Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure

Paul D. T. Huibers[†] and Alan R. Katritzky*[,‡]

Department of Chemical Engineering, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139-4307, and Department of Chemistry, University of Florida,
Gainesville, Florida 32611-7200

Received June 20, 1997[⊗]

The aqueous solubilities of a set of 109 hydrocarbons and 132 halogenated hydrocarbons (total 241) are correlated by a three term equation using descriptors calculated solely from molecular structure, with a correlation coefficient (R) of 0.979 and a standard error (s) of 0.386 log units. This equation allows the estimation of aqueous solubilities of hydrocarbons and halogenated hydrocarbons (including polychlorinated biphenyls). The key descriptor is the molecular volume, modified by topological and electrostatic terms. The use of descriptors calculated only from molecular structure eliminates the need for experimental determination of properties for use in the correlation and allows for the estimation of aqueous solubility for molecules not yet synthesized or isolated.

INTRODUCTION

The aqueous solubility of organic compounds is an important molecular property, playing a large role in the behavior of compounds in many areas of interest. In modeling the environmental impact of a contaminant, along with the soil-water absorption coefficient, the solubility is a key term in the understanding of transport mechanisms and distribution in groundwater. Given the importance of solubility, a means of prediction based solely on molecular structure should prove a useful tool, as many compounds exist for which the solubility simply is not available. Whereas a general equation would be of the greatest use, the present study is limited to hydrocarbons and halogenated hydrocarbons which were expected to be advantageous in obtaining a significant correlation, as the elimination of compounds that will undergo specific interactions with water, such as hydrogen bonding, simplifies the nature of the interactions that must be accounted for. Our study still covers a diverse and important group of possible compounds of great significance for better understanding the environmental impact, because these types of compounds are often the most long-lived of environmental contaminants due to their comparatively low level of biodegradability when compared to oxygen or nitrogen containing compounds.

Many different approaches to the prediction of aqueous solubility can be found in the literature, as summarized by Yalkowsky and Banerjee.[1] These approaches can be categorized as follows: (i) correlations with experimentally determined physicochemical quantities such as partition coefficient, chromatographic retention time, melting point, boiling point, molar volume (derived from liquid density), or parachor (derived from density and surface tension), but these require a sufficient quantity of the purified compound to be available and are not applicable for compounds not yet synthesized or isolated; (ii) group contributions derived from measured aqueous solubility, but these fail to account for the presence of neighboring groups or conformational influences; and (iii) parameters calculated only from molecular structure, such as molecular surface area, molecular volume, and topological indices, which is the most general approach.

Several related but distinct parameters have been commonly used to measure aqueous solubility. Solubilities of gases and vapors are expressed in terms of the dimensionless Ostwald solubility coefficient (L_w), defined as the ratio between the concentration of the solute in solution and the concentration of the solute in the gas phase,[2] or the Henry's Law constant, essentially an air/water partition coefficient, which has units of pressure.[3] Solubilities of liquids and solids are described by the solubility (S_w), defined as the concentration (in units of moles or weight of solute per weight or volume of solution) of solute in the aqueous phase, at equilibrium with a pure solute phase.

Aqueous solubilities (S_w) for organic compounds have been predicted using models that fall into the three approaches mentioned above, and we now summarize some of the most important. There are several examples of the use of other physical property measurements in predictions of type i. Yalkowsky et al. used melting point (mp) and either molecular surface area (MSA) or octanol/water partition coefficient for predicting S_w for polycyclic aromatic hydrocarbons[4] and halobenzenes.[5] These authors suggested that the melting point term accounted for the lattice energy in breaking apart a solid so that it could be solvated. Dunnivant et al.[6] used mp, MSA, and a topological descriptor for predicting S_w for polychlorinated biphenyls (PCBs). Amidon et al.[7] used MSA for predicting S_w for a wide range of structures. Kamlet et al.[8] used a linear solvation energy relationship (LSER) approach to predict S_w for aliphatic and aromatic hydrocarbons. A group contribution approach of

[†] Massachusetts Institute of Technology.
[‡] University of Florida.
[⊗] Abstract published in *Advance ACS Abstracts,* December 15, 1997.

S0095-2338(97)00043-7 CCC: $15.00
Published on Web 03/23/1998

Table 1. Examples of Previous QSPR Studies for Prediction of Aqueous Solubility[a]

N												
HC	Ar	XHC	PCB	OHC	outlier		R^2	*s*	no. of descriptors	correlation type	descriptor class	ref
			22			22	0.984	0.165	3	i	GP	6
8		16		91		115	0.992	0.126	3	i	GP	8
	33	19		18		70	0.983	0.216	3	i	GP	8
	32					32	0.976	0.271	2	i	GP	4
	32					32	0.979	0.251	2	i	P	4
		35				35	0.994	0.116	2	i	GP	5
9	14	20	9	14		66			U	ii	U	9
18						18	0.977	0.165		ii	G	7
					0	331	0.965	0.299	18	iii	CTGEQ	10
69	21	75	2	91	42	258	0.937	0.374	9	iii	TGE	11
69	22				32	91	0.978	0.168	9	iii	CTGE	11
		76	3		1	79	0.975	0.180	9	iii	CTGE	11
				92	5	92	0.975	0.167	9	iii	CTGE	11
58	10	22	15	18	4	123	0.980	0.277	9	iii	CTGE	12
58	10	22		18	19	107	0.974	0.201	9	iii	CTGE	12
	17	78			0	95	0.952	0.347	3	iii	CT	13
	17	78		50	0	145	0.926	0.318	5	iii	CT	13
66	45	104	35		9	241	0.959	0.386	3	iii	TGE	this study

[a] Structure subclasses are aliphatic hydrocarbons (HC), aromatic hydrocarbons (Ar), halogenated hydrocarbons (XHC), polychlorinated biphenyls (PCB), and oxygen and nitrogen containing compounds (OHC). Descriptor types used include constitutional (C), topological (T), geometrical (G), electrostatic (E), quantum-chemical (Q), and physical property based (P). The UNIFAC approach is denoted (U).

type ii using the UNIFAC method has been developed by Kan and Tomson[9] for predicting both S_w and solubilities in organic solvents. With regard to approach iii, S_w predictions have been made by several groups, using descriptors calculated only from molecular structure. Bodor and Huang[10] investigated a large 331 molecule set containing halogenated and oxygenated hydrocarbons, and used 18 descriptors of various types to arrive at a low standard error of 0.299. Nelson and Jurs (11) investigated an equally diverse set of 238 molecules, and with 9 descriptors achieved very low standard errors for sets of molecules broken down into classes by composition (hydrocarbons, halogenated hydrocarbons, ethers, and alcohols). Sutter and Jurs[12] also used 9 descriptors to achieve a standard error of 0.277 for a diverse set of 123 molecules. Nirmalakhandan and Speece[13] developed an approach using molecular connectivity indices combined with a modified polarizability, where the polarizability term was calculated by a group contribution method, different for each set of compounds studied. These studies are summarized in Table 1. It should be noted that no effort is made in any of these studies to estimate the experimental error of the solubility values used, as no multiple values were considered. This is important, as it makes no sense to create a correlation with a smaller standard error than the experimental error.

Our group previously correlated L_w, the Ostwald solubility coefficient, for gas solubilities of 95 diverse hydrocarbons in water with two parameters (eq 1), where G_I is the

$$-\log L_w = -(1.37 \pm 0.06) + (0.0067 \pm 0.0001)G_I - (0.050 \pm 0.001)^0\mathrm{CIC} \quad (1)$$

$$R^2 = 0.9765, \quad F = 1988, \quad s = 0.45, \quad N = 95$$

gravitation index and ^{0}CIC is the complementary information content.[2] For the aqueous solubilities (L_w) of the vapors of a group of 406 diverse organic compounds (including structures containing N, O, S, and halogen atoms), we derived the five parameter equation (eq 2), where HDCA2 is the partial charge weighted normalized hydrogen bonding donor surface area, N_O and N_N are counts of oxygen and nitrogen atoms, $E_{HOMO} - E_{LUMO}$ is the energy gap between the HOMO and the LUMO, PCWT^E is the most negative partial charge weighted topological electronic index, and N_{rings} is the number of rings.

$$-\log L_w = (2.82 \pm 0.22) + (41.61 \pm 1.11)\mathrm{HDCA2} + (0.71 \pm 0.02)(N_O + 2N_N) - (0.17 \pm 0.02)(E_{HOMO} - E_{LUMO}) + (0.13 \pm 0.01)\mathrm{PCWT}^E + (0.79 \pm 0.06)N_{rings} \quad (2)$$

$$R^2 = 0.9407, \quad F = 1269, \quad s = 0.73, \quad N = 406$$

In the present study we demonstrate that the aqueous solubilities (S_w) of hydrocarbons and halogenated hydrocarbons can be estimated with a multiple linear regression using just three geometrical, topological, and constitutional descriptors. It is important to produce regressions with as few parameters as possible, so that the contribution of each descriptor may be interpreted more clearly. Topological descriptors have proven useful for the prediction of many properties and activities, as described by Kier and Hall.[14,15] The three descriptors necessary for a good fit to the available solubility data can all be calculated from the molecular structure directly, requiring no measurements of physical properties of the compounds to be investigated.

METHODS

Computational Methods. The quantitative structure−property relationships were developed using CODESSA.[16] This program performs the calculation of descriptors and searches for the best multiple linear relationships between calculated descriptors and experimental property data, as described in previous papers.[2,17] Briefly, the three-dimensional molecular structures of the molecules were drawn and preoptimized using a molecular-mechanics based program.[18] Quantum chemical parameters derived from AM1 molecular orbitals were calculated using MOPAC 6.0,[19] a standard semiempirical quantum-chemical code. The MOPAC output files were supplied to CODESSA to calculate five types of molecular descriptors: constitutional, topological, geometrical, electrostatic, and quantum-chemical.[16,17] A total of 404 descriptors were generated for each molecule. The number of descriptors to be used in the search for optimal correlations was reduced by eliminating highly correlated descriptors, descriptors that are defined for only a subset of the molecules, and descriptors that individually correlate poorly with the solubility. With those remaining, the best multiple linear regressions were identified.

Data Sources. The aqueous solubility data in the present study (Table 2) was compiled from several literature sources. Several previous studies compiled large sets of measurements.[9−12,20] We tracked most of these S_w values back to the original references and added several additional original references, covering aliphatic and aromatic hydrocarbons,[21−24] halogenated hydrocarbons[5,23,25] and PCBs.[6,25,26] For measurements published before 1960, two studies were used that tabulated original measurements from many sources.[27,28] A summary of S_w measurements for all compounds derived from these sources is listed in Table 2. Values used were reported for 25 °C and 1 atm pressure. For cases where multiple solubility values were available, the average was used to generate the regression coefficients.

No solubilities were used for compounds that exist as gases under the conditions of measurement. This eliminated all hydrocarbons with fewer than five carbons, 2,2-dimethylpropane, and several one and two carbon halogenated hydrocarbons. The treatment of the solubility of gases in water in terms of S_w is influenced by different intermolecular forces, as the cohesive energy of the liquid is not accounted for;[2] constrast L_w and see discussion above.

Estimation of Error in Literature Values. A regression developed from measured property values is limited by the accuracy of the experimental measurements that are being used. Knowledge of the experimental error of the solubility measurements is important to this work, because attempts to create regression equations with a variance smaller than the experimental variance would be misleading. Such regressions may be making predictions for some systematic error in the measurement techniques. The magnitude of the error in S_w is difficult to determine, as measurements are often reported in the literature without an error estimate. Even with published error estimates, the variation between research groups can be significantly larger. For many substances, only a single reported value can be found in the literature, so calculating the variance between a number of independent measurements is not possible.

Of the 250 structures listed in Table 2, 65 had solubility data measured by more than one researcher. From the differences in these measurements, some insight into the experimental error may be gained. The standard deviation of the data set may be estimated, given the following assumptions. We assume that there is no systematic error between the different researchers and the different experimental techniques, that the variance of the measurements for different researchers is equal, and that the error distribution is Gaussian. For the 65 duplicate measurements in Table 2, the standard deviation can then be estimated using eq 3.

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{S_{i,1} - S_{i,2}}{2}\right)^2 \qquad (3)$$

This equation is equivalent to the standard equation for the calculation of variance (s^2), where the sums of differences between the solubilities of each molecule ($S_{i,j}$) and the average of reported solubilities (S_{av}) for each molecule simplifies to the difference between the two reported solubility values ($S_{i,1} - S_{i,2}$). With application of eq 3, the estimated experimental error (s) is a minimum of 0.16 log units. The actual error in the data set may be significantly larger than this value, as the set of 65 values is inherently biased, by generally including only measurements of compounds that are more readily available. The error in the S_w measurements for some of the more unusual compounds with only one published measurement available would be expected to be larger, as additional error due to the synthesis and purification of these compounds could be expected. That this is true is suggested by comparing the differences in measured values where two measurements are available. The largest differences are for less usual compounds, while common alkanes and such solvents as benzene and toluene have very small relative differences between measurements.

RESULTS AND DISCUSSION

Best Multiple Linear Regression. The CODESSA program was used to analyze the solubilities and descriptor values for 241 molecular structures, as described in Computational Methods and in previous efforts.[2,17] Selective elimination of the 404 available descriptors resulted in 185 remaining descriptors, which were used to identify a three term regression that fit the aqueous solubility data well. The use of more descriptors resulted in higher correlation coefficients, but lower F statistic values, suggesting that the additional descriptors were not contributing to improve the fit to the actual property but rather to the error in the measurements. In eq 4, S_w is the solubility (moles/liter),

$$-\log S_w = -(0.13 \pm 0.11) + (0.0437 \pm 0.0007)\mathrm{MV} - (0.258 \pm 0.031)^0\mathrm{BIC} + (0.0523 \pm 0.0047)\mathrm{PNSA} \qquad (4)$$

$$R^2 = 0.959, \quad F = 1861, \quad s = 0.386, \quad N = 241$$

MV is the molecular volume (Å^3), ^{0}BIC is the structural information content of 0th order, and PNSA is the atomic charge weighted partial negative surface area. The statistical terms are the correlation coefficient (R), F statistic (F), standard error (s), and the number of molecules used to calculate the regression (N). It was observed that certain structures were outliers in several simple one and two descriptor regressions, having a large difference between their calculated and experimental solubility values. Nine structures (3.6% of the data set) were removed as indicated in Table 2. A possible reason for the problems with these compounds may simply be that the measured values in the literature are not accurate, as multiple measurements are not available to assess the accuracy of the measurements. Some of the solubility values relied on come from references as far back as the 1920's, and by virtue of appearing in certain compilations,[26,27] these values have been accepted in many quantitative structure−property relationship (QSPR) studies. As these few molecules appeared as outliers in the regressions performed using all 250 structures, it was desirable to improve the regressions by eliminating these few outliers that were possibly erroneous measurements and contributed

Table 2. Aqueous Solubility Values for 250 Compounds[a]

structure name	− log *S*	ref	− log *S*	ref	eq 4	resid	structure name	− log *S*	ref	− log *S*	ref	eq 4	resid
Hydrocarbons													
C5													
n-pentane	3.27	20	3.25	22	2.91	−0.35	2-methyl-2-butene	2.56	27			2.21	−0.35
2-methylbutane	3.18	20	3.18	21	2.94	−0.24	3-methyl-1-butene	2.73	20			2.11	−0.62
cyclopentane	2.65	20	2.64	21	2.60	−0.05	cyclopentene	2.10	20			1.83	−0.27
1-pentene	2.68	20			2.08	−0.60	1,4-pentadiene	2.09	20			1.63	−0.46
trans-2-pentene	2.54	20			2.22	−0.32	2-methyl-1,3-butadiene	2.03	20			1.68	−0.35
2-methyl-1-butene	2.73	11			2.11	−0.62							
C6													
n-hexane	3.96	20	3.82	22	3.63	−0.26	2-hexene	3.10	11			2.82	−0.28
2-methylpentane	3.79	20	3.82	21	3.59	−0.21	2-methyl-1-pentene	3.03	20			2.72	−0.31
3-methylpentane	3.83	20	3.82	21	3.56	−0.26	4-methyl-1-pentene	3.24	20			2.71	−0.53
2,2-dimethylbutane	3.67	20	3.61	21	3.55	−0.09	cyclohexene	2.59	20			2.45	−0.14
2,3-dimethylbutane	3.61	11	3.66	21	3.48	−0.15	1,4-cyclohexadiene	2.06	20			1.99	−0.07
cyclohexane	3.19	20			3.25	0.06	1,5-hexadiene	2.69	20			2.33	−0.36
methylcyclopentane	3.30	20	3.31	21	3.18	−0.12	2,3-dimethyl-1,3-butadiene	2.40	27			2.31	−0.09
1-hexene	3.23	20	3.08	22	2.66	−0.50	benzene	1.64	20	1.65	21	2.15	0.51
C7													
n-heptane	4.53	20	4.65	21	4.16	−0.43	cycloheptane	3.51	20			3.83	0.32
2-methylhexane	4.60	21			4.23	−0.37	1-heptene	3.73	22			3.37	−0.36
3-methylhexane	4.58	21			4.16	−0.42	*trans*-2-heptene	3.82	20			3.50	−0.32
2,2-dimethylpentane	3.67	20	4.36	21	4.13	0.12	cycloheptene	3.16	20			3.11	−0.05
2,3-dimethylpentane	4.28	21			4.17	−0.11	1-methylcyclohexene	3.27	20			3.09	−0.18
2,4-dimethylpentane	4.39	20	4.36	21	4.18	−0.20	1,6-heptadiene	3.34	20			2.78	−0.56
3,3-dimethylpentane	4.23	21			4.14	−0.09	cycloheptatriene	2.17	20	1.16	11	2.48	0.81
methylcyclohexane	3.85	20	3.79	21	3.85	0.03	toluene	2.25	20	2.22	21	2.52	0.28
C8													
n-octane	5.24	20	5.43	21	4.85	−0.48	propylcyclopentane	4.74	21			4.44	−0.30
3-methylheptane	5.16	21			4.80	−0.36	1-octene	4.62	20	4.44	22	3.89	−0.64
2,2,4-trimethylpentane	4.67	20	5.00	21	4.77	−0.07	ethylbenzene	2.84	20	2.91	21	3.08	0.20
2,3,4-trimethylpentane	4.93	21			4.68	−0.25	1,2-dimethylbenzene	2.78	20	2.81	21	3.07	0.27
cyclooctane	4.15	20			4.50	0.35	1,3-dimethylbenzene	2.82	22	2.90	21	3.12	0.26
cis-1,2-dimethylcyclohexane	4.27	20			4.40	0.13	1,4-dimethylbenzene	2.69	22	2.83	21	3.07	0.31
1,4-dimethylcyclohexane	4.47	21			4.46	−0.01	4-vinylcyclohexene	3.34	20			3.10	−0.24
1,1,3-trimethylcyclopentane	4.48	21			4.40	−0.08	styrene	2.54	26			2.89	0.35
C9													
n-nonane	6.02	21			5.49	−0.53	isopropylbenzene (2-propyl)	3.38	20	3.40	21	3.57	0.18
4-methyloctane	6.05	21			5.46	−0.49	1,2,3-trimethylbenzene	3.26	22			3.62	0.36
2,2,5-trimethylhexane	5.05	20			5.38	0.33	1,2,4-trimethylbenzene	3.32	20	3.37	21	3.62	0.27
1,1,3-trimethylcyclohexane	4.85	21			5.02	0.17	1,3,5-trimethylbenzene	3.09	26			3.69	0.60
1,1,4-trimethylcyclohexane	5.22	11			5.04	−0.18	1-ethyl-2-methylbenzene	3.21	22			3.65	0.44
1-nonene	5.05	22			4.59	−0.46	indan	3.13	21	3.03	23	3.31	0.23
n-propylbenzene	3.34	20	3.36	22	3.71	0.36							
C10													
n-decane	6.98	11			6.12	−0.86	*tert*-butylbenzene	3.60	20			4.13	0.53
pentylcyclopentane	6.08	21			5.74	−0.34	1-methyl-4-isopropylbenzene	3.76	11			4.24	0.48
decalin	5.19	11			5.31	0.12	1,2,4,5-tetramethylbenzene	3.84	26	4.59	21	4.27	0.05
1-decene	4.39	11			5.17	0.78	1,2,3,4-tetrahydronaphthalene	3.49	26			3.91	0.42
n-butylbenzene	3.94	20	3.99	22	4.26	0.30	naphthalene	3.57	26	3.61	23	3.74	0.15
sec-butylbenzene (2-butyl)	3.67	20			4.20	0.53							
C11													
n-undecane	7.59	11			6.74	−0.85	pentamethylbenzene	3.98	26			4.82	0.84
2-methyldecalin	6.57	11			5.96	−0.61	1-methylnaphthalene	3.71	23			4.12	0.41
n-pentylbenzene	4.59	22			4.85	0.26	2-methylnaphthalene	3.77	23			4.22	0.45
tert-amylbenzene	4.15	20	4.15	26	4.71	0.56							
C12													
n-dodecane	7.67	11			7.41	−0.26	1,4-dimethylnaphthalene	4.14	23			4.66	0.52
n-hexylbenzene	5.20	22			5.48	0.28	1,5-dimethylnaphthalene	4.68	23			4.62	−0.06
biphenyl	4.35	23	4.46	26	4.59	0.18	2,3-dimethylnaphthalene	4.72	23			4.70	−0.02
1-ethylnaphthalene	4.16	23			4.68	0.52	2,6-dimethylnaphthalene	4.89	23			4.70	−0.19
2-ethylnaphthalene	4.29	11			4.69	0.40	acenaphthene	4.59	23	4.40	26	4.36	−0.14
1,3-dimethylnaphthalene	4.29	23			4.72	0.43							
C13													
diphenylmethane	4.06	26	4.70	10	5.24	0.86	1,4,5-trimethylnaphthalene	4.91	23			5.15	0.24
fluorene	4.93	23			4.82	−0.11							
C14+													
phenanthrene	5.15	23	5.05	26	5.27	0.17	*n*-octadecane[b]	8.08	19			11.21	3.13
anthracene	6.38	23	6.35	26	5.50	−0.87	benz[*a*]anthracene	7.33	26			6.96	−0.37
n-hexadecane[b]	7.80	19			9.95	2.15	benz[*a*]pyrene	8.22	19			7.74	−0.48
pyrene	6.18	23	6.09	26	6.07	−0.07							

Table 2. (Continued)

structure name	− log *S*	ref	− log *S*	ref	eq 4	resid	structure name	− log *S*	ref	− log *S*	ref	eq 4	resid
Halogenated Hydrocarbons													
C1													
iodomethane	1.00	26			1.10	0.10	bromochloromethane	0.89	22			0.91	0.02
dichloromethane	0.63	26			0.62	−0.01	tribromomethane	1.91	26			2.61	0.70
dibromomethane	1.18	26			1.67	0.49	trichloromethane	1.21	26			1.51	0.30
diiodomethane	2.35	26			2.17	−0.18	tetrachloromethane	2.30	26			2.93	0.63
C2													
bromoethane	1.08	26			1.39	0.31	1,1,2,2-tetrabromoethane[b]	2.73	26			3.96	1.23
iodoethane	1.60	26			1.90	0.30	pentachloroethane	2.64	26	2.61	27	3.52	0.90
1,1-dichloroethane	1.29	26			1.27	−0.02	2-bromo-2-chloro-1,1,1-trifluoroethane	1.70	11			1.46	−0.24
1,2-dichloroethane	1.06	26			1.08	0.02	1,1,2-trichlorotrifluoroethane	3.04	27			2.39	−0.65
1,2-dibromoethane	1.67	26			1.97	0.30	1,1,2,2-tetrachlorodifluoroethane	3.19	27			3.03	−0.16
1-chloro-2-fluoroethane	0.51	11			0.51	0.00	hexachloroethane	3.67	27	4.47	3	4.77	0.70
1-bromo-2-chloroethane	1.32	26			1.33	0.01	*cis*-1,2-dichloroethene	1.10	27	1.44	26	1.35	0.08
1,1,1-trichloroethane	2.01	26			1.97	−0.04	*trans*-1,2-dichloroethene	1.19	27			1.36	0.17
1,1,2-trichloroethane	1.48	26			1.67	0.19	trichloroethene	2.12	26	1.98	22	2.32	0.27
1-chloro-1,1-difluoroethane	1.20	11			0.53	−0.67	tetrachloroethene	3.04	27			3.31	0.27
1,1,1,2-tetrachloroethane	2.18	26			2.62	0.44	tetrafluoroethene	1.60	11			1.27	−0.33
1,1,2,2-tetrachloroethane	2.77	26	1.76	27	2.45	0.18							
C3													
1-chloropropane	1.46	26			1.54	0.08	1,3-dichloropropane	1.62	26			1.60	−0.02
2-chloropropane	1.41	26			1.54	0.13	1,2-dibromopropane	2.14	27			2.63	0.49
1-bromopropane	1.70	26			2.01	0.31	1,3-dibromopropane	2.08	27			2.55	0.47
2-bromopropane	1.59	26			1.93	0.34	1-bromo-3-chloropropane	1.85	22			1.92	0.07
1-iodopropane	2.20	26			2.51	0.31	3-chloropropene	1.28	27	1.60	11	2.23	0.79
2-iodopropane	2.09	26			2.47	0.38	3-bromopropene	1.50	22			1.28	−0.22
1,2-dichloropropane	1.61	26			1.67	0.06							
C4													
1-chlorobutane	2.03	22	2.16	26	2.16	0.07	1-bromo-2-methylpropane	2.43	26			2.58	0.15
1-bromobutane	2.20	22	2.36	26	2.61	0.33	1,1-dichlorobutane	2.40	27			2.51	0.11
1-iodobutane	2.94	26			3.16	0.22	4-bromo-1-butene	2.25	22			1.76	−0.49
1-chloro-2-methylpropane	2.00	26			2.18	0.18	hexachloro-1,3-butadiene	4.91	11			5.63	0.72
2-chloro-2-methylpropane	2.13	26			2.20	0.07							
C5													
1-chloropentane	2.73	27			2.82	0.09	1-bromopentane	3.08	22			3.25	0.17
2-chloropentane	2.63	27			2.84	0.21	1-bromo-3-methylbutane	2.88	27			3.16	0.28
3-chloropentane	2.63	27			2.85	0.22							
C6													
1-bromohexane	3.81	22			3.87	0.06	1,2-dichlorobenzene	3.01	26	3.20	24	3.11	0.01
lindane	4.59	19			5.19	0.60	1,3-dichlorobenzene	3.08	26	3.07	24	3.08	0.00
bromobenzene	2.68	26	2.64	5	2.76	0.10	1,4-dichlorobenzene	3.28	26	3.68	24	3.00	−0.48
chlorobenzene	2.44	26	2.58	22	2.44	−0.07	1,2-dibromobenzene	3.50	5			3.51	0.01
fluorobenzene	1.87	26			1.88	0.01	1,3-dibromobenzene	3.38	5			3.61	0.23
iodobenzene	3.11	26	3.01	22	2.99	−0.07							
C6													
1,4-dibromobenzene	4.07	27			3.52	−0.55	2-fluorochlorobenzene	2.54	22			2.38	−0.16
1,2-difluorobenzene	2.00	5			1.96	−0.04	3-fluorochlorobenzene	2.54	22			2.38	−0.16
1,3-difluorobenzene	2.00	5			1.85	−0.15	1,2,3-trichlorobenzene	4.17	24	3.76	5	4.58	0.61
1,4-difluorobenzene	1.97	5			1.87	−0.10	1,2,4-trichlorobenzene	3.72	5	3.60	24	3.79	0.13
1,2-diiodobenzene	4.24	5			3.97	−0.27	1,3,5-trichlorobenzene	4.64	24	4.44	5	3.78	−0.76
1,3-diiodobenzene	4.57	5			4.10	−0.47	1,2,4-tribromobenzene	4.50	5			4.40	−0.10
1,4-diiodobenzene[b]	5.25	5			3.98	−1.27	1,3,5-tribromobenzene[b]	5.60	5			4.39	−1.21
2-bromochlorobenzene	3.19	5			3.18	−0.01	1,2,3,4-tetrachlorobenzene	4.70	5	4.25	24	4.52	0.05
3-bromochlorobenzene	3.21	5			3.17	−0.04	1,2,3,5-tetrachlorobenzene	4.79	5	4.87	24	4.53	−0.30
4-bromochlorobenzene	3.63	5			3.15	−0.48	1,2,4,5-tetrachlorobenzene	5.56	5	4.96	24	4.58	−0.68
4-bromoiodobenzene	4.56	5			3.62	−0.94	1,2,4,5-tetrabromobenzene[b]	6.98	5			5.25	−1.73
2-chloroiodobenzene	3.54	5			3.45	−0.09	pentachlorobenzene	5.65	5	5.48	24	5.35	−0.22
3-chloroiodobenzene	3.55	5			3.41	−0.14	hexachlorobenzene	6.78	24			6.27	−0.51
4-chloroiodobenzene	4.03	5			3.37	−0.66							
C7													
1-chloroheptane	4.00	22			4.08	0.08							
1-bromoheptane	4.43	22			4.53	0.10							
1-iodoheptane	4.81	22			5.05	0.24							
α-chlorotoluene	2.43	11			2.54	0.11							
α,α,α-trifluorotoluene	2.51	11			1.92	−0.59							
C8													
1-bromo-2-ethylbenzene	3.67	27			3.68	0.01	1-bromooctane	5.06	22			5.15	0.09
C9													
1-bromo-2-isopropylbenzene	4.19	27			4.23	0.04							

Table 2. (Continued)

structure name	− log S	ref	− log S	ref	eq 4	resid
Halogenated Hydrocarbons (continued)						
2-PCB	4.57	24			4.91	0.34
3-PCB	5.16	11			4.95	−0.21
2,2′-PCB	5.27	6			5.56	0.29
2,4-PCB	5.28	6			5.44	0.16
2,4′-PCB	5.07	10			5.45	0.38
2,5-PCB	5.06	24	5.26	6	5.50	0.34
2,6-PCB	5.21	24	4.97	6	5.35	0.26
3,3′-PCB	5.80	6			5.42	−0.38
4,4′-PCB[b]	6.79	6			5.44	−1.35
2,2′,5-PCB	5.70	6			6.15	0.45
2,3′,5-PCB	6.01	6			6.02	0.02
2,4,4′-PCB	6.34	6			6.01	−0.33
2,4,5-PCB	6.20	24			6.12	−0.08
2,4,6-PCB	6.06	24	6.01	6	6.01	−0.02
2,2′,3,3′-PCB	7.27	6			6.97	−0.30
2,2′,4′,5-PCB	7.25	24			6.63	−0.62
2,2′,5,5′-PCB	6.43	6			6.81	0.38
C10+						
2,2′,5,6′-PCB	6.79	6			6.78	0.00
2,2′,6,6′-PCB	7.39	6			6.88	−0.51
2,3,4,5-PCB	7.14	24			6.79	−0.35
2,3,5,6-PCB	7.32	6			6.94	−0.39
2,3′,4,4′-PCB	7.70	19			6.79	−0.91
3,3′,4,4′-PCB[b]	8.73	6			6.83	−1.90
3,3′,5,5′-PCB[b]	8.37	6			6.68	−1.69
2,2′,4,5,5′-PCB	7.23	24	7.68	6	7.57	0.11
2,3,4,5,6-PCB	7.77	24	7.91	6	7.38	−0.46
2,2′,3,3′,4,4′-PCB	9.11	24	9.01	6	8.34	−0.72
2,2′,3,3′,6,6′-PCB	7.78	24			8.12	0.34
2,2′,4,4′,5,5′-PCB	8.62	6			8.34	−0.28
2,2′,4,4′,6,6′-PCB	8.95	24	8.20	6	8.32	−0.26
2,2′,3,3′,4,4′,6-PCB	8.26	24			9.08	0.82
2,2′,3,3′,5,5′,6,6′-PCB	9.04	24			9.85	0.81
2,2′,3,3′,4,5,5′,6,6′-PCB	10.41	24			10.53	0.12

[a] The solubility is tabulated as the negative logarithm of the solubility (mol/L). [b] Values not used in the calculation of regression (eq 4).

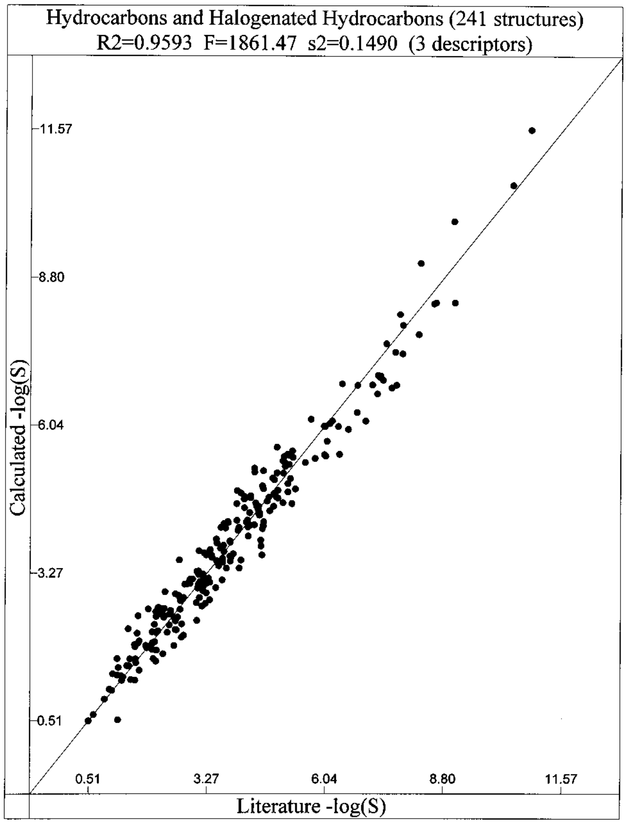

Figure 1. Scatter plot of the calculated vs literature aqueous solubility values for 241 hydrocarbons and halogenated hydrocarbons.

nothing to the improvement of the model in general. The regression calculated with the remaining 241 structures had a significantly improved correlation, with the correlation coefficient (R^2) increasing from 0.933 to 0.959 and the standard error (s) falling from 0.511 to 0.386 log units. The scatter plot of calculated vs experimental aqueous solubility for eq 4 can be seen in Figure 1.

Calculation of Descriptors. Several methods for the calculation of molecular volume are summarized by Yalkowsky and Banerjee.[1] The molecular volume for this study was calculated by considering all atoms in the molecule to be spheres with radii equal to the atomic van der Waals radii. To handle the problem of multiple intersecting spheres, a rectangular box containing the molecule was divided into a fine grid of elements in Cartesian coordinates, and the volume of each element that fell within the van der Waals radius of one of the atoms was included in the molecular volume. This approach can correctly calculate the molecular volume for any number of overlapping spheres (Table 3), with the accuracy determined by the size of the grid elements.[16] For the molecules in this study the molecular volume ranged from 52.9 to 298 Å^3.

The bonding information content[29] is a topological descriptor which encodes both the degree of branching and the number of different atom types in the molecule. It is calculated by eq 5, where i is the number of classes of atoms, n is the number of atoms, n_i is the number of atoms in class i, and q is the number of edges (bonds) in the structural graph of the molecule.

$$^0\mathrm{BIC} = -\left(\frac{1}{\log_2 q}\right)\sum_i \frac{n_i}{n} \log_2\left(\frac{n_i}{n}\right) \qquad (5)$$

The final descriptor, the atomic charge weighted partial negative surface area (PNSA),[30] is a sum of atomic surface areas weighted both by the surface charge of the atom ($-A_i$) and the partial atomic charge (Q_i^-) for the negatively charged surfaces only (eq 6). The atomic area is a solvent accessible area as calculated by MOPAC.[19]

$$\mathrm{PNSA} = \sum(-A_i Q_i^-) \qquad (6)$$

The intercorrelations between the descriptors are +0.69 for MV and ^{0}BIC, −0.18 for MV and PNSA, and −0.42 for ^{0}BIC and PNSA.

Interpretation of Descriptors. The molecular volume is clearly the most important descriptor for aqueous solubility, as can be seen by its high t-test value (the ratio of the coefficient to the coefficient error). In order for a solute to enter into aqueous solution, a cavity must be formed in the solvent for the solute molecule to occupy. Water as a solvent

Table 3. Descriptor Values for 250 Hydrocarbons and Halogenated Hydrocarbons

structure name	MV	^{0}BIC	PNSA	structure name	MV	^{0}BIC	PNSA
Hydrocarbons							
C5							
n-pentane	95.5253	3.7144	−3.3886	2-methyl-2-butene	90.0233	4.8929	−6.2249
2-methylbutane	96.4419	3.7144	−3.4161	3-methyl-1-butene	90.5076	4.8929	−8.5291
cyclopentane	85.5575	3.5257	−1.8368	cyclopentene	79.7471	4.6889	−5.8791
1-pentene	90.7837	4.8929	−9.4598	1,4-pentadiene	84.6252	4.4926	−14.8815
trans-2-pentene	90.1433	4.8929	−6.2080	2-methyl-1,3-butadiene	84.2849	4.4926	−13.6969
2-methyl-1-butene	90.2794	4.8929	−8.3689				
C6							
n-hexane	113.4246	4.1493	−2.2615	2-hexene	107.1182	5.3919	−6.4762
2-methylpentane	113.4448	4.1493	−3.0474	2-methyl-1-pentene	107.2063	5.3919	−8.3831
3-methylpentane	113.3607	4.1493	−3.5601	4-methyl-1-pentene	107.4462	5.3919	−8.8579
2,2-dimethylbutane	112.9923	4.1493	−3.5319	cyclohexene	95.5734	5.1952	−4.7922
2,3-dimethylbutane	112.5482	4.1493	−4.4566	1,4-cyclohexadiene	89.8911	5.0699	−9.5368
cyclohexane	101.8841	3.9639	−0.8278	1,5-hexadiene	101.6200	5.3190	−11.6609
methylcyclopentane	102.1679	3.9639	−2.4461	2,3-dimethyl-1,3-butadiene	101.6159	5.3190	−12.0518
1-hexene	107.2741	5.3919	−9.6492	benzene	87.8504	3.3473	−13.2120
C7							
n-heptane	129.2815	4.5724	−3.3878	cycloheptane	118.4947	4.3904	−1.4755
2-methylhexane	130.2774	4.5724	−2.7708	1-heptene	124.4731	5.8599	−8.1782
3-methylhexane	130.3215	4.5724	−4.2597	*trans*-2-heptene	124.3891	5.8599	−5.6061
2,2-dimethylpentane	130.0214	4.5724	−4.5526	cycloheptene	112.4404	5.6690	−3.9805
2,3-dimethylpentane	130.2135	4.5724	−3.9122	1-methylcyclohexene	112.6922	5.6690	−4.6388
2,4-dimethylpentane	130.1854	4.5724	−3.7747	1,6-heptadiene	118.6749	5.9800	−14.0564
3,3-dimethylpentane	130.2295	4.5724	−4.5949	cycloheptatriene	100.1073	4.8872	−9.6523
methylcyclohexane	118.6505	4.3904	−1.2507	toluene	105.0931	4.8872	−13.0038
C8							
n-octane	147.4000	4.9857	−3.2465	propylcyclopentane	136.4580	4.8068	−2.8458
3-methylheptane	147.1720	4.9857	−4.0957	1-octene	141.7431	6.3068	−10.3430
2,2,4-trimethylpentane	146.9882	4.9857	−4.5431	ethylbenzene	122.2162	5.8345	−11.9414
2,3,4-trimethylpentane	147.0521	4.9857	−6.2606	1,2-dimethylbenzene (*o*-xylene)	121.5917	5.8345	−11.5833
cyclooctane	135.9902	4.8068	−1.3756	1,3-dimethylbenzene (*m*-xylene)	121.7118	5.8345	−10.7182
cis-1,2-dimethylcyclohexane	135.6064	4.8068	−2.8437	1,4-dimethylbenzene (*p*-xylene)	121.9961	5.8345	−11.9780
1,4-dimethylcyclohexane	135.5865	4.8068	−1.6720	4-vinylcyclohexene	124.0410	6.3442	−10.4657
1,1,3-trimethylcyclopentane	136.2302	4.8068	−3.4335	styrene	115.5496	4.0000	−19.0612
C9							
n-nonane	164.4147	5.3904	−3.2482	isopropylbenzene (2-propyl)	139.0725	6.5921	−12.8835
4-methyloctane	164.4867	5.3904	−3.8347	1,2,3-trimethylbenzene	138.0971	6.5921	−11.1964
2,2,5-trimethylhexane	164.1507	5.3904	−5.2138	1,2,4-trimethylbenzene	137.9770	6.5921	−11.0892
1,1,3-trimethylcyclohexane	152.4572	5.2144	−3.0935	1,3,5-trimethylbenzene	137.9211	6.5921	−9.5835
1,1,4-trimethylcyclohexane	151.9133	5.2144	−2.2011	1-ethyl-2-methylbenzene	138.6247	6.5921	−10.9689
1-nonene	158.3780	6.7381	−8.7236	indan	128.1985	6.2997	−10.1520
n-propylbenzene	139.1165	6.5921	−10.1593				
C10							
n-decane	181.2692	5.7877	−3.3885	*tert*-butylbenzene	155.4514	7.2468	−12.6387
pentylcyclopentane	171.2229	5.6143	−3.0838	1-methyl-4-isopropylbenzene	155.4995	7.2468	−10.6197
decalin	158.2260	5.4195	−1.4273	1,2,4,5-tetramethylbenzene	154.6920	7.2468	−9.3568
1-decene	175.7165	7.1569	−10.2003	1,2,3,4-tetrahydronaphthalene	144.3657	6.9808	−8.8965
n-butylbenzene	156.2910	7.2468	−10.8247	naphthalene	134.0108	4.1995	−17.2143
sec-butylbenzene (2-butyl)	155.9433	7.2468	−11.6896				
C11							
n-undecane	197.9002	6.1783	−3.5306	pentamethylbenzene	170.3234	7.8367	−9.0112
2-methyldecalin	175.2046	5.8176	−1.1398	1-methylnaphthalene	152.4571	5.7855	−17.5935
n-pentylbenzene	173.4457	7.8367	−10.9212	2-methylnaphthalene	152.4251	5.7855	−15.5820
tert-amylbenzene	172.2263	7.8367	−12.6374				
C12							
n-dodecane	215.2266	6.5631	−3.2836	1,4-dimethylnaphthalene	168.6403	6.8478	−15.4746
n-hexylbenzene	190.2683	8.3818	−10.3456	1,5-dimethylnaphthalene	168.7722	6.8478	−16.2854
biphenyl	162.8391	4.8344	−21.9849	2,3-dimethylnaphthalene	168.6843	6.8478	−14.8512
1-ethylnaphthalene	168.9960	6.8478	−15.4896	2,6-dimethylnaphthalene	169.3319	6.8478	−15.3678
2-ethylnaphthalene	169.0041	6.8478	−15.2766	acenaphthene	158.2060	6.4709	−14.3777
1,3-dimethylnaphthalene	168.7283	6.8478	−14.4060				
C13							
diphenylmethane	182.2929	6.3946	−18.0721	1,4,5-trimethylnaphthalene	184.4276	7.7183	−15.0004
fluorene	171.2029	5.9871	−18.7436				
C14+							
phenanthrene	180.0333	5.0031	−22.3700	*n*-octadecane	315.7193	8.7751	−3.5308
anthracene	180.0734	5.0031	−18.1587	benz[*a*]anthracene	226.0642	5.7744	−24.8181
n-hexadecane	282.2855	8.0537	−3.3895	benz[*a*]pyrene	242.8749	5.9076	−23.2103
pyrene	196.8919	5.1446	−20.5790				

Table 3. (Continued)

structure name	MV	^{0}BIC	PNSA	structure name	MV	^{0}BIC	PNSA
Halogenated Hydrocarbons							
C1							
iodomethane	52.8737	3.4274	−3.6877	bromochloromethane	61.0451	4.8048	−7.3558
dichloromethane	56.2855	3.8048	−13.7905	tribromomethane (bromoform)	82.9085	3.4274	0.0000
dibromomethane	64.7691	3.8048	−0.7919	trichloromethane (chloroform)	70.0552	3.4274	−10.2417
diiodomethane	76.5778	3.8048	−1.2007	tetrachloromethane	84.5290	1.8048	−3.1972
C2							
bromoethane	63.5531	3.7011	−5.7654	1,1,2,2-tetrabromoethane	118.9269	4.2745	0.0000
iodoethane	69.7671	3.7011	−1.2186	pentachloroethane	115.7253	3.7011	−8.5441
1,1-dichloroethane	73.5326	4.2745	−13.6256	2-bromo-2-chloro-1,1,1-trifluoroethane	94.3971	6.1428	−18.1981
1,2-dichloroethane	73.6887	4.2745	−17.2983	1,1,2-trichlorotrifluoroethane	103.9408	4.4491	−16.6306
1,2-dibromoethane	82.2721	4.2745	−7.5026	1,1,2,2-tetrachlorodifluoroethane	112.4721	4.2745	−12.3363
1-chloro-2-fluoroethane	64.7171	4.9869	−17.2168	hexachloroethane	129.6736	2.3119	−3.1115
1-bromo-2-chloroethane	77.7383	4.9869	−12.3622	*cis*-1,2-dichloroethene	66.9941	4.0956	−7.3489
1,1,1-trichloroethane	87.2541	4.4491	−10.7250	*trans*-1,2-dichloroethene	67.0181	4.0956	−7.1615
1,1,2-trichloroethane	87.6064	4.4491	−16.7757	trichloroethene	80.6636	3.7705	−1.9627
1-chloro-1,1-difluoroethane	70.0192	5.4304	−19.1101	tetrachloroethene	94.2086	2.3729	−1.2160
1,1,1,2-tetrachloroethane	101.8801	4.2745	−11.4404	tetrafluoroethene	61.9531	2.3729	−13.2875
1,1,2,2-tetrachloroethane	101.3115	4.2745	−14.1924				
C3							
1-chloropropane	76.1178	4.1083	−11.4059	1,3-dichloropropane	90.2633	4.7530	−18.8145
2-chloropropane	76.1738	4.1083	−11.3057	1,2-dibromopropane	99.3230	4.7530	−6.6435
1-bromopropane	80.7315	4.1083	−6.2752	1,3-dibromopropane	98.7467	4.7530	−7.6671
2-bromopropane	80.3114	4.1083	−7.3138	1-bromo-3-chloropropane	94.5091	5.3550	−13.2789
1-iodopropane	86.4137	4.1083	−1.3526	3-chloropropene	107.4021	5.2164	−18.7578
2-iodopropane	86.4540	4.1083	−2.2551	3-bromopropene (allyl bromide)	74.8090	4.9732	−10.9821
1,2-dichloropropane	90.6716	4.7530	−17.8446				
C4							
1-chlorobutane	93.4325	4.5329	−11.7982	1-bromo-2-methylpropane	97.7784	4.5329	−7.5265
1-bromobutane	97.7463	4.5329	−6.8387	1,1-dichlorobutane	107.2220	5.2164	−13.3573
1-iodobutane	103.5446	4.5329	−1.1313	4-bromo-1-butene	91.9721	5.5986	−12.9782
1-chloro-2-methylpropane	93.4727	4.5329	−11.5381	hexachloro-1,3-butadiene	150.1745	3.0630	−0.1612
2-chloro-2-methylpropane	93.2606	4.5329	−11.0256				
C5							
1-chloropentane	110.2192	4.9559	−11.1893	1-bromopentane	114.7652	4.9559	−6.6924
2-chloropentane	110.4194	4.9559	−10.9505	1-bromo-3-methylbutane	114.4210	4.9559	−8.2019
3-chloropentane	110.5115	4.9559	−10.7688				
C6							
1-bromohexane	131.5886	5.3728	−6.8165	3-bromochlorobenzene	119.7111	5.4421	−10.0826
lindane (hexachlorocyclohexane)	186.4664	6.8417	−20.2520	4-bromochlorobenzene	119.7872	5.4421	−10.5334
bromobenzene	106.3697	4.4352	−11.6256	4-bromoiodobenzene	129.9654	5.4421	−10.0812
chlorobenzene	101.1917	4.4352	−13.4000	2-chloroiodobenzene	125.7734	5.4421	−9.7903
fluorobenzene	92.7603	4.4352	−17.1424	3-chloroiodobenzene	125.8377	5.4421	−10.5207
iodobenzene	112.2361	4.4352	−12.1135	4-chloroiodobenzene	125.5212	5.4421	−11.0579
1,2-dichlorobenzene (*o*-)	114.8131	4.8842	−9.7795	2-fluorochlorobenzene	124.2611	7.0603	−20.9582
1,3-dichlorobenzene (*m*-)	114.5328	4.8842	−10.2594	3-fluorochlorobenzene	124.4691	7.0603	−21.0884
1,4-dichlorobenzene (*p*-)	114.5289	4.8842	−11.6052	1,2,3-trichlorobenzene	141.5507	4.8842	−4.1487
1,2-dibromobenzene	124.1290	4.8842	−10.0535	1,2,4-trichlorobenzene	127.9620	5.0210	−7.0745
1,3-dibromobenzene	124.0929	4.8842	−7.9746	1,3,5-trichlorobenzene	127.9701	5.0210	−7.3147
1,4-dibromobenzene	123.8567	4.8842	−9.5689	1,2,4-tribromobenzene	141.7669	5.0210	−6.9190
1,2-difluorobenzene	98.1826	4.8842	−17.9616	1,3,5-tribromobenzene	141.3634	5.0210	−6.7790
1,3-difluorobenzene	98.0664	4.8842	−19.9594	1,2,3,4-tetrachlorobenzene	141.4908	4.8842	−5.0599
1,4-difluorobenzene	97.9303	4.8842	−19.4960	1,2,3,5-tetrachlorobenzene	141.5747	4.8842	−5.1187
1,2-diiodobenzene	135.7422	4.8842	−10.9448	1,2,4,5-tetrachlorobenzene	141.5707	4.8842	−4.1473
1,3-diiodobenzene	135.7902	4.8842	−8.3419	1,2,4,5-tetrabromobenzene	159.2175	4.8842	−6.0285
1,4-diiodobenzene	135.6183	4.8842	−10.5935	pentachlorobenzene	154.8713	4.4352	−2.7605
2-bromochlorobenzene	120.1273	5.4421	−10.2178	hexachlorobenzene	168.0480	3.3473	−1.3639
C7							
1-chloroheptane	144.0339	5.7826	−11.3558	α-chlorotoluene	119.2469	6.0002	−18.9057
1-bromoheptane	148.6991	5.7826	−6.5328	α,α,α-trifluorotoluene	121.0917	6.8415	−28.2731
1-iodoheptane	154.5520	5.7826	−1.5056				
C8							
1-bromo-2-ethylbenzene	140.0760	6.9592	−9.8622	1-bromooctane	165.8099	6.1855	−7.0905
C9							
1-bromo-2-isopropylbenzene	157.2106	7.7226	−9.9068				
C10+							
2-PCB	175.7600	5.8712	−21.5136	2,2′,6,6′-PCB	217.8731	6.9808	−13.4889
3-PCB	176.2237	5.8712	−21.0297	2,3,4,5-PCB	215.5619	6.9808	−13.2686
2,2′-PCB	190.5761	6.4303	−18.6780	2,3,5,6-PCB	217.2894	6.9808	−11.9768

Table 3. (Continued)

structure name	MV	^{0}BIC	PNSA	structure name	MV	^{0}BIC	PNSA
Halogenated Hydrocarbons (continued)							
C10+ (continued)							
2,4-PCB	189.0686	6.4303	−19.6737	2,3′,4,4′-PCB	215.6658	6.9808	−13.4661
2,4′-PCB	188.8688	6.4303	−19.3146	3,3′,4,4′-PCB	216.2858	6.9808	−13.1458
2,5-PCB	188.7687	6.4303	−18.3855	3,3′,5,5′-PCB	215.9100	6.9808	−15.6621
2,6-PCB	188.3809	6.4303	−20.9131	2,2′,4,5,5′-PCB	230.7100	7.0450	−10.7725
3,3′-PCB	189.5123	6.4303	−20.4704	2,3,4,5,6-PCB	228.0430	7.0450	−12.1982
4,4′-PCB	189.2003	6.4303	−19.8242	2,2′,3,3′,4,4′-PCB	244.7305	6.9808	−8.0777
2,2′,5-PCB	204.1687	6.7826	−17.0172	2,2′,3,3′,6,6′-PCB	242.7716	6.9808	−10.6546
2,3′,5-PCB	202.3051	6.7826	−17.8680	2,2′,4,4′,5,5′-PCB	244.1187	6.9808	−7.5567
2,4,4′-PCB	202.4332	6.7826	−18.1428	2,2′,4,4′,6,6′-PCB	244.4065	6.9808	−8.1535
2,4,5-PCB	202.2493	6.7826	−15.9822	2,2′,3,3′,4,4′,6-PCB	258.0671	6.7826	−6.0467
2,4,6-PCB	202.0055	6.7826	−17.8488	2,2′,3,3′,5,5′,6,6′-PCB	271.3278	6.4303	−4.0201
2,2′,3,3′-PCB	217.9371	6.9808	−11.8653	2,2′,3,3′,4,5,5′,6,6′-PCB	282.7735	5.8712	−3.3542
2,2′,4′,5-PCB	215.0143	6.9808	−15.9817	decachloro-PCB	297.7333	4.8344	−1.2059
2,2′,5,5′-PCB	217.2537	6.9808	−14.3453	*p*,*p*′-DDT	267.2780	10.1786	−19.2229
2,2′,5,6′-PCB	217.6654	6.9808	−15.1992				

Table 4. Coefficients and Statistical Parameters for the Blind Test Regressions[a]

groups	R^2	F	s^2	intercept	MV	^{0}BIC	PNSA
ABC	*0.959*	*1861*	*0.149*	*-0.13 ± 0.11*	*(4.37 ± 0.07) × 10^{-2}*	*-0.26 ± 0.03*	*0.052 ± 0.005*
AB	0.962	1328	0.144	-0.15 ± 0.14	(4.41 ± 0.09) × 10^{-2}	-0.27 ± 0.04	0.052 ± 0.006
AC	0.955	1122	0.156	-0.06 ± 0.14	(4.31 ± 0.10) × 10^{-2}	-0.26 ± 0.04	0.053 ± 0.006
BC	0.960	1260	0.150	-0.20 ± 0.14	(4.38 ± 0.09) × 10^{-2}	-0.25 ± 0.04	0.052 ± 0.006

[a] The 241 structures were broken into three sets of 80 or 81 structures, called A, B, and C. The three descriptors were molecular volume, bonding information content (order 0), and the atomic charge weighted partial negative surface area.

would much prefer to interact with itself or other hydrogen bonding or ionic species than with a nonpolar solute, so there is an increasing penalty (and thus lower solubility) for larger hydrocarbon solutes. Molecular volume is a key term in several different approaches to property prediction, such as linear solvation energy relationships (LSER)[8] and various group contribution methods such as UNIFAC.[1] In the present study, we showed that molecular volume by itself predicts solubility by eq 7 to reasonable accuracy, although not as well as the three descriptor relationship of eq 4.

$$-\log S_{\mathrm{w}} = -(1.27 \pm 0.12) + (0.0380 \pm 0.0008)\mathrm{MV} \tag{7}$$

$$R^2 = 0.904, \quad F = 2253, \quad s = 0.590, \quad N = 241$$

The major problem with molecular volume as the sole descriptor is that it does not account for steric interactions or conformational effects. In this study it also does not account for increased solubility due to favorable dipole−dipole interactions between the halogen atoms and water. The final two terms in the three descriptor regression, the bonding information content and the partial negative surface area, specifically correct for these deficiencies of the molecular volume. Both of these additional terms serve to decrease $-\log S_{\mathrm{w}}$ and thus increase solubility. The information content descriptor increases with both unsaturation and the number of different atom types and thus the product of the descriptor and its negative coefficient decreases with complexity. The partial negative surface area term decreases with an increase in atoms with negatively charged surfaces. It is intuitively expected that the presence of these features would increase the aqueous solubility of a compound. The PNSA should be directly related to the hydrogen bond or Lewis basicity of the molecule. A larger (in magnitude) value of PNSA should and does lead to a larger S_{w}.

Blind Tests of Descriptors. To further test the ability of the three descriptors to predict aqueous solubility, tests were conducted to determine whether regression models made with a subset of the structures would accurately predict the aqueous solubility values of the remaining structures. The general applicability of the QSPR approach would be established if the aqueous solubility of a large set of compounds could be accurately predicted, given regressions developed with another, exclusive set of compounds. The 241 structures were broken into three sets of 80 or 81 structures, regression models were made from two of the sets, and the solubilities of the third set were predicted. These three sets (A, B, and C) contained equal numbers of hydrocarbons and halogenated hydrocarbons.

Several interesting observations can be made of the regressions in Table 4. First, the statistical ratings of the regressions are all approximately equal, suggesting that the regressions made with any two-thirds of the structure set are of equal predictive ability as the regression made with all structures. Second, the calculated coefficients of the descriptors for the two group regressions are all within the error estimate of the coefficients for the regression made with all structures, suggesting that the coefficient values are reliable. Finally, when considering the correlation coefficients of the predictions for the blind tests in Table 5, it is apparent that the ability of the regressions made using two-thirds of the structures to predict the aqueous solubility for the excluded third is essentially equal. The average correlation coefficient for the blind cases (AB → C, AC → B, BC → A) was equal to the correlation coefficient ($R^2 = 0.959$) for the regression created using all structures, when used to predict all structures (ABC → ABC).

Another measure of the quality of the regression is the cross-validated correlation coefficient (R^2_{cv}). For each experimental data point, the regression is recalculated with

Table 5. Correlation Coefficients for Aqueous Solubility Blind Test Regressions[a]

groups used in regression	groups predicted	R^2	groups used in regression	groups predicted	R^2
ABC	*ABC*	*0.959*	AB	C	0.953
ABC	A	0.949	AC	B	0.967
ABC	B	0.967	BC	A	0.957
ABC	C	0.953			

[a] Predictions are made with regressions calculated from subsets of the 241 structure set.

the same descriptors but for the data set without this point. The obtained regression is used to predict the value of this point, and the set of estimated solubilities calculated in this manner is correlated with the experimental solubility values.[16] For the best regression (eq 4), $R^2_{cv} = 0.958$, as compared to $R^2 = 0.959$, indicating the high quality of the regression equation.

CONCLUSION

A quantitative structure−property relationship approach was used to predict the aqueous solubility of hydrocarbons and halogenated hydrocarbons for a diverse set of 241 compounds. Key to this effort was the attempt to produce regressions with as few descriptors as possible. Predictions using just three terms can estimate aqueous solubility with a correlation coefficient of 0.979, having a standard error of 0.386 log units. Molecular volume is the key descriptor, with corrections using one topological and one electrostatic descriptor to account for features that increase the solubility of the molecules. This predictive equation compares favorably with previously developed relations as may be judged from Table 1. Thus, in comparison with the first six correlations of Table 1, our equation describes a much larger and much more diverse data set. In comparison with the other equations, the number of descriptors is drastically reduced from nine or more to just three. Furthermore, our equation allows the estimation of aqueous solubility given only the molecular structure and should be applicable to as yet unstudied hydrocarbons and halogenated hydrocarbons.

Although the scope of the present paper is limited to a QSPR correlation of aqueous solubility, it is of considerable interest to note that our conclusions are generally in line with the general treatment of solubility advocated by Kamlet and Taft.[8] These authors postulated that solvent/solute based properties can be broken down into the effects of three types of interactions: cavity effects, polarizability and dipolarity, and hydrogen bonding. Two of the three descriptors that we have found correlate with the cavity effect and with hydrogen bonding. We hope that the present study will assist and stimulate further work into the general treatment of solubility.

ACKNOWLEDGMENT

The authors would like to thank Professor Mati Karelson of Tartu University, Estonia, for his helpful comments and suggestions.

REFERENCES AND NOTES

(1) Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Dekker: New York, 1992.

(2) Katritzky, A. R.; Mu, L.; Karelson, M. A QSPR Study of the Solubility of Gases and Vapors in Water. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1162.

(3) Mackay, D.; Shiu, W. Y. A Critical Review of Henry's Law Constants for Chemicals of Environmental Interest. *J. Phys. Chem. Ref. Data* **1981**, *10*, 1175−1199.

(4) Yalkowsky, S. H.; Valvani, S. C. Solubilities and Partitioning. 2. Relationships between Aqueous Solubilities, Partition Coefficients, and Molecular Surface Areas of Rigid Aromatic Hydrocarbons. *J. Chem. Eng. Data* **1979**, *24*, 127−129.

(5) Yalkowsky, S. H.; Orr, R. J.; Valvani, S. C. Solubility and Partitioning. 3. The Solubility of Halobenzenes in Water. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 351−353.

(6) Dunnivant, F. M.; Elzerman, A. W.; Jurs, P. C.; Hasan, M. N. Quantitative Structure−Property Relationships for Aqueous Solubilities and Henry's Law Constants of Polychlorinated Biphenyls. *Environ. Sci. Technol.* **1992**, *26*, 1567−1573.

(7) Amidon, G. L.; Yalkowsky, S. H.; Anik, S. T.; Valvani, S. C. Solubility of Nonelectrolytes in Polar Solvents. V: Estimation of the Solubility of Aliphatic Monofunctional Compounds in Water Using a Molecular Surface Area Approach. *J. Phys. Chem.* **1975**, *79*, 2239−2246.

(8) Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Carr, P. W.; Doherty, R. F.; Taft, R. W. Linear Solvation Energy Relationships. 41. Important Differences between Aqueous Solubility Relationships for Aliphatic and Aromatic Solutes. *J. Phys. Chem.* **1987**, *91*, 1996−2004.

(9) Kan, A. T.; Tomson, M. B. UNIFAC Prediction of Aqueous and Nonaqueous Solubilities of Chemicals with Environmental Interest. *Environ. Sci. Technol.* **1996**, *30*, 1369−1376.

(10) Bodor, N.; Huang, M.-J. A New Method for the Estimation of the Aqueous Solubility of Organic Compounds. *J. Pharm. Sci.* **1992**, *81*, 954−960.

(11) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.

(12) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds Using a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.

(13) Nirmalakhandan, N. N.; Speece, R. E. Prediction of Aqueous Solubility of Organic Chemicals Based on Molecular Structure. *Environ. Sci. Technol.* **1988**, *22*, 328−338.

(14) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(15) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.

(16) Katritzky, A. R.; Lobanov, V.; Karelson, M. *CODESSA Version 2.0 Users Manual*; University of Florida: Gainesville, FL, 1994.

(17) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.

(18) *PCMODEL molecular modeling software*; Serena Software: Bloomington, IN, 1987−92.

(19) Stewart, J. J. P. MOPAC 6.0., Quantum Chemistry Program Exchange No. 455, Bloomington, IN, 1989. See: http://ccl.osc.edu/ccl/qcpe/.

(20) Schwarzenbach, R. P.; Gschwend, P. M.; Imboden, D. M. *Environmental Organic Chemistry*; Wiley: New York, 1993.

(21) McAuliffe, C. Solubility in Water of Paraffin, Cycloparaffin, Olefin, Acetylene, Cycloolefin, and Aromatic Hydrocarbons. *J. Phys. Chem.* **1966**, *70*, 1267−1275.

(22) Price, L. C. Aqueous Solubility of Petroleum as Applied to Its Origin and Primary Migration. *Am. Assoc. Pet. Geol. Bull.* **1976**, *60*, 213−244.

(23) Tewari, Y. B.; Miller, M. M.; Wasik, S. P.; Martire, D. E. Aqueous Solubility and Octanol/Water Partition Coefficient of Organic Compounds at 25.0 °C. *J. Chem. Eng. Data* **1982**, *27*, 451−454.

(24) Mackay, D.; Shiu, W. Y. Aqueous Solubility of Polynuclear Aromatic Hydrocarbons. *J. Chem. Eng. Data* **1977**, *22*, 399−402.

(25) Miller, M. M.; Ghodbane, S.; Wasik, S. P.; Tewari, Y. B.; Martire, D. E. Aqueous Solubilities, Octanol/Water Partition Coefficients, and Entropies of Melting of Chlorinated Benzenes and Biphenyls. *J. Chem. Eng. Data* **1984**, *29*, 184−190.

(26) Li, A.; Andren, A. W. Solubility of Polychlorinated Biphenyls in Water/Alcohol Mixtures. 1. Experimental Data. *Environ. Sci. Technol.* **1994**, *28*, 47−52.

(27) Deno, N. C.; Berkheimer, H. E. Activity Coefficients as a Function of Structure and Media. *J. Chem. Eng. Data* **1960**, *5*, 1−5.

(28) Irmann, F. Eine einfache Korrelation zwischen Wasserlösichkeit und Struktur von Kohlenwasserstoffen und Halogenkohlenwasserstoffen. *Chem.-Ing.-Tech.* **1965**, *37*, 789−798.

(29) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Comparative Study of Lipophilicity versus Topological Molecular Descriptors in Biological Correlations. *J. Pharm. Sci.* **1984**, *73*, 429−437.

(30) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

CI9700438