

A QSPR Study of the Solubility of Gases and Vapors in Water

Alan R. Katritzky* and Lan Mu

Florida Institute of Heterocyclic Compounds, Department of Chemistry, University of Florida,
Gainesville, Florida 32611-7200

Mati Karelson*

Department of Chemistry, University of Tartu, 2 Jakobi Street, Tartu EE2400, Estonia

Received June 20, 1996[®]

QSPR correlation equations were developed for the prediction of the solubilities of organic gases and vapors in water. A two-parameter correlation with the squared correlation coefficient $R^2 = 0.977$ gives excellent predictions for 95 alkanes, cycloalkanes, alkenes, alkylarenes, and alkynes. A satisfactory description ($R^2 = 0.941$) of the gas solubilities of 406 organic compounds with a large structural variability was obtained using a five-parameter QSPR equation. Notably, all the parameters involved in these equations can be derived solely from the chemical structure of the compounds which makes them very useful for the prediction of the solubilities of unknown or unavailable compounds.

INTRODUCTION

The solubility of nonelectrolytes in water and aqueous solutions is of significant chemical and thermodynamical interest as well as of great practical importance. Notably, the rationalization and modeling of the environmental fate of pollutants, including the prediction of soil/sediment adsorption coefficients and bioconcentration factors for nonionic pesticides from aqueous solutions, require a knowledge of their solubilities together with other properties.¹ Biomedical applications of aqueous solubilities include their use in predictions of the suitability of gaseous anesthetics, blood substitutes, and oxygen carriers.^{2,3}

Aqueous solutions, including the aqueous solutions of gases, have many peculiarities not generally observed for other solvents.⁴ The tendency of a molecule to pass from the gas phase to a dilute aqueous solution has been proposed as a measure of the hydrophilic character of the gas.⁴ Solute–solute interactions determine the solubilities of gases, liquids, or solids in a given liquid or solution. However, the solubilities of gases are more directly related to the condensed phase solute–solvent interactions than the solubilities of liquids or solids which also involve the solute–solute interactions. Consequently, gas solubilities provide a more appropriate experimental characteristic for the construction of theoretical models of solute–solvent interactions in condensed media.

$$L = \frac{\text{concn of solute in soln}}{\text{concn of solute in the gas phase}} \quad (1)$$

$$H = \frac{\text{fugacity of solute in the gas phase}}{\text{activity of solute in soln}} \quad (2)$$

The Ostwald solubility coefficient (L) is defined as the ratio of the equilibrium concentrations of a gaseous compound in the liquid and in the gas phase (eq 1), where a superscript w (L^w) usually denotes water as a solvent. Another commonly used gas solubility parameter is the

Henry's law constant H (eq 2), which is approximately equal to L^{-1} . Analytical constraints in dealing with concentrations at the nanogram level limit the availability of reliable experimental solubility data for many chemical compounds. Therefore it is of great practical importance to develop theoretical approaches for the accurate prediction of the solubilities of gases in liquids.

Rigorous statistical mechanical treatments of gas–liquid solubilities are based on the calculation of chemical potential for the solute in the gas and in the solvent phases from the respective partition functions, which should account for all the relevant energetic and probabilistic factors. The classical approximation associates the solubility with the average interaction potential of a solute with the entire solvent at some coupling strength.^{2,5} However, the corresponding equations are cumbersome and, in addition, require the parameterization of the solute–solvent and solvent–solvent interaction potentials. Accurate calculation of the solvation free energy or even the calculation of the relative free energies of solvation of structurally related compounds using molecular dynamics or Monte Carlo simulation methods^{6–8} is arduous for several reasons, including the very large size of the configuration space accessible to solvent molecules and the approximate nature of the intermolecular force fields used.

Therefore, continuous efforts have been made to derive analytical expressions of solubility using simplified physical models. Such theoretical tools, developed primarily for the description of hydrophobic effects, include the classical electrostatic models of Onsager⁹ and Kirkwood,¹⁰ scaled particle theory (SPT),^{2,5,11} and semiempirical perturbation theory.^{5,12} The widely used approach developed by Eley¹³ is based on the division of the solution process into two consecutive steps, of which the first is associated with cavity formation in the solvent, and the second accounts for the physical interactions between a solute molecule immersed into the cavity and the surrounding solvent molecules. The first step should reflect the short-range repulsive part of the potential, whereas the second involves the attractive long-

[®] Abstract published in *Advance ACS Abstracts*, November 1, 1996.

range solute-solvent interaction potential. In the electrostatic approach,^{9,10} the excess chemical potential of a solute molecule in a solvent is expressed as an integral involving the solvent-solute pair distribution function, and solvent-solute interaction potential. This technique is also referred to as the charging technique and has been successful in studying solutions in which the solutes are relatively small.

Scaled particle theory (SPT)^{2,5,11} characterizes the solvent using a hard-sphere molecular diameter obtained from the heat of vaporization and thermal expansivity of the solvent. The SPT technique has also been extended to real fluids and applied to the solubilities of gases in liquids.^{2,5,14} Other statistical-mechanical models of gas solubility include the thermodynamic statistical model, based on the distribution of molecular populations among quantized discrete energy levels^{15,16} and the statistical thermodynamic lattice-gas model.¹⁷

While these theoretical approaches to describe and understand the processes involving solubility and the underlying effects are well established from the viewpoint of statistical mechanics and thermodynamics, it is still difficult to quantify the solubilities of most real systems from such physical models. In fact, the application of these theoretical models has been restricted mainly to the calculation of the solubilities of the inert gases and of small gaseous hydrocarbons. Their application to larger and more complex molecules has not been successful due to their limitations, particularly in regards to the complexity of the estimation of the appropriate entropic contributions to the Gibbs free energies of solvation.

An alternative approach to the prediction of gas solubilities in liquids is based on the development of quantitative structure-property relationships (QSPR). Although this approach has been successfully applied for the correlation of many diverse physical properties of chemical compounds,¹⁸⁻²⁴ relatively few attempts have been made to correlate or to predict the solubility of gases in water. Hine and Mookerjee⁴ first reported empirically based group and bond contribution schemes. Their group scheme reproduced solubility values of 292 diverse compounds with a standard error of 0.12 log units using 69 group contribution factors. Their bond contribution scheme reproduced solubility values of 263 diverse solutes with a standard error of 0.42 log units using 34 bonds. A group contribution scheme was also developed by Cabani *et al.*²⁵ who derived 28 group contributions to reproduce 209 log L^w values of diverse compounds to within 0.09 log units. Because of the large number of groups or bonds involved in these schemes, neither the group-contribution method nor the bond-contribution method conveys much understanding of the physical nature of the relationship between the molecular structure and interactions and the solubility of gases in water. Moreover, their application to the prediction of solubilities of compounds containing structural functionality not included in the original set is questionable at best.

Nirmalakhandan and Speece²⁶ developed a predictive model involving three structurally determined descriptors: the valence connectivity index (introduced by Kier and Hall²⁷), a molecular polarizability descriptor Φ ,²⁸ and an indicator variable, I , for the presence of an electronegative atom. The solubility data for 180 diverse compounds was reproduced with a standard error of 0.262 log units. However, the polarizability descriptor was calculated on the basis of an atomic contribution scheme involving another

11 empirical parameters, which effectively increases the number of parameters employed in the correlation.

Later, Russell, Dixon, and Jurs²⁹ correlated the logarithms of Henry's law constant, log H , of 63 diverse gases in water, using five theoretically calculated descriptors. Their linear regression model had a correlation coefficient of 0.978, a standard deviation of 0.375 log units, and an F -statistic of 250. It was suggested by the authors that the factors influencing the solubility of gases in water were related to the solute bulk, the lipophilicity, and the polarizability. However, the data set used²⁹ was rather limited due to its restricted number of compounds included. Study of a larger data set is necessary to provide more profound understanding of the gas solubility.

Most recently, Abraham *et al.*³⁰ predicted the solubility of 408 diverse gases in water with the five LSER (Linear Solvation Energy Relationship) descriptors, comprising the excess molar fraction R_2 , calculated from the experimental molar refraction, the experimentally determined dipolarity/polarizability π_2^H and effective hydrogen-bond acidity $\Sigma\alpha_2^H$ and basicity $\Sigma\beta_2^H$, and the McGowan characteristic volume V_x (calculated from some tabulated atomic increments). The regression model had an R of 0.9976 with a standard deviation of 0.151 log units and an F -value of 16 810. Although this correlation equation can be interpreted term-by-term using well-established chemical principles, the LSER descriptors are limited in their ability to make *a priori* predictions since four experimentally determined values are required for each compound. Also, since the resulting correlations do not relate the property to the molecular structural information, it is difficult to elucidate how molecular structures affect the observed property.

The primary objective of the present work is the development of QSPR equations for the aqueous solubility of a wide variety of gaseous compounds. It is expected that in addition to their applicabilities to predictions, those equations should provide useful information about the physical mechanisms determining the solubility of gases in water. Notably, our previous study of the normal boiling points of a diverse set of organic molecules^{24a} suggested that the molecular size-dependent dispersion and cavity-formation effects can be represented by a limited set of structural parameters, whereas the specific hydrogen-bonding interactions in the molecule can be effectively described by the area weighted surface charge on the hydrogen bonding donor atoms. Again, our work on nonionic surfactants demonstrated that critical micelle concentrations can be adequately predicted by two topological fragment descriptors for a hydrophobic tail together with a single fragment descriptor for a hydrophilic head.^{24b} Thus, the second objective of the present work is to assess the applicability of such theoretical molecular descriptors for the prediction of another condensed phase property, the solubility of gases in liquids.

RESULTS AND DISCUSSION

The set of 408 organic compounds, previously used by Abraham *et al.*³⁰ in their LSER treatment, was chosen for the QSPR treatment. Two compounds (sulfur hexafluoride and triethyl phosphate) were eliminated from Abraham's original set since the AM1 semiempirical method used in this work for the calculation of theoretical descriptors gives

inadequate geometry and charge distribution for these compounds. The remaining set of 406 compounds is still structurally very diverse and includes saturated and unsaturated hydrocarbons, halogenated compounds, and compounds containing hydroxyl, cyano, amino, nitro, thio, ester, ether, carbonyl, and carboxyl functional groups and furan, pyran, pyridine, and pyrazine rings. However, the set is compact enough to allow the calculation of the descriptors and the development of the QSPR equations within a reasonable time frame. The structures were drawn and preoptimized by the MMX molecular mechanics method using the PCMODEL³¹ program. The final geometry optimization of compounds was performed on an IBM RISC/6000 Model 320 computer using the semiempirical quantum-chemical AM1 parameterization with the MOPAC 6.0 program modified by the inclusion of the self-consistent reaction field (SCRF) method.³² The MOPAC results on individual compounds were loaded into the CODESSA program³³ along with the experimental solubility data.

The CODESSA program implements procedures which enable the calculation of a large selection of descriptors including the number of atoms and bonds of different types, the molecular weight, molecular volume, and many other geometrical characteristics of the molecule, such as Jurs' shape parameters,³⁴ and a large variety of topological indices (Wiener indices,³⁵ Randić connectivity indices,³⁶ Kier and Hall connectivity indices,²⁷ and information content indices.³⁷) Numerous quantum-chemical descriptors are extracted and calculated from the AM1 output.^{22,24,33,38,39} The CPSA descriptors proposed by Jurs *et al.*^{40,41} are also included in the CODESSA program based on AM1 calculated atomic partial charges. Quantum-chemical descriptors used in this work also included the most positive and the most negative Mulliken net atomic charges, frontier molecular orbital (FMO) energies, and the respective Fukui FMO nucleophilic, electrophilic, and one-electron reactivity indices. The total dipole moment of the molecule, dipole moment components, and molecular bond orders were also used as descriptors. Additional, more specific, descriptors included the valence state energies of atoms and total Coulombic and exchange energies between atoms in the molecule. The zero-point energy, the calculated electronic and vibrational transition energies, the rotational, vibrational, translational, internal, and total enthalpies, entropies, and heat capacities were also used.

Altogether more than 600 molecular descriptors were calculated using CODESSA for all 406 compounds. The correlation analysis to find the best QSPR model of a given size was carried out using the procedure based on the stepwise scale addition methods (see Appendix).³⁹ A preselection of descriptors was implemented.

Similarly to our previous study,^{24a} we started with a limited subset of hydrocarbons to avoid initial complexities that arise from specific interactions between the solute and solvent (water). In that way, the structural features attributable to size/mass/bulk related cohesiveness, as defined by Cramer,⁴² should become more apparent. This hydrocarbon subset contained 95 alkanes, cycloalkanes, alkenes, alkylarenes, and alkynes. The best two-parameter equation obtained ($R^2 = 0.9765$, $F = 1988$, $s = 0.20$, averaged absolute errors = 0.15 Figure 1) involved the gravitational index over all bonded atoms in the molecule³³ and the complementary information content (^0CIC) (order 0).²⁷ The gravitational

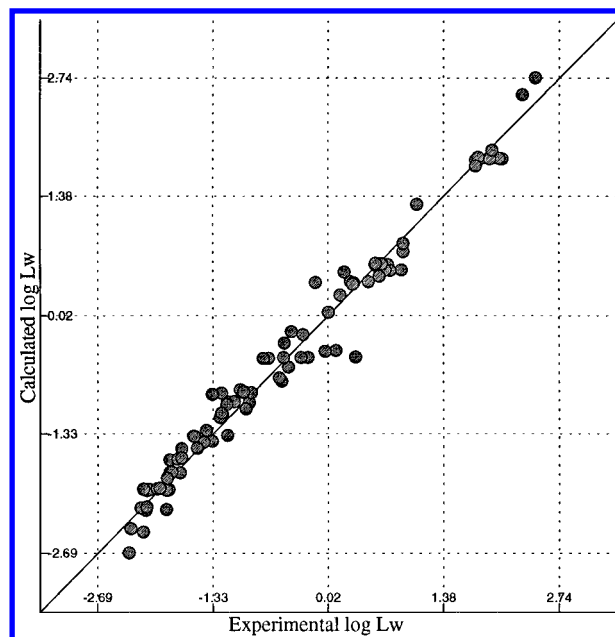


Figure 1. The calculated *vs.* experimental gas solubilities of 95 hydrocarbons using the best two-parameter correlation equation.

Table 1. The Best Two-Parameter Correlation of the Gas Solubility for the Data Set of 95 Hydrocarbons

descriptor	$X \pm X$	<i>t</i> -test	R_2	R_{cv}^2
intercept	-1.37 ± 0.06	-23.16		
G_1	0.0067 ± 0.0001	60.80	0.3997	0.3748
^0CIC	-0.050 ± 0.001	-47.55	0.9765	0.9746

index (G_1) is defined by eq (3), where m_i and m_j are the atomic masses of the bonded atoms and r_{ij} the respective bond lengths, and reflects the effective mass distribution in the molecule. The ^0CIC (order 0) is defined by eq (4), where n_i is the number of atoms in the i th class, and n is the total number of atoms in the molecule. The complementary information content (^0CIC) encodes the degree of branching of a hydrocarbon molecule.

$$G_1 = \sum_{ij}^{\text{all bonded atoms}} \frac{m_i m_j}{r_{ij}^2} \quad (3)$$

$$^0\text{CIC} = \log_2 n - ^0\text{IC} \quad (4)$$

$$^0\text{IC} = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n} \quad (5)$$

Remarkably, these two descriptors (G_1 and ^0CIC) individually have poor correlations with gas solubility ($R^2 = 0.3997$ and 0.0453, respectively). The successive correlation coefficients (R^2), cross-validated correlation coefficients (R_{cv}^2), and Student *t*-test values for these two descriptors are given in Table 1. As reported in our previous study on the normal boiling points,^{24a} the gravitation index (G_1) effectively describes the intermolecular dispersion forces in the bulk solvent media. However, this descriptor varies little among isomers, and thus the additional ^0CIC index accounts for the differences in the shape of these molecules. Evidently the combination of these two descriptors, comprising size and shape information about molecules, adequately represents the

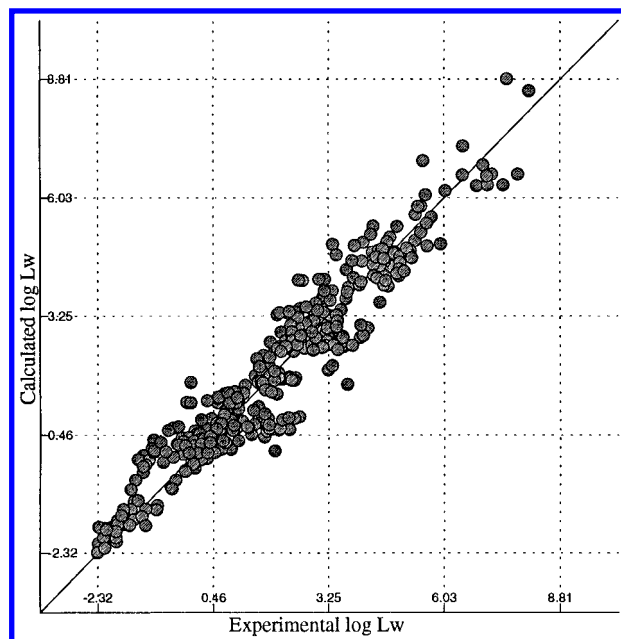


Figure 2. The calculated *vs.* experimental gas solubilities of 406 organic compounds using the preliminary five-parameter correlation equation from Table 2.

effective dispersion and cavity formation effects for the solvation of nonpolar solutes in water.

$$\text{HDCA}(2) = \sum \frac{q_D \sqrt{S_D}}{\sqrt{S_{\text{tot}}}} \quad (6)$$

Further treatment then proceeded with the complete set of 406 compounds to give the five-parameter regression model with $R^2 = 0.9158$ (Figure 2). The most important descriptor in this regression is the hydrogen-bonding related descriptor, HDCA (2), defined by eq (6), where q_D is the partial charge on the hydrogen bonding donor (H) atom(s), S_D denotes the exposed surface area of this atom and S_{tot} is the total molecular surface area, calculated from the van der Waals' radii of the atoms (overlapping spheres). The summation in eq 6 is performed over the number of simultaneously possible hydrogen bonding donor and acceptor pairs per solute molecule. Also, hydrogen atoms attached to carbons connected directly to carbonyl or cyano groups were included as possible hydrogen bonding donor centers (their effectiveness is, of course, much smaller because of the smaller partial charge on them). The additional descriptors in the five-parameter equation (*cf.* Table 2) are as follows: the energy gap between HOMO and LUMO ($E_{\text{HOMO}} - E_{\text{LUMO}}$), the numbers of nitrogen atoms and of oxygen atoms in the molecule, and the most negative partial charge weighted topological electronic index^{43a} (PCWT^E) defined by eq 7, where q_i and q_j are the Zefirov partial charges^{43b} of the bonded atoms, r_{ij} is the respective bond lengths, and Q_{min} is the most negative partial charge.

$$\text{PCWT}^E = \frac{1}{Q_{\text{min}}} \sum_{i < j} \frac{|q_i - q_j|}{r_{ij}^2} \quad (7)$$

The successive values of R^2 , R_{cv}^2 , and the Fisher indexes (F) for this correlation are listed in Table 2. The HOMO – LUMO energy gap relates to the dispersion energy of polar solutes in solution,^{44,45} whereas the most negative partial

Table 2. The Preliminary Five-Parameter Correlation of the Gas Solubility for the Data Set of 406 Diverse Structures

descriptor	R^2	R_{cv}^2	F
intercept			
HDCA (2)	0.5225	0.5182	442.1
$E_{\text{HOMO}} - E_{\text{LUMO}}$	0.6940	0.6901	457.1
N(N)	0.7899	0.7863	504.7
PCWT ^E	0.8805	0.8771	739.2
N(O)	0.9158	0.9130	870.3

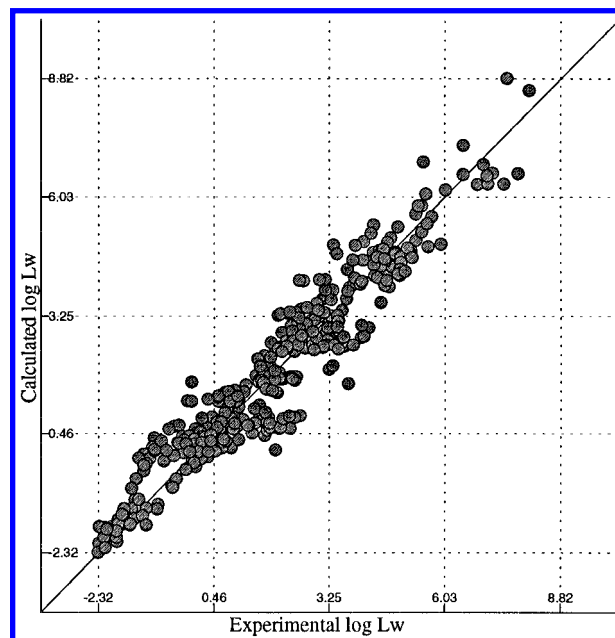


Figure 3. The calculated *vs.* experimental gas solubilities of 406 organic compounds using the best four-parameter correlation equation from Table 3.

Table 3. The Final Best Five-Parameter Correlation of the Gas Solubility for the Data Set of 406 Diverse Structures

descriptor	$X \pm \Delta X$	t-test	R^2	R_{cv}^2
intercept	2.82 ± 0.22	12.92		
HDCA (2)	41.61 ± 1.11	37.44	0.5225	0.5182
N(O) + 2*N(N)	0.71 ± 0.02	28.41	0.6845	0.6801
$E_{\text{HOMO}} - E_{\text{LUMO}}$	-0.17 ± 0.02	-9.42	0.8792	0.8763
PCWT ^E	0.13 ± 0.01	19.03	0.9158	0.9134
N_{rings}	0.79 ± 0.06	12.96	0.9407	0.9386

charge weighted topological electronic index corresponds to the electrostatic part of the solute–solvent interaction. The HDCA(2) was shown to adequately represent the specific hydrogen-bonding interactions in pure liquids.²⁴

The inclusion of the number of oxygen and nitrogen atoms in the correlation model (*cf.* Table 2) may be due to a deficiency of AM1 calculated charges to describe properly the electrostatic and hydrogen bonding interactions with the involvement of these elements in the solute molecules. Notably, the use of a linear combination of the number of nitrogen and oxygen atoms [$2*N(N) + N(O)$] yielded a four-parameter correlation with $R^2 = 0.9158$ identical to that of the five-parameter correlation given in Table 2 ($R^2 = 0.9158$), Figure 3. Proceeding from this four-parameter equation, it was found that still another descriptor, the number of rings, offers a statistically significant improvement to the regression model. Thus, the final five-parameter correlation given in Table 3 has $R^2 = 0.9407$, $F = 1269$, and $s = 0.53$ (averaged absolute error = 0.42) (Figure 4). The successive regression coefficients for each descriptor and the respective standard

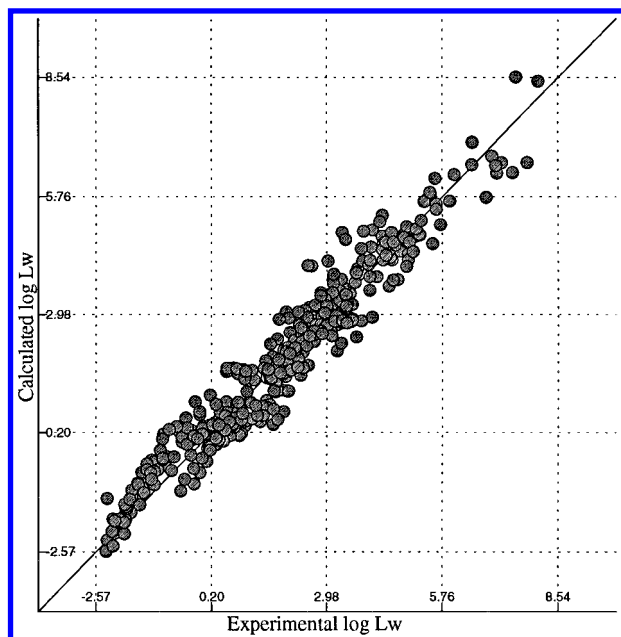


Figure 4. The calculated *vs.* experimental gas solubilities of 406 organic compounds using the final five-parameter correlation equation from Table 3.

Table 4. The One-Parameter Correlation of the Gas Solubility in Hexadecane for the Data Set of 392 Diverse Compounds and Their Hydrocarbon Subset

descriptor	$X \pm \Delta X$	t-test	R^2	F	s
intercept ^a	0.015 ± 0.006	0.27			
intercept ^b	0.08 ± 0.08	1.01			
G_1^a	0.0055 ± 0.0001	43.69	0.8329	1908.7	0.53
G_1^b	0.0065 ± 0.0001	95.30	0.9903	9082.5	0.15

^a For 392 nonfluorinated diverse compounds. ^b For the subset of 95 hydrocarbons.

errors, the R^2 and R_{cv}^2 values, and the Student t-test for this equation are given in Table 3.

None of the five descriptors in our final model rigorously account for the mass/size and shape related intermolecular interactions, and therefore the calculated solubilities for compounds with the same functionality, but different size, do not vary appropriately. For example, the solubilities calculated for benzene and biphenyl differ by only 0.07 log units, but their actual solubility difference is 1.32 log units. Also, the calculated difference in the solubility of formaldehyde and acetaldehyde is 0.10 log units, while the observed difference is 0.55 log units.

An attempt was made to select the best five-parameter model with mass/size and shape (G_1 and 0CIC) descriptors being forced into the regression equation for all 406 compounds. However, these two descriptors were found to have insignificant impact as compared to the hydrogen-bonding related descriptors. Evidently, the intermolecular interaction mechanisms determining the solubility of compounds in water are different for hydrocarbons as compared to molecules containing polar groups. In low polarity solvents, such differences are less apparent: for instance, we found that the solubility of (i) 392 diverse gases and vapors in hexadecane (the data are cited from ref 30) and (ii) their hydrocarbon subset are both well correlated with the gravitation index (G_1) (Table 4).

Although our five-parameter model is statistically less precise than the equation developed by Abraham *et al.*³⁰

utilizing five empirical LSER descriptors, our equation is more general because it uses solely descriptors which can be derived directly from the structure. Thus, our equation can be used for the calculation of solubilities of all compounds including those for which some of the experimentally measured parameters required by Abraham's equation are not available. As compared to the five-parameter model by Jurs and co-workers,²⁹ the correlations developed by us are applicable to a much wider variety of chemical structures. The stabilities of the correlation equations presented in Tables 1–3 are characterized by the corresponding cross-validated correlation coefficients (R_{cv}^2),³³ which are almost identical to the correlation coefficients themselves.

$$\log L^w = (2.65 \pm 0.22) + (42.37 \pm 1.11)HDCA(2) + (0.65 \pm 0.02)[N(O) + 2 \cdot N(N)] + (-0.16 \pm 0.02)(E_{HOMO} - E_{LUMO}) + (0.12 \pm 0.01)PCWT^E + (0.82 \pm 0.06)N_{rings} \quad (8)$$

$$R^2 = 0.9420 \quad F = 1300 \quad s = 0.52 \text{ (averaged absolute error} = 0.42)$$

As already mentioned, all of the structural descriptors used in our correlation equations were calculated from the AM1^{32a} optimized molecular geometries and charge distributions obtained in solution using the self-consistent reaction field (SCRF) approach.^{32b} To investigate how the descriptor values and correlation equations depend on the solvent, we recalculated the descriptors and regressions for the isolated molecules using the same (AM1) parameterization. The five-parameter correlation equation thus developed was very similar to that obtained using SCRF molecular descriptors (eq 8). This result encourages the use of the quantum-chemical data derived from isolated molecules for the description of solvation processes. However, a systematic study on the basis of wider selection of solvational properties is needed for the generalization of this conclusion.

CONCLUSIONS

We have been able to describe satisfactorily the solubilities of structurally widely variable organic molecules in water using correlation equations with a limited number of theoretical molecular descriptors. These descriptors are mostly derived from the quantum-chemical charge distribution of the molecules and have definite physical meaning corresponding to different solute–solvent interactions in solutions. Notably, all the descriptors involved are calculated solely from the chemical structure of compounds, and, therefore, the prediction of the aqueous gas solubility can now be made for any compound, even for those not yet synthesized. Therefore, the equations developed by us should be useful in the development of new technologically and biomedically important materials and compounds.

ACKNOWLEDGMENT

We much appreciate helpful discussions with Dr. V. Lobanov.

APPENDIX. THE PROCEDURE BASED ON THE LINEAR REGRESSION TECHNIQUE TO FIND THE BEST QSPR MODEL OF A GIVEN SIZE

The strategy used to develop the best multilinear QSPR equations from the very large pool of descriptors was based

on a heuristic method. The following procedure was carried out using this approach:

(1) To reduce the number of descriptors in the starting set, the following criteria were applied to eliminate descriptors for which:

- (a) the F -value for the one-parameter correlation with the descriptor was below 1,
- (b) the correlation coefficient for this one-parameter equation was less than R_{\min} , a user-defined value for insignificant correlations ($R_{\min} = 0.01$ was used in the present work),
- (c) the t -value for descriptor in the one-parameter correlation was less than $t_1 = 1.5$,
- (d) the descriptor which was highly intercorrelated with another descriptor was characterized by a higher single-parameter correlation coefficient value for a given property.

All the remaining descriptors were then listed in decreasing order according to the correlation coefficient of the corresponding one-parameter correlation equation.

(2) Starting with the first descriptor from this preselected list of descriptors, two-parameter correlations were calculated using the following pairs: (a) the first descriptor with each of the remaining descriptors, the second descriptor with each of the remaining descriptors, etc. This procedure was continued until for some n th descriptor no correlations with an F -value above one-third of the maximum F -value for a given set had been found. Pairs of descriptors with an intercorrelation above R_{nc} were not considered. The NS (number of descriptor sets to select for the next recursion level) best pairs showing the highest F -values in the two-parameter correlations were selected and processed further as the *working* sets.

(3) Each of the remaining descriptors, if not correlated over R_{nc} with the descriptors already included, were added to the selected *working* set of descriptors. When the resulting correlation had an F -value above $F_{\text{working}}n/(n+1)$ (where n is a number of descriptors in the *working* set plus one), i.e., when this correlation was more significant than the *working* correlation, this extended set of descriptors was treated further.

(4) After all the descriptors have been processed one-by-one, if the maximum number of descriptors ND_{\max} allowed by the user have not yet been achieved, then the NS best extended *working* sets, i.e., the sets with the highest F -values, were submitted to the procedure of step 2. Otherwise the procedure was completed and the correlation with the highest F -value was stored as the final result.

REFERENCES AND NOTES

- (1) Briggs, G. G. Theoretical and Experimental Relationships Between Soil Absorption, Octanol-Water Partition Coefficients, Water Solubilities, Bioconcentration Factors and Parachor. *J. Agric. Food Chem.* **1981**, *29*, 1050–1059.
- (2) Pollack, G. L. Why Gases Dissolve in Liquids. *Science* **1991**, *251*, 1323–1330.
- (3) Imbriani, M. The Pulmonary Capillary Blood Flow (Q_c) at Rest and Work Measured Simultaneously by 3 Inert Substances. *G. Ital. Med. Lav.* **1993**, *15*, 133–138.
- (4) Hine, H.; Mookerjee, P. K. The Intrinsic Hydrophilic Character of Organic Compounds. Correlations in Terms of Structural Contributions. *J. Org. Chem.* **1975**, *40*, 292–298.
- (5) Wilhelm, E.; Battino, R.; Wilcock, R. J. Low-Pressure Solubility of Gases in Liquid Water. *Chem. Rev.* **1977**, *77*, 219–262.
- (6) Swope, W. C.; Andersen, H. C. A Molecular Dynamics Method for Calculating the Solubility of Gases in Liquids and the Hydrophobic Hydration of Inert-Gas Atoms in Aqueous Solution. *J. Phys. Chem.* **1984**, *88*, 6548–6556.
- (7) Watanabe, K.; Andersen, H. C. Molecular Dynamics Study of the Hydrophobic Interaction in an Aqueous Solution of Krypton. *J. Phys. Chem.* **1986**, *90*, 795–802.
- (8) Jorgensen, W. L.; Blake, J. F.; Buckner, J. K. Free Energy of TiP4P Water and the Free Energies of Hydration of CH_4 and Cl^- from Statistical Perturbation Theory. *Chem. Phys.* **1989**, *129*, 193–200.
- (9) Onsager, L. Theories of Concentrated Electrolytes. *Chem. Rev.* **1933**, *13*, 73–89.
- (10) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.
- (11) Pierotti, R. A. The Solubility of Gases in Liquids. *J. Phys. Chem.* **1963**, *67*, 1840–1845.
- (12) Crovetto, R.; Fernandez-Prini, R.; Japas, M. L. Contribution of the Cavity-Formation or Hard-Sphere Term to the Solubility of Simple Gases in Water. *J. Phys. Chem.* **1982**, *86*, 4094–4095.
- (13) Eley, D. D. On The Solubility of Gases. Part I. the Inert Gases in Water. *Trans. Faraday Soc.* **1939**, *35*, 1281–1293. On The Solubility of Gases. Part II. a Comparison of Organic Solvents with Water. *Trans. Faraday Soc.* **1939**, *35*, 1421–1432.
- (14) Guttman, H. J.; Anderson, C. F.; Record, M. T., Jr. Analyses of Thermodynamic Data for Concentrated Hemoglobin Solutions Using Scaled Particle Theory: Implications for a Simple Two-State Model of Water in Thermodynamic Analyses of Crowding In Vitro and In Vivo. *Biophys. J.* **1995**, *68*, 835–846.
- (15) Braibanti, A.; Fiescaro, E.; Dallavalle, F.; Lamb, J. D.; Oscarson, J. L.; Rao, R. S. Molecular Thermodynamic Model for the Solubility of Noble Gases in Water. *J. Phys. Chem.* **1994**, *98*, 626–634.
- (16) Braibanti, A.; Fiescaro, E.; Ghiozzi, A.; Compari, C.; Panelli, M. Solubility of Inert Gases and Liquid Hydrocarbons in Water. *J. Solution Chem.* **1995**, *24*, 703–718.
- (17) Besseling, N. A. M.; Lyklema, J. Hydrophobic Hydration of Small Apolar Molecules and Extended Surfaces: A Molecular Model. *Pure Appl. Chem.* **1995**, *67*, 881–888.
- (18) Horvath, A. L. *Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds*; Elsevier: Amsterdam, 1992.
- (19) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841–850.
- (20) Hall, L. H.; Kier, L. B. Electrotological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
- (21) 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.
- (22) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factor Using a General Quantitative Structure–Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799–1807.
- (23) 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.
- (24) (a) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. The Correlation of Boiling Points with Molecular Structure. Part I: A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400–10407. (b) Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O. and Karelson, M. Prediction of Critical Micelle Concentration Using a Quantitative Structure–Property Relationship Approach. 1. Nonionic Surfactants. *Langmuir* **1996**, *12*, 1462–1470.
- (25) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *J. Solution Chem.* **1981**, *10*, 563–595.
- (26) Nirmalakhandan, N. N.; Speece, R. E. QSAR Model for Predicting Henry's Constant. *Environ. Sci. Technol.* **1988**, *22*, 1349–1357.
- (27) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Wiley: New York, 1986.
- (28) Horvath, A. L. *Halogenated Hydrocarbons*; Marcel Dekker: New York, 1982.
- (29) Russell, C. J.; Dixon, S. L.; Jurs, P. C. Computer-Assisted Study of the Relationship Between Molecular Structure and Henry's Law Constant. *Anal. Chem.* **1992**, *64*, 1350–1355.
- (30) Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. Hydrogen Bonding. Part 34. The Factors that Influence the Solubility of Gases and Vapours in Water at 298 K, and a New Method for its Determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1777–1791.
- (31) *PCMODEL*, 5th ed.; Serena Software: Bloomington, IN, 1992.
- (32) (a) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F. and Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909. (b) Karelson, M. M.; Tamm, T.; Katritzky, A. R.; Cato, S. J.; Zerner, M. C. M. O.

- Calculations Applicable to Condensed Phases: The Combination of Self-Consistent Reaction Field Theory with Semi-Empirical Quantum-Chemical Methods. *Tetrahedron Comput. Methodol.* **1989**, 2, 295–304.
- (33) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA Training Manual*; 1995.
- (34) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/Activity and Structure/Property Relationships. *Anal. Chim. Acta* **1987**, 199, 99–109.
- (35) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, 69, 17–20.
- (36) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, 97, 6609–6615.
- (37) Stankevich, M. I.; Stankevich, I. V.; Zefirov, N. S. Topological Indices in Organic Chemistry. *Russ. Chem. Rev. (Engl. Transl.)* **1988**, 57, 191–208.
- (38) Murugan, R.; Grendze, M. P.; Toomey, J. E., Jr.; Katritzky, A. R.; Karelson, M.; Lobanov, V.; Rachwal, P. Predicting Physical Properties from Molecular Structure. *CHEMTECH* **1994**, 24, 17.
- (39) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, 96, 1027–1043.
- (40) 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.
- (41) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 306–316.
- (42) Cramer, R. D. III BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid State. *J. Am. Chem. Soc.* **1980**, 102, 1837–1849.
- (43) (a) Osmialowski, K.; Halkiewicz, J.; Kaliszan, R. Quantum Chemical Parameters in Correlation Analysis of Gas-Liquid Chromatographic Retention Indices of Amines. *J. Chromatogr.* **1986**, 361, 63–69. (b) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle. *Dokl. Akad. Nauk (Engl. Transl.)* **1987**, 296, 883–887.
- (44) Rosch, N.; Zerner, M. C. Calculation of Dispersion Energy Shifts in Molecular Electronic Spectra. *J. Phys. Chem.* **1994**, 98, 5817–5823.
- (45) Amos, A. T.; Burrows, B. L. Solvent-Shift Effects on Electronic Spectra and Excited-State Dipole Moments and Polarizabilities. *Adv. Quantum Chem.* **1973**, 7, 289–313.

CI960081B