496	0.2147	216.80	[28]
BENZ[<i>e</i>]PYRENE ^a C ₂₀ H ₁₂	-9.466 ^g	252.32	176.5	8.72	0.2000	2.520	0.2259	229.85	
BENZ[<i>b</i>]FLUORANTHENE ^a C ₂₀ H ₁₂	-9.971 ^f	252.32	168.0		0.0000	2.602	0.2288	240.37	
BENZ[<i>j</i>]FLUORANTHENE ^a C ₂₀ H ₁₂	-9.749 ^f	252.32	166.0		0.0000	2.581	0.2287	237.27	
BENZ[<i>k</i>]FLUORANTHENE ^a C ₂₀ H ₁₂	-10.244 ^f	252.32	217.0		0.1875	2.626	0.2285	246.48	
INDENO[1,2,3- <i>cd</i>]PYRENE ^a C ₂₂ H ₁₂	-10.907 ^f	276.35	162.0 ^l	11.80 ^l	0.0000	2.745	0.2449	278.10	
DIBENZ[<i>a,j</i>]ANTHRACENE ^a C ₂₂ H ₁₄	-9.253 ^h	278.36	197.5		0.1719	2.848	0.2523	329.41	

a) Not employed in solubility correlations.

b) Logarithm of aqueous mole fraction solubility.

Taken, except as noted, from Reference [7].

c) Reference [3].

d) Reference [1].

e) Reference [4].

f) Reference [34].

g) Reference [35].

h) Reference [36].

i) Taken, except where noted, from Reference [37].

j) Reference [38].

k) Reference [39].

l) Reference [41].

m) Taken, except as noted, from Reference [38].

n) Reference [40].

o) Molecular Redundancy Index.

thermodynamic considerations a number of workers have accounted for this energy by incorporating an additional term, the so called triple point melting temperature, into the correlation expression [1, 3, 8]. At room temperature (25 °C) this augmented relationship can be expressed as

$$\log X = a_0 + a_1 (\text{Molecular Descriptor}) + a_2 (\Delta S_f)(\text{MP-25})$$

in which a_0 , a_1 and a_2 are constants, MP is the triple point melting temperature and ΔS_f is the entropy of fusion which, due to frequent inavailability, is normally assumed to be constant and incorporated into a_2 .

Sets $[\alpha, \alpha, \alpha, \alpha]$ $[\beta, \beta, \beta, \beta]$ $[\gamma, \gamma]$
 Multiplicity 4 4 2

Figure 1. Decomposition of the naphthalene molecule into sets of topologically equivalent atoms (C_α , C_β , C_γ , H_α , H_β) and their multiplicities.

At this writing no attempt to compare the applicability of molecular surface area with respect to other molecular descriptors has appeared in the published literature. Similarly, the apparently obvious relationships between solubility and molecular shape and molecular symmetry have yet to be investigated. In this report, therefore, we document efforts to establish a correlation between the aqueous solubility of the first 15 PAH compounds listed in Table 1 and a variety of molecular size descriptors, including molecular surface area, molecular volume and molecular weight. In addition the applicability of the melting points and entropy of fusion will be explored along with molecular topology and symmetry correction terms. Since these latter terms require knowledge of the conformational arrangement of individual atoms within a given molecule it was decided to restrict our initial investigations to the planar rigid alternant and non-alternant polycyclic aromatic hydrocarbons.

2 Materials and Methods

Mole fraction solubilities (X) at 25 °C, melting points (MP) and the entropies of fusion (ΔS_f) for the 25 compounds considered in this study were obtained from the literature. Total molecular surface areas (TSA) were computed using the program SAREA [10]. Van der Waals molecular volumes (V_w) were computed by the program MOLV [11], which uses an iterative summation of discrete increments within the confines of a given molecular unit. As it was desired to compare the viability of each descriptor, atomic coordinates used as input for TSA and V_w calculations were derived, whenever possible, from x-ray or neutron diffraction studies [12–28]. In the case of benzo[ghi]perylene and several of the molecules not included in the original regression analysis, for which this information was not available, suitable atomic coordinates were determined on the basis of analogous molecular structures employing chemically reasonable interatomic bond distances and angles. The van der Waals radii employed in both surface area and volume calculations were 1.70 Å for aromatic carbon and 1.20 Å for hydrogen.

In this study, we define a molecular redundancy index (R), derived from information theory and molecular graph theory [29–32], as a molecular symmetry descriptor. The redundancy index, a quantity indicating the symmetry and complexity of a molecule, can be computed from the following formula:

$$R = \frac{\sum_i n_i \log n_i}{N \log N}$$

where n_i is the number of atoms in the i th atom set, i is the number of different atom sets, and $N = \sum_i n_i$ is the

total number of atoms in the molecule. Considering as an example the naphthalene molecule, $C_{10}H_8$ (Figure 1), the summation extends over three types of non-equivalent carbon atoms (C_α , C_β , C_γ) and two types of non-equivalent hydrogen (H_α , H_β) atoms. The resultant calculation becomes:

$$R = \frac{n_{C_\alpha} \log n_{C_\alpha} + n_{C_\beta} \log n_{C_\beta} + n_{C_\gamma} \log n_{C_\gamma} + n_{H_\alpha} \log n_{H_\alpha} + n_{H_\beta} \log n_{H_\beta}}{(n_{C_\alpha} + n_{C_\beta} + n_{C_\gamma} + n_{H_\alpha} + n_{H_\beta}) \log(n_{C_\alpha} + n_{C_\beta} + n_{C_\gamma} + n_{H_\alpha} + n_{H_\beta})}$$

$$= (4 \log 4 + 4 \log 4 + 2 \log 2 + 4 \log 4 + 4 \log 4) / (18 \log 18) = 0.4530.$$

We also introduce a novel molecular size descriptor, r_d , which is defined for a molecule composed of j atoms as:

$$r_d = \sum_i \sum_{j>i} r_{ij}$$

in which r_{ij} is defined for a given solute molecule as the interatomic distance between any two atoms in the structure. Such a descriptor was found to be particularly useful in distinguishing structural isomers.

Following the computation and tabulation of the molecular and thermodynamic input data multiple regression analysis was performed using the program HANSCH3X [33]. With this program it is possible to systematically select, either individually, or by collection from as many as thirty dependent variables to be correlated against a chosen independent variable. In this study the logarithm of mole fraction aqueous solubility ($\log X$) was examined in relation to numerous calculated molecular size descriptors, including the molecular weight (MW), TSA, V_w , and r_d . Additional terms considered included the thermodynamic (ΔS_f , MP) and symmetry (R) correction terms along with various geometric combinations of the symmetry and molecular size descriptors. Regression equations were limited to contain no more than four terms and all possible linear combinations of the input data were evaluated.

3 Results and Discussions

Correlations exhibiting significant statistical merit are presented in Table 2. These equations represent the most favorable correlations obtained from the approximately 10000 permutations of input variables evaluated. The criteria employed in the decision to report a given correlation expression were: first that equations reported were required to exhibit favorable standard deviations (s) and correlation coefficients and second that the 95 percent confidence intervals of the individual regression coefficients could not be inordinately large with respect to the coefficients themselves. In many instances the regression statistics could be favorably affected by augmentation of the data set to include more compounds. However, in the interests of evaluating the effectiveness of the molecular and

Table 2. Correlation Equations (Equations (1)–(30)).

Eqn. #	MW $\times 10^2$	TSA	V _W	MT $\times 10^3$	$\Delta S_F \cdot MT$ $\times 10$	R · MW $\times 10^3$	R · TSA $\times 10$	R · V _W	R · r _d $\times 10^2$	const.	s	r ²	F (P, Q)
(1)	–3.55 (47)									–0.86 (1.04)	0.487	0.953	265 (1,13)
(2)	–3.04 (90)			–3.45 (5.18)						–1.38 (1.27)	0.468	0.960	145 (2,12)
(3)	–2.94 (52)				–4.79 (3.04)					–1.29 (82)	0.360	0.976	249 (2,12)
(4)	–2.29 (86)			–10.0 (6.0)		10.3 (7.1)				–2.51 (1.24)	0.352	0.980	174 (3,11)
(5)	–2.23 (92)			–9.72 (6.14)			9.85 (7.24)			–2.69 (1.39)	0.363	0.978	164 (3,11)
(6)	–2.28 (88)			–9.78 (6.06)				11.1 (8.0)		–2.56 (1.30)	0.359	0.978	167 (3,11)
(7)	–2.50 (70)			–10.9 (5.6)					1.23 (71)	–1.90 (93)	0.321	0.982	210 (3,11)
(8)	–3.05 (1.11)			–11.1 (5.4)		–21.7 (34.3)			3.59 (3.78)	–0.52 (2.36)	0.307	0.986	172 (4,10)
(9)	–3.15 (1.15)			–11.2 (5.3)			–19.0 (27.2)		3.29 (3.02)	–0.25 (2.53)	0.302	0.986	179 (4,10)
(10)	–3.13 (1.08)			–11.3 (5.3)				–24.9 (33.6)	3.67 (3.35)	–0.28 (2.36)	0.298	0.986	183 (4,10)
(11)		–4.30 (61)								1.21 (1.38)	0.517	0.947	234 (1,13)
(12)		–3.39 (90)		–5.22 (4.26)						0.05 (1.49)	0.426	0.966	176 (2,12)
(13)		–3.59 (76)			–4.52 (3.65)					0.43 (1.31)	0.425	0.968	177 (2,12)
(14)		–2.62 (89)		–10.9 (5.2)		9.35 (6.71)				–1.32 (1.52)	0.327	0.982	202 (3,11)
(15)		–2.57 (95)		–10.5 (5.3)			8.96 (6.85)			–1.50 (1.68)	0.336	0.982	191 (3,11)
(16)		–2.61 (91)		–10.6 (5.2)				10.1 (7.5)		–1.37 (1.58)	0.332	0.982	196 (3,11)
(17)		–2.82 (68)		–11.9 (4.7)					1.16 (63)	–0.70 (1.08)	0.284	0.986	270 (3,11)
(18)		–3.89 (95)		–12.7 (3.6)			–27.6 (20.3)		4.12 (2.23)	2.17 (2.27)	0.215	0.994	355 (4,10)
(19)		–3.91 (88)		–12.6 (3.4)		–37.0 (24.6)			5.13 (2.68)	2.15 (2.05)	0.204	0.994	393 (4,10)
(20)		–3.89 (85)		–13.0 (3.4)				–37.0 (23.9)	4.75 (2.37)	2.19 (2.02)	0.201	0.994	406 (4,10)
(21)			–42.8 (5.8)							–0.05 (1.15)	0.494	0.953	258 (1,13)
(22)			–35.1 (9.8)	–4.33 (4.67)						–0.82 (1.34)	0.444	0.964	161 (2,12)
(23)			–35.5 (6.6)		–4.72 (3.20)					–0.63 (97)	0.377	0.974	226 (2,12)
(24)			–26.8 (9.1)	–10.5 (5.3)		9.99 (6.58)				–2.05 (1.28)	0.327	0.982	202 (3,11)
(25)			–26.2 (9.8)	–10.1 (5.4)			9.57 (6.74)			–2.23 (1.43)	0.338	0.982	189 (3,11)
(26)			–26.7 (9.4)	–10.2 (5.4)				10.8 (7.4)		–2.10 (1.34)	0.333	0.982	195 (3,11)
(27)			–29.1 (7.1)	–11.5 (4.8)					1.22 (63)	–1.43 (92)	0.286	0.986	266 (3,11)
(28)			–38.2 (9.7)	–12.3 (4.0)				–31.0 (26.6)	4.25 (2.66)	0.73 (2.00)	0.232	0.992	305 (4,10)
(29)			–38.6 (10.5)	–12.1 (4.0)			–24.0 (21.8)		3.82 (2.41)	0.80 (2.17)	0.237	0.992	292 (4,10)
(30)			–37.9 (10.2)	–12.0 (4.1)			–29.6 (27.6)		4.43 (3.03)	0.59 (2.04)	0.239	0.992	286 (4,10)

MT = MP – 25

F(P, Q) = F statistic with P and Q degrees of freedom.

Table 3. Comparison Between the Calculated and Experimental Solubilities.

Compound	LogX experimental – logX calculated from correlation equation									
	3	4	7	12	13	17	18	23	27	28
Benzene	0.20	0.32	0.30	0.12	–0.06	0.12	0.09	0.07	0.21	0.18
Naphthalene	0.07	0.06	0.08	–0.01	–0.06	–0.02	–0.05	0.00	0.02	0.04
Fluorene	0.02	0.20	0.12	0.32	0.32	0.32	0.10	0.14	0.20	–0.03
Anthracene	–0.29	–0.33	–0.24	–0.45	–0.23	–0.09	0.15	–0.31	–0.21	–0.04
Phenanthrene	0.02	0.04	–0.03	0.23	0.22	0.06	–0.12	0.16	0.05	–0.09
Pyrene	–0.03	–0.21	–0.13	0.00	–0.07	–0.19	0.06	–0.07	–0.17	0.07
Fluoranthene	0.07	–0.09	–0.12	0.15	0.21	–0.12	–0.19	0.09	–0.15	–0.23
Chrysene	–0.54	–0.28	–0.25	–0.62	–0.51	–0.17	–0.18	–0.53	–0.21	–0.22
Triphenylene	0.59	–0.23	–0.25	0.52	0.69	–0.15	0.11	0.71	–0.15	0.06
Benz[a]anthracene	–0.25	0.08	0.07	–0.14	–0.03	0.14	–0.02	–0.15	0.11	–0.01
Benzo[a]pyrene	–0.17	0.26	0.31	–0.20	–0.25	0.11	0.09	–0.22	0.22	0.25
Perylene	–0.18	–0.70	–0.67	–0.59	–0.22	–0.62	–0.45	–0.16	–0.62	–0.47
Benzo[ghi]perylene	–0.44	0.04	0.13	–0.37	–0.68	–0.08	0.05	–0.51	0.07	0.26
Dibenz[a,h]anthracene	0.59	0.29	0.22	0.35	0.76	0.36	0.32	0.62	0.25	0.09
Coronene	0.34	0.53	0.44	0.69	–0.09	0.33	0.04	0.14	0.37	0.15
Acenaphthene	0.06	0.12	0.03	0.11	0.16	0.08	–0.23	0.04	0.01	–0.33
Benzo[a]fluorene		0.54	0.53	0.28		0.77	0.58		0.63	0.44
Benzo[b]fluorene		0.46	0.46	–0.11		0.55	0.31		0.45	0.23
Naphthacene		–0.29	–0.27	–0.29		0.27	0.55		0.01	0.10
Benzo[e]pyrene	–0.12	–0.18	–0.18	–0.18	–0.26	–0.39	–0.40	–0.20	–0.29	–0.26
Benzo[b]fluoranthene		–0.26	–0.21	–0.45		–0.24	–0.19		–0.25	–0.21
Benzo[j]fluoranthene		–0.05	–0.01	–0.31		–0.10	–0.08		–0.05	–0.06
Benzo[k]fluoranthene		–0.52	–0.52	–0.39		–0.39	–0.29		–0.53	–0.53
Indeno[1,2,3-cd]pyrene	–0.71	–0.70	–0.61	–0.93	–0.76	–0.84	–0.65	–0.82	–0.78	–0.60
Dibenz[a,j]anthracene		0.87	0.78	1.26		0.87	0.88		0.80	0.71

thermodynamic descriptors as predictive tools we elected to determine regression coefficients based upon the first 15 compounds appearing in Table 1. These linear and multilinear correlations were in turn utilized to predict the solubilities of the 10 remaining compounds. Comparisons between the experimental and calculated values are listed in Table 3 for ten of the correlation equations. For the majority of compounds agreement between experimental and predicted solubility is quite good; in all instances well within one order of magnitude and, in most instances, differing by no more than a factor of two or three.

An interesting discovery which was not originally anticipated is the surprisingly similar results obtained for the three molecular size descriptors, MW, TSA and V_w . This is particularly significant in light of the considerable effort and computational time involved in obtaining the latter two quantities for a given compound. Moreover, considering the inherent inaccuracies inevitably incurred by the use of the somewhat arbitrary van der Waals radii and the spherical atom assumption employed in the calculations, molecular weight provides a far more precise size descriptor than do the calculated surface areas or molecular volumes. Therefore, for this classification of compounds, Equation (1) provides a good first approximation for estimating the aqueous solubility at 25°C

$$\log X = -0.86 - 0.0355 (\text{MW}) \quad (1)$$

As mentioned previously, crystal lattice forces do play an important role in determining the solubility of solids. This is particularly true in the case of high melting substances. Through basic thermodynamics, it can be shown that an additional term involving the solute's entropy of fusion and melting point temperature is required to properly describe the solubility of crystalline compounds. Introduction of the ΔS_f (MP-25)

term into Equation (1) does provide a better mathematical representation of the experimental data

$$\log X = -1.29 - 0.0294 (\text{MW}) - 0.000479 \Delta S_f \cdot (\text{MP-25}) \quad (3)$$

in comparing the statistical fits produced in Equation (3) and Equation (2), as demonstrated by a reduced standard deviation (0.360 compared to 0.487). Expressions utilizing TSA (Equation (13)) and V_w (Equation (23)) as molecular size descriptors give very similar results as shown in Figure 2. This is as expected

Figure 2. Comparison of observed and predicted logX values.

Figure 3. Linear correlations between molecular volume, molecular surface area and molecular weight.

since, as depicted in Figure 3 the molecular weight, total surface area and molecular volume can be shown to be linearly related to each other.

$$\log X = 0.43 - 3.59 (\text{TSA}) - 0.000452 \Delta S_f \cdot (\text{MP-25}) \quad (13)$$

$$\log X = -0.63 - 35.5 (V_w) - 0.000472 \Delta S_f \cdot (\text{MP-25}) \quad (23)$$

It is readily demonstrated that the correlations presented here between the three major molecular size descriptors (MW, TSA and V_w) apply to members of this classification of compounds other than the 15 employed in the original regression analysis. In Table 4 we compare the computed values of two of these descriptors (TSA and V_w) with the values obtained from the linear molecular weight correlations. These comparisons further confirm that molecular surface area and van der Waals volume for the PAH's are estimable as linear functions of their molecular weight. Ultimately, it is conceivable that the molecular weight may be used in this straightforward manner to provide a reliable and facile means of estimating both the quantities of molecular surface area and molecular volume. The molecular weight correlations, however, do not allow one to calculate molecular surface area (or molecular volume) differences between structural PAH isomers.

Table 5. Squared Correlation Coefficients.

	MW	TSA	V_w	MT	ΔS_f	R	r_d	$\Delta S_f \cdot \text{MT}$	$R \cdot \text{MW}$	$R \cdot \text{TSA}$	$R \cdot V_w$	$R \cdot r_d$
MW	1.00											
TSA	0.99	1.00										
V_w	1.00	1.00	1.00									
MT	0.74	0.68	0.71	1.00								
ΔS_f	0.07	0.12	0.09	0.02	1.00							
R	0.12	0.17	0.15	0.00	0.19	1.00						
r_d	0.94	0.92	0.93	0.72	0.04	0.10	1.00					
$\Delta S_f \cdot \text{MT}$	0.55	0.57	0.55	0.69	0.25	0.01	0.58	1.00				
$R \cdot \text{MW}$	0.09	0.06	0.07	0.39	0.02	0.56	0.07	0.17	1.00			
$R \cdot \text{TSA}$	0.03	0.01	0.02	0.28	0.03	0.69	0.03	0.12	0.98	1.00		
$R \cdot V_w$	0.06	0.04	0.05	0.34	0.02	0.62	0.05	0.14	1.00	0.99	1.00	
$R \cdot r_d$	0.31	0.26	0.29	0.64	0.00	0.29	0.29	0.33	0.91	0.82	0.88	1.00

Table 4. Estimation of Total Surface Areas and van der Waals Volumes.

Compound	Calculated From SAREA Program TSA	MOLV Program V_w	Estimated From Molecular Weight Correlations	
			TSA ^{a)}	V_w ^{b)}
Acenaphthene	1.780	0.1460	1.759	0.1469
Benzo[a]fluorene	2.370	0.2022	2.267	0.1983
Benzo[b]fluorene	2.306	0.1977	2.267	0.1983
Naphthacene	2.496	0.2147	2.365	0.2083
Benzo[e]pyrene	2.520	0.2259	2.562	0.2282
Benzo[b]fluoranthene	2.602	0.2288	2.562	0.2282
Benzo[j]fluoranthene	2.581	0.2287	2.562	0.2282
Benzo[k]fluoranthene	2.626	0.2285	2.562	0.2282
Indeno[1,2,3-cd]pyrene	2.745	0.2449	2.759	0.2481
Dibenz[a,j]anthracene	2.848	0.2523	2.775	0.2498

$$a) \text{ TSA} = 0.4975 + 0.008182 (\text{MW}) \quad S = 0.0586 \quad r^2 = 0.986$$

$$(0.1245) \quad (0.000567)$$

$$F(1,13) = 972.9$$

$$b) \text{ V}_w = 0.01912 + 0.0008286 (\text{MW}) \quad S = 0.00281 \quad r^2 = 0.999$$

$$(0.00598) \quad (0.0000272)$$

$$F(1,13) = 4317.4$$

It should be noted that Equations (12) and (13) are identical in concept to expressions proposed by Yalkowsky and co-workers [2]–[4].

$$\log X = 0.05 - 3.39 (\text{TSA}) - 0.00522 (\text{MP-25}) \quad (12)$$

Surprisingly, these two equations do exhibit comparable standard deviations (0.426 and 0.425 respectively), despite the fact that the entropies of fusion of the 15 PAH compounds considered here are not constant, as has been popularly assumed, but, rather, range from 6.45 to 13.99 cal/mol-K (cf. Table 1).

Improved correlations can be obtained by combining the symmetry and molecular size descriptors. Inspection of Table 2 reveals that Equation (17) represents one of the best 3-parameter expressions (Equations (7), (17), and (27))

$$\log X = -0.70 - 2.82 (\text{TSA}) - 0.0119 (\text{MP-25})$$

$$+ 0.0116 (R \cdot r_d) \quad (17)$$

and for 4-parameter expressions Equation (18) represents one of the three best correlations (Equations (18)–(20))

$$\log X = 2.17 - 3.89 (\text{TSA}) - 0.0127 (\text{MP-25}) - 2.76 (R \cdot \text{TSA})$$

$$+ 0.0412 (R \cdot r_d) \quad (18)$$

Both Equations (17) and (18) are superior to the 2-parameter expressions of Yalkowsky and co-workers. Furthermore, we have been able to replace the ΔS_f (MP-25) term with the simpler (MP-25) term, combined with a molecular topological descriptor term, which is readily calculable and does not rely upon the availability of experimentally determined measurements. Although the experimental triple point melting temperature is still required in the correlations presented here, our preliminary QSAR calculations indicate that this quantity can itself be predicted solely on the basis of structural information [42]. Dependent upon application, if one is willing to sacrifice a small degree of predictive accuracy for greater calculational simplicity, then Equation (4)

$$\log X = -2.50 - 0.0229(\text{MW}) - 0.0100(\text{MP-25}) + 0.0103(\text{R} \cdot \text{MW}) \quad (4)$$

may be the logical choice. As one might surmise, selection of a predictive expression does depend, to a large extent, on how accurately one wishes to estimate the solubility and how much effort one is willing to spend.

4 Acknowledgements

The authors wish to thank Dr. Ferenc Darvas for helpful discussions and comments. The support from National Institutes of Health (Grant No. GM-31257) is gratefully acknowledged.

5 References

- [1] D. Mackay and W. Y. Shiu, *J. Chem. Eng. Data* 22, 399 (1977).
- [2] S. C. Valvani, S. H. Yalkowsky, and G. L. Amidon, *J. Phys. Chem.* 80, 829 (1976).
- [3] S. H. Yalkowsky and S. C. Valvani, *J. Chem. Eng. Data* 24, 127 (1979).
- [4] S. H. Yalkowsky and S. C. Valvani, *J. Pharm. Sci.* 69, 912 (1980).
- [5] G. L. Amidon and S. T. Anik, *J. Chem. Eng. Data* 26, 28 (1981).
- [6] R. P. Eganhouse and J. P. Calder, *Geochim. Cosmochim. Acta* 40, 555 (1976).
- [7] S. T. Anik, Ph. D. Thesis, University of Wisconsin, Madison (1978).
- [8] D. Mackay, R. Mascarenhaus, W. Y. Shiu, S. C. Valvani, and S. H. Yalkowsky, *Chemosphere* 9, 257 (1980).
- [9] C. McAuliffe, *J. Phys. Chem.* 70, 1267 (1966).
- [10] R. S. Pearlmann, *QCPE Program No. 413*, Indiana University.
- [11] R. J. Baker and C.-c. Tsai, *unpublished*.

- [12] E. G. Cox, D. W. J. Cruickshank, and J. A. S. Smith, *Proc. Roy. Soc. (London) Ser. A* 247, 1 (1958).
- [13] V. I. Ponomarev, O. S. Filipenko, and L. O. Otovmayan, *Sov. Phys. Crystallogr.* 21, 215 (1976).
- [14] D. M. Burns and J. Iball, *Proc. Roy. Soc. (London)* 227, 200 (1955).
- [15] R. Mason, *Acta Cryst.* 17, 547 (1964).
- [16] M. I. Kay, Y. Okaya, and D. E. Cox, *Acta Cryst. B* 27, 26 (1971).
- [17] A. C. Hazell, F. K. Larsen, and M. S. Lehmann, *Acta Cryst. B* 28, 2977 (1972).
- [18] A. C. Hazell, D. W. Jones, and J. M. Snowden, *Acta Cryst. B* 33, 1516 (1977).
- [19] D. M. Burns and J. Iball, *Proc. Roy. Soc. (London)* 257, 491 (1960).
- [20] G. Ferraris, D. W. Jones, and J. Yerkess, *Z. Kristallogr.* 138, 113 (1973).
- [21] R. Foster, J. Iball, S. N. Scrimgeour, and B. C. Williams, *J. Chem. Soc. Perkin II*, 682 (1976).
- [22] J. Iball, S. N. Scrimgeour, and D. W. Young, *Acta Cryst. B* 32, 328 (1976).
- [23] A. Camerman and J. Trotter, *Proc. Roy. Soc. (London) Ser. A*, 279, 129 (1964).
- [24] J. Iball, C. H. Morgan, and D. E. Zacharias, *J. Chem. Soc. Perkin II*, 1271 (1975).
- [25] J. M. Robertson and J. G. White, *J. Chem. Soc.*, 607 (1945).
- [26] H. W. W. Ehrlich, *Acta Cryst.* 10, 699 (1957).
- [27] R. Bohme and E. Wilhelm, *Cryst. Struct. Commun.* 9, 933 (1980).
- [28] J. M. Robertson, V. C. Sinclair, and J. Trotter, *Acta Cryst.* 14, 697 (1961).
- [29] C. A. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL (1949).
- [30] L. Brillouin, *Science and Information Theory*, Academic Press, New York 1962.
- [31] N. Rashevsky, *Bull. Math. Biophys.* 17, 229 (1955).
- [32] L. B. Kier, *J. Pharm. Sci.* 69, 807 (1980).
- [33] C. Hansch, A. Leo, D. A. Elkins, and N. D. Dreyer, *private communication*.
- [34] S. A. Wise, W. J. Bonnett, F. R. Guenther, and W. E. May, *J. Chromatogr. Sci.* 19, 457 (1981).
- [35] F. P. Schwarz, *J. Chem. Eng. Data* 22, 273 (1977).
- [36] R. Krashnoshchekova and M. Gubergrits, *Org. React. (Tartu)* 13, 432 (1976).
- [37] R. C. Weast (Ed.), *CRC Handbook of Chemistry and Physics*, 63rd edition, CRC Press Inc., Boca Raton, FL. (1982).
- [38] F. Casselato, C. Vecchi, and A. Girelli, *Thermochim. Acta* 6, 361 (1973).
- [39] J. R. Pollock and R. Stevens (Eds.), *Dictionary of Organic Compounds*, 4th ed., 1, 342 (1965).
- [40] S. C. Mraw and D. F. Orourke, *J. Chem. Thermodynamics* 13, 199 (1981).
- [41] G. W. Smith, *Mol. Cryst., Liq. Cryst.* 64, 15 (1980).
- [42] C.-c. Tsai, R. J. Baker, and W. E. Acree, Jr., manuscript in preparation.

(Received on August 16th 1983, accepted on October 6th 1983)