New Developments in Hydrogen Bonding Acidity and Basicity of Small Organic Molecules for the Prediction of Physical and ADMET Properties. Part 2. The Universal Solvation Equation

Polina V. Oliferenko,[‡] Alexander A. Oliferenko,[‡] Gennadiy Poda,[§] Vladimir A. Palyulin,[‡] Nikolay S. Zefirov,[‡] and Alan R. Katritzky*,[||]

Department of Chemistry, Moscow State University, Moscow, 119992 Russia, Structural & Computational Chemistry Group, Pfizer Global Research & Development, Chesterfield, Missouri 63017, and Department of Chemistry, University of Florida, Gainesville, Florida 32611-7200

Received September 5, 2008

Theoretical quantifications of hydrogen bonding (HB) basicities and acidities, originally developed for aliphatic systems (*J. Chem. Inf. Comput. Sci*. 2004, *44*, 1042−1055), are now extended to cover aromatic, heterocyclic, anionic, cationic and zwitter-ionic molecular fragments, thus encompassing a majority of druggable chemical space. The addition of terms accounting for cavity formation, polarity, hydrophobicity, and resonance allowed us to derive a new equation able to predict accurately free energies of solvation of diverse solutes, interphase transfers, and aqueous solubilities (log S_w). We thus provide a "universal solvation equation" (USE) available for the accurate estimation of desolvation energies in protein−ligand docking, for the prediction of many physical and ADMET properties, and for studying fluid phase equilibria.

1. INTRODUCTION

The capacity of small organic molecules to form hydrogen bonds (HB) affects very significantly a wide range of their physical properties and is thus linked directly to many important properties, including aqueous solubility, boiling, and melting points, viscosity, solvatochromism, lipophilicity (frequently measured by the octanol−water partition coefficient, log *P*, or its ionic counterpart octanol−water distribution coefficient log *D*, measured at a given pH), cell permeability, human intestinal absorption (HIA), blood−brain barrier (BBB) permeation, nonspecific plasma protein binding, metabolic oxidation caused by cytochromes P450, and many others.[1,2] HB capacity governs the transport properties of drugs starting from their absorption in the human intestine, entering the portal vein and, after the first passing through the liver, entering the blood stream, becoming distributed in a large number of organs, tissues, and target cells and eventually reaching specific protein targets in the human body to exhibiting the pharmacological activity, finally undergoing metabolism and elimination through kidneys or liver.

Quantification of HB capacity in the liquid phase (later rationalized in terms of linear solvation energy relationships, LSER) began with solvatochromic and thermodynamic studies in solutions by Taft, Kamlet, Abraham, and others.[3−8] Significant increase of computer speed, parallel computing, and developments in computational chemistry initiated theoretical efforts to extend the LSER methodology beyond experimentally available compounds by deriving theoretical characteristics, mimicking the LSER descriptors, from the MO-LCAO,[9,10] AIM,[11] and MEP theories.[12,13] Numerous molecular descriptors not closely related to LSER have been designed and applied to develop quantitative structure−property relationships (QSPR). Topological, fragmental, geometrical, electrostatic, and quantum mechanical descriptors have been instrumental in the prediction of solvation properties.[14−20] These general purpose descriptors implemented, for example, in the Codessa-Pro[21] and DRAGON[22] systems benefit from a straightforward theoretical basis and a high speed of computation. However, despite the vast number of such descriptors, they could encounter difficulties in describing complex patterns of solute−solvent interactions. The situation, however, may be assisted by the introduction of a few highly specific descriptors tailor-made for modeling specific intermolecular interactions. Several successful attempts are known where this approach has been taken. Rather complex but very informative descriptors, accounting for configuration statistics, were derived by Duffy and Jorgensen[23] from Monte Carlo (MC) simulations and quantum mechanical calculations. Their HB and hydrophobicity descriptors showed good correlations with the solvation free energies for a diverse data set of drug-like and lead-like compounds. Another complex descriptor, the screening charge density, which is available in COSMO-RS developed by Klamt,[24] is based on quantum mechanical calculations and sophisticated treatment of charges at molecular surfaces. Our previous study[25] and the present study follow the same idea of designing a few smart descriptors capable to reproduce rather accurately specific variations in solvation free energies and other interphase transfer properties.

All the aforementioned approaches to predict solvation properties are reasonably accurate and applicable in general cases. Although very accurate, the Abraham's LSER method[3]

* Corresponding author phone: +1 (352) 392-0554; fax: +1 (352) 392-9199; e-mail: katritzky@chem.ufl.edu.
[‡] Moscow State University.
[§] Pfizer Global Research & Development.
[||] University of Florida.

10.1021/ci800323q CCC: $40.75
Published on Web 02/20/2009

Table 1. Original Electronegativity and Hardness Parameters and Calculated Intrinsic HB Basicity Values for Atoms in Various Valence States

atom type	χ_{lp}	η_{lp}	χ_σ	η_σ/χ_π	B_{in}
$F_{te^2te^2te^2te}$	18.16[a]	8.18[a]	13.24[a]	8.63[a]/−	3.68
$Cl_{te^2te^2te^2te}$	12.27[a]	5.46[a]	10.52[a]	5.61[a]/−	2.43
$Br_{te^2te^2te^2te}$	11.36[a]	4.51[a]	9.03[a]	4.73[a]/−	1.79
$I_{te^2te^2te^2te}$	10.35[a]	3.96[a]	8.66[a]	4.46[a]/−	1.52
$N_{te^2tetete}$	4.57	9.58	11.49	7.456/−	20.08
$N_{tr^2trtrp_\pi}$	5.34	9.769	12.82	7.8/7.9	17.87
$N_{trtrtrp_\pi^2(Pyrrole)}$	3.585[b]	8.375[b]	12.26	7.466/−	19.57
$N_{di^2dip_\pi p_\pi}$	7.36[b]	9.88[b]	15.59	8.33/7.8	13.26
$O_{te^2te^2tete}$	7.88	10.825	15.25	9.15/−	14.87
$O_{tr^2tr^2trp_\pi}$	8.96	11.19	17.06	9.6/10.1	13.98
$O_{tr^2trtrp_\pi^2(Furan)}$	8.765[b]	10.875[b]	16.73	9.42/−	13.49
$S_{te^2te^2tete}$	6.59	7.045	10.14	5.36/−	7.53
$S_{tr^2tr^2trp_\pi}$	7.044	7.3	10.89	5.45/7.7	7.57
$S_{tr^2trtrp_\pi^2(Thiophene)}$	7.24	7.09	10.88	5.39/−	6.94
$S_{te^2teteted_\pi(DMSO)}$	6.635[b]	7.005[b]	11.92	6.21/−1.1	7.40

[a] Taken from Oliferenko et al. *J. Phys. Org. Chem.* **2001**, *14*, 355−369. [b] Values are calculated from valence state energies of respective cations, ref 33.

and similar experiment-based methods[4−8] are intrinsically confined by the limited set of experimentally studied compounds. General QSPR approaches based on the whole-molecule theoretically derived descriptors do not have such intrinsic applicability problems, but as was mentioned above, they are less sensitive and, therefore, less accurate compared to empirical LSER techniques. By contrast, descriptors derived from MC simulations were intentionally designed to relate to HB and hydrophobic contacts. They are very effective in predicting solvation properties of complex drug-like and lead-like molecules, but at the same time, they are computationally expensive and, thus, difficult to calculate for large data sets. The latter is also true for COSMO-RS calculations.

On the basis of these facts, we recognize the large unmet need for computational drug design in fast-computed, theoretically derived HB descriptors that accurately predict HB capacity of heteroatoms and ionic centers. The first paper in this series[25] reported a computational formalism applicable to aliphatic compounds. We now extend the formalism to aromatic, conjugated and heterocyclic systems as well as to charged and zwitter-ionic species. Our present aim is to design accurate HB acidity and basicity descriptors and to provide theoretical background for calculation of solvation properties that we outline in our "universal solvation equation" (USE). We believe that USE, applicable to aliphatic, aromatic, and heterocyclic compounds in their neutral as well as ionic forms (including zwitterions) will be useful in creation of powerful QSPR models to predict a large number of physical properties, ADMET end points of interest, and in the development of more accurate protein−ligand docking applications.

2. METHODOLOGY

2.1. Hydrogen Bond Basicity of Neutral Molecules. *2.1.1. Intrinsic Basicity.* A detailed discussion on heteroatom hydrogen bond basicity is given in our previous HB paper published in 2004 in this Journal.[25] It was assumed that the intrinsic ability of a heteroatom to donate and to retain the electron density (designated as B_{in}) depends inversely on the electronegativity and directly on the hardness of its lone pair. In this work we take squared hardness reduced by electronegativity (eq 1) to put the HB basicity on the energy scale.[26] Also eq 1 may be thought of as a function inverse to the electrophilicity index suggested by Parr and co-workers.[26]

$$B_{in} = \frac{\eta_{lp}^2}{\chi_{lp}} \tag{1}$$

Here, χ_{lp} means the lone pair electronegativity, and η_{lp} is the corresponding hardness value. Parameters of the method are the orbital electronegativity (EN) and hardness values proposed by Hinze and Jaffe[27] and then further developed by Bergmann and Hinze.[28]

2.1.2. Molecular Environment. Influence of the molecular environment is accounted for by a simple function of distance (topological distance as a first approximation) weighted by electronegativities of neighboring atoms. These through-bond sigma contributions to hydrogen bond basicity are written as follows:

$$B_\sigma = \frac{1}{2}\sum_i^N \frac{\chi_{\sigma(i)}}{d_{(i)}^2}, \quad \begin{cases} \chi_{\sigma(i)} = \chi_{\sigma(i)}, & i \equiv \mathrm{Csp^3} \\ \chi_{\sigma(i)} = 0, & i \equiv \mathrm{Csp^2, Csp} \\ \chi_{\sigma(i)} = -\chi_{\sigma(i)}, & i \equiv \text{heteroatom} \end{cases} \tag{2}$$

where $\chi_{\sigma(i)}$ and $d_{(i)}$ are the electronegativities of the bonded orbital and the effective topological distance, respectively. For parametrization, the effective topological distance for the atom neighboring the basic site is set at 1.2. Subsequent distances are calculated as $d_{(i>1)} = d_{(i)}^0 - 0.5$, where $d_{(i)}^0$ is the number of bonds in a shortest walk between the basic center and the *i*-th atom. In eq 2, the summation runs over all atoms positioned at particular topological distances from the *i*-th basic center, with *N* being the number of atoms. Contributions from heteroatoms are assumed to be negative because of the electron-withdrawing or the $-I$-effect. The saturated carbon atom contributes positively because of its electron-donating or the $+I$-effect, while the influence of unsaturated carbon atoms is taken neutral. Effective atomic basicity, *B*, corrected for molecular environment is taken as the sum of the intrinsic value and the through-bond contributions:

$$B = B_{in} + B_\sigma = \frac{\eta_{lp}^2}{\chi_{lp}} + \frac{1}{2}\sum_i^N \frac{\chi_{\sigma(i)}}{d_{(i)}^2} \tag{3}$$

The total basicity of a molecule is calculated as the sum of effective atomic basicities of all basic centers. Electronegativity and hardness parameters for atoms in different valence states along with the calculated intrinsic components are given in Table 1.

The incorporation of aromatic, heterocyclic, and complex multifunctional compounds into the computational formalism is a challenging task because of the multitude of different resonance forms violating common basicity patterns. We found that considering several resonance forms (the resonance correction) significantly improves correlations. A simple way to calculate the resonance correction was determined to be (eq 4) the ratio of the π-orbital electronegativity of an electron-withdrawing atom, χ_π^A, and the lone pair (or π-orbital) electronegativity of an electron-donating atom, χ_{lp}^D (or χ_π^D)

$$k = \frac{\chi_{\pi}^{A}}{\frac{2}{n_{lp}^{D}} \cdot \chi_{lp,\pi}^{D}} \quad (4)$$

where n_{lp}^{D} is the effective quantum number of a lone pair-donating atom. Chemical groups needing the correction were found to have $k > 1.3$ for lp/π and $k > 5$ for π/π-systems, respectively. These are the amide, nitro, cyanamide, phosphaneoxide, sulfoxide, and haloanhydride groups.

Common resonance situations include either bond order alteration or charged resonance forms or both; we therefore consider resonating atoms and groups in terms of valence schemes. The percentage of the resonance forms for a basic center, r, is obtained (eq 5) as the reduced sum of weighted π-bond orders of all bonds adjacent to the basic center:[29]

$$r = \sum_{i} \frac{b_i}{\frac{1}{n}\sum_{j=1}^{N} b_j} \quad (5)$$

where b_i is the π-bond order of the i-th bond and n is the number of multiple bonds at the basic center. The summation is performed over all bonds of the basic center in the conjugated system. For a double bond, the π-bond order is taken as is. π-Bond orders were calculated with the SHMO program designed by Rauk;[30] however, any simple Hückel method implementation can be used. The Coulomb α and the resonance β integrals chosen empirically for some complex bonding situations are given in Table 1S of the Supporting Information.

2.1.3. Corrections for Multiple Resonance Forms. Resonance-corrected intrinsic basicity is formed by summation of basicities of neutral and resonance forms weighted (eqs 6 and 7) by their percentage contributions.

$$B_{in}^{corr} = B_{in} \cdot r + B_{in(A^-)} \cdot (1 - r)$$

for basic centers accepting electrons from other atoms (6)

$$B_{in}^{corr} = B_{in} \cdot (1 - r)$$

for basic centers donating electrons to other atoms (7)

Here $B_{in(A^-)}$ is the intrinsic basicity of a negatively charged resonating atom. Although resonance states are not observable, we find it useful to deduce parameters for some of them from physical grounds. Physically related parameters would be electronegativities and hardnesses of the respective atomic states, for example, O^-, N^-, N^+, S^-. HB basicity of a positively charged atom is assumed effectively equal to zero. Derivation of the basicity of negatively charged atoms needs special parametrization. Because of the lack of negatively charged valence states in the tabulation of Hinze and Jaffe,[27] we choose an alternative way. We fit basicity values of the anions OH^-, NH_2^-, and SH^-, the simplest patterns of the corresponding resonating states, using a trial-and-error method to incorporate them into the general training set. This strategy known as back-calculation, was successfully employed by Abraham et al.[31] and by Leahy et al.[32] in their developments of predictive LSER equations. A brief outline of this method is as follows. Let us assume an equation to correlate a given property with n variables (molecular descriptors) and a molecule with a known property and ($n - 1$) descriptors, then the given property can be calculated by substituting the known values in the equation and solving the latter for the unknown variable. Reliable solubility data are available for anions in aqueous solutions. The Gibbs free energies, $\Delta G_s^*(A^-)$, were recalculated into logarithmic partition coefficients, log $L_{w/g}$, and the data sets containing neutral and ionizable molecules were combined. The best correlation ($R^2 = 0.993$) for a set containing OH^-, NH_2^-, SH^-, as well as 219 neutral molecules, was achieved when B_{in} of these anions were equal to 610, 550, and 430 basicity units, respectively. This operation gives basicity values of charged species 1 order of magnitude larger than those of neutral molecules. Since substituting them in eq 6 overestimates the intrinsic basicity of a negatively charged resonating atom by approximately 10-fold, the final $B_{in(A^-)}$ values are reduced by 10. The π-bond orders and corrected basicity values calculated for various groups are shown in Figure 1 and in more detail in Table 2S of the Supporting Information.

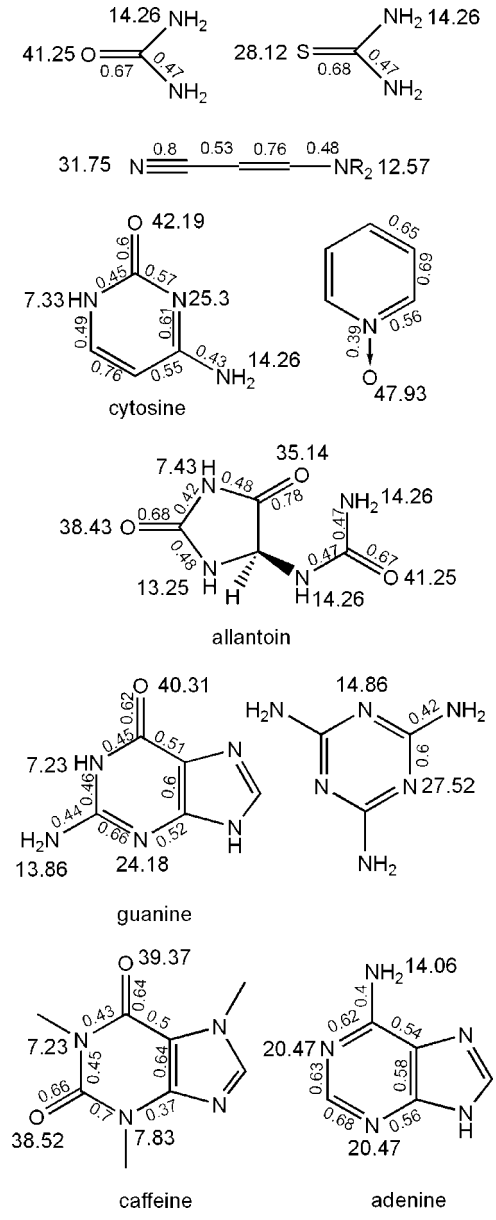

Figure 1. π-Bond orders and corrected intrinsic basicity values calculated for typical HB acceptor centers.

We note that some chemical functional groups need special treatment. For the nitro group, original bond orders were used, and only one oxygen atom was taken into account. No correction is applied if the overall corrected basicity is

smaller than the initial one like, for example, in the cases of phenol and pyrrole. In resonance situations an increase in electron density at one basic site is not always accompanied by an equal decrease at a less electronegative heteroatom. Thus corrected molecular basicity turns out to be even lower (e.g., pyrrole) or remains the same (e.g., imidazole).

2.2. Hydrogen Bond Basicity of Anions. Molecular environment effects are also applicable to anions (which are treated here as deprotonated acids), although the signs of such influences may be reversed. The effect of alkyl substituents on an anionic center is now negative as the carbon's EN is higher than that of any negatively charged atom. On the other hand, the σ-influence of an anionic center itself on neighboring atoms vanishes as the EN of a negatively charged atom is close to zero. Uncharged heteroatoms are treated as described earlier by eqs 1−3. In case when the oxygen bears a fractional formal charge (ClO_4^-, NO_3^-, HCO_2^-) or is attached to a carbon atom ($AlkO^-$, ArO^-), its basicity is calculated as the sum of B_{in} values of constituent heteroatoms corrected for the squared root of the number of contributing resonance forms. The above considerations can be summarized in the eqs 8−10.

$$B_{(A^-)} = 10 \cdot (B_{in(A^-)} + B_{\sigma(A^-)}) \quad (8)$$

$$B_{(A^-group)} = \frac{1}{\sqrt{N}} \cdot \left(B_{(A^-)} + \sum B_{(ar)}\right) \quad (9)$$

$$B = B_{(A^-)} + \sum_i B_i \text{ or } B = B_{(A^-group)} + \sum_i B_i \quad (10)$$

Here, $B_{(A^-)}$ is the basicity of an anionic center in saturated molecules, and $B_{\sigma(A^-)}$ is the basicity caused by the influence of aliphatic environment calculated by eq 2. $B_{(ar)}$ is the basicity of heteroatoms conjugated with an anionic center, $B_{(A^-group)}$ is the basicity of a group of heteroatoms sharing a negative charge, and N is the number of resonance forms. Finally, the total basicity of an anion, B, is calculated as the sum of basicities of all anionic and neutral basic centers.

2.3. Hydrogen Bond Acidity of Neutral Molecules. HB acidity is not merely a property of the hydrogen atom but also, to a large extent, of the atom to which it is bonded. Here, HB acidity defined in terms of electronic characteristics of neighboring atoms, and "acidic center" means any atom to which a hydrogen atom is attached. Specifically, it is formulated as the ratio of squared bonded orbital electronegativity (as opposed to lone pair EN in the case of basicity) and bonded orbital hardness.

$$A_{in} = \frac{\chi_\sigma^2}{2\eta_\sigma} \cdot \frac{1}{\alpha} \quad (11)$$

where χ_σ and η_σ are the electronegativity and hardness of a σ orbital of the atom bonded to the acidic hydrogen. Atomic polarizability, α, was introduced to increase the A_{in} variation within one row of the Periodic Table and to discriminate atoms that belong to different rows.

Polar influences exerted by molecular environment are accounted for in a manner similar to the basicity treatment. Opposite to basicity, alkyl substituents influence acidity negatively because of their electron-donating effect, whereas most heteroatoms withdraw electron density from an acidic center and increase its strength. Thus, the HB acidity term encompassing a σ-bonded molecular network is written (eq 12) as follows:

$$A_\sigma = \frac{1}{2}\sum_i^N \frac{\chi_{\sigma(i)}}{d_{(i)}^2} \begin{cases} \chi_{\sigma(i)} = -\chi_{\sigma(i)} & i \equiv Csp^3 \\ \chi_{\sigma(i)} = 0 & i \equiv Csp^2, Csp \\ \chi_{\sigma(i)} = \chi_{\sigma(i)} & i \equiv \text{heteroatom} \end{cases} \quad (12)$$

where $\chi_{\sigma(i)}$ and $d_{(i)}$ are the σ orbital electronegativity of i-th substituent atom and the effective topological distance between the hydrogen and the i-th atom, respectively. Intrinsic acidity and contributions from all atoms form the final acidity value A of an acidic center.

$$A = A_{in} + A_\sigma = \frac{\chi_\sigma^2}{2\eta_\sigma\alpha} + \frac{1}{2}\sum_i^N \frac{\chi_{\sigma(i)}}{d_{(i)}^2} \quad (13)$$

Total acidity of molecules having multiple acidic centers is composed by summing up contributions from all acidic sites.

The resonance correction applied above for the treatment of HB basicity is found useful in case of acidity for bonding situations where the acidic hydrogen atoms and high level of resonance are involved. Predominantly these occur for primary and secondary amides. The intrinsic acidity is corrected (analogously to Eqs 6 and 7) by using Eq. 14:

$$A_{in}^{corr} = A_{in} \cdot (1 - r) + 0.5 \cdot A_{in(BH^+)} \cdot r \quad (14)$$

Here, $A_{in(BH^+)}$ is the intrinsic acidity of a negatively charged resonating atom, and r is calculated according to eq 5.

2.4. Hydrogen Bond Acidity of Cations. Derivation of the HB acidity of cations (here by "cation" we mean a protonated base) is less difficult compared to that for anions, because EN and hardness values are readily available from electronic characteristics of positively charged valence states.[33] The acidity of a positively charged species is computed similarly to that of neutral species except that the σ-influence of molecular environment is accounted for by eqs 15 and 16 in a manner similar to that of anions.

$$A_{in(BH^+)} = \frac{\chi_\sigma^2}{\eta_\sigma} \quad (15)$$

$$A = 10 \cdot \frac{1}{\sqrt{N}} \cdot (A_{in(BH^+)} - A_{\sigma(BH^+)}) \quad (16)$$

Here, N is the number of resonance forms available for the acidic center. The σ-influence of molecular environment, $A_{\sigma(BH^+)}$, is calculated by eq 12 in a manner similar to that for neutral molecules. Respective σ-orbital EN and hardness parameters for neutral and positively charged atoms are listed in Table 2.

2.5. Universal Solvation Equation. The LFER/LSER approach assumes that any complex free energy-related property can be expressed as a linear combination of simpler properties governed by changes of free energy of some elementary processes. Mimicking elementary acts of general intermolecular interactions and even specific molecular recognition acts can be achieved through using appropriate molecular descriptors. This is typical QSAR/QSPR methodology that inherited a lot of features from chemometrics.[34] Normally dozens or even hundreds of descriptors are available for designing multilinear QSAR/QSPR models,[35] and the best models are selected based primarily on statistical merits. However, in the LSER-related studies, both experimental and theoretical, it has become possible to identify limited, fixed sets of descriptors, which reflect major

Table 2. Original Electronegativity and Hardness Parameters and Calculated Intrinsic HB Acidity for Different Acidic Centers

atom type	χ_σ	η_σ/χ_π	α^b, A^3	A_{in}
$F_{te^2te^2te^2te}$	13.24[c]	8.63[c]/−	0.296	34.31
$Cl_{te^2te^2te^2te}$	10.52[c]	5.61[c]/−	2.315	4.26
$Br_{te^2te^2te^2te}$	9.03[c]	4.73[c]/−	3.013	2.86
$I_{te^2te^2te^2te}$	8.66[c]	4.46[c]/−	5.415	1.55
$N_{te^2tetete}$	11.49	7.456/−	0.964	9.18
$N_{trtrtrp^2_{\pi(Pyrrole)}}$	12.26	7.466/−	1.03	9.77
N_{nitro}	18.84[a]	8.78[a]/−	n/a	n/a
$O_{te^2te^2tete}$	15.25	9.15/−	0.637	19.95
$S_{te^2te^2tete}$	10.14	5.361/−	3.00	3.20
$S_{teteteted_\pi d_\pi}$	13.64	6.95/1.27	n/a	n/a
$S_{ohohohohoh(SF6)}$	9.52	6.39/−	n/a	n/a
$C_{tetetete}$	8.06	6.565/−	1.061	n/a
$C_{trtrtrp_\pi}$	8.87	6.77/5.67	1.352	n/a
$C_{d_id_ip_\pi p_\pi}$	10.47	6.972/5.73	1.283	6.13
$N^+_{trtrtr_\pi}$	24.855[d]	9.765[d]/−	n/a	63.26
$O^{+_2}_{te^2tetete}$	29.505[d]	10.805[d]/−	n/a	80.57
$O^{+_2}_{tr^2trtrp_\pi}$	31.32[d]	11.17[d]/−	n/a	87.82
$S^{+_2}_{te^2tetete}$	20.65[d]	7.01[d]/−	n/a	60.83
$P_{teteteted_\pi}$	11.7	6.01/1.779	n/a	n/a
$P^{+_+}_{tetetete}$	18.1[d]	6.01[d]/−	n/a	54.51

[a] Electronegativity (hardness) of the nitrogen σ-orbital is averaged over the respective values of ionic and neutral forms. [b] Here α is an additive atomic polarizability: Miller, K. J. *J. Am. Chem. Soc.* **1990**, *112*, 8533−8542. [c] Taken from Oliferenko et al. *J. Phys. Org. Chem.* **2001**, *14*, 355−369. [d] Values are calculated from valence state energies of respective cations, ref 33.

contributions to solvation energies.[4,6,9,25] The well-known solvation equation developed by Abraham[4] includes five experimentally derived descriptors

$$\log SP = c + rR_2 + s\pi_2^{\mathbf{H}} + a\alpha_2^{\mathbf{H}} + b\beta_2^{\mathbf{H}} + v\mathbf{V}_x^{\mathbf{H}} \quad (17)$$

where R_2 is the excess molar refraction, $\pi_2^{\mathbf{H}}$ is the solute dipolarity−polarizability, $\alpha_2^{\mathbf{H}}$ and $\beta_2^{\mathbf{H}}$ are the solute HB acidities and basicities, respectively, and $\mathbf{V}_x^{\mathbf{H}}$ is the McGowan characteristic volume.[36] SP means some solvation-related property, and the regression coefficients (r, s, a, b, and v) are not just statistically fitted constants but rather coefficients obtained as a result of a large-scale training process.

As shown previously, theoretical counterparts of the descriptors comprising eq 17 have been identified and have produced successful correlations. It is noteworthy, that our descriptors A and B described above can be considered as corresponding to the Abraham's α_2 and β_2 parameters. A bilinear correlation to predict the gas−water partition coefficients, log $L_{w/g}$, with only two of our parameters, A and B, gives $R^2 = 0.72$ for a data set consisting of 219 compounds. The role of these two descriptors in predictions of the octanol−water partition coefficient, log $P_{o/w}$, and aqueous solubility, log S_w, is more modest: the squared correlation coefficients R^2 is equal 0.35 and 0.13 for two data sets consisting of 516 and 288 compounds, respectively.

To understand the whole complexity of intermolecular interactions in a solution requires to introduce into the equation additional terms similar to those present in the Abraham's solvation equation. We propose using five additional descriptors in such an equation. As a measure of molecular volume we use the semiempirically calculated α polarizability as the principal contribution and two more adjuvant descriptors: the hydrophobic term H and the steric correction S. Polarity P calculated as the sum of bond polarizabilities can be related to the solute dipolarity-polarizability term, $\pi_2^{\mathbf{H}}$. Finally, the π-energy E_π calculated by the Hückel method can be considered as a measure of the excess molar refraction, R_2.

Table 3. Polarizabilities of Various Types of Bonds and Functional Groups

bond	b_L	bond	b_L	bond	b_L	bond	b_L
C−F	1.25	C_{ar}−N	1.5	S−F	1.5	P→O	2.0
C_{ar}−F	0.75	C−S	1.7	P−O	0.86	C−CN	3.6
C−Cl	3.2	C_{ar}−S	4.74	N−P	1.0	C_{ar}−CN	5.75
C_{ar}−Cl	4.2	$C-S_{thioph}$	1.47	C_{ar}−P	1.2	$C-NO_2$	3.3
C−Br	5.0	$C-N_{pyrrol}$	0.75	C═O	2.3	$C_{ar}-NO_2$	5.5
C_{ar}−Br	6.3	$C-O_{furane}$	0.25	C═N	0.71	C_{ar}−N═O	4.5
C−I	6.75	O−O	0.62	C═S	2.6	$C_{ar}-CO_2$	5.85
C_{ar}−I	9.2	N−N	0.62	N═O	2.0	SO_2	3.0
C−O	0.8	N−O	0.56	S═O	1.5	CSO_2	3.1
C_{ar}−O	1.4	S−S	4.66	N═N	2.85	C_2SO_2	3.55
C−N	0.57	S−N	2.32	N→O	2.1	O_3PO	4.16

In such a way the Universal Solvation Equation, USE, written in the symbolic form includes seven descriptors and an intercept.

$$\log SP = \text{const} + A + B + \alpha + P + H + S + E_\pi \quad (18)$$

Coefficients of eq 18 are found by the method of multiple linear regression. The terms entering eq 18 (except for A and B already discussed above in this paper) are described below.

2.5.1. Polarizability. The total molecular polarizability α has been calculated at the AM1 semiempirical level of theory using MOPAC. Atomic units have been used (bohr^3).

2.5.2. Polarity. Polarity of a molecule, P, is defined as the sum of polarizabilities of all bonds bearing at least one heteroatom

$$P = \sum_i b_{L(i)} \quad (19)$$

Here $b_{L(i)}$is the polarizability of the i-th bond along its direction. Bonds formed by carbon atoms only, or containing hydrogen atoms, are not taken into account. Polarizability values for various bonds types taken from the compilation provided by Vereshchagin[37] are listed in Table 3. Descriptor P is especially suitable in improving correlation results of polyhalogenated and other highly polar compounds, otherwise found to be strongly underestimated.

2.5.3. Hydrophobic term. This descriptor accounts for saturated alkyl chains in a molecule that matches two requirements: (i) no neighboring heteroatoms (with the exception of the carbon atom bonded to the nearest heteroatom) and (ii) not lying in a path joining two or more heteroatoms. The hydrophobic term, H, consists of two parts. The first one considers the type of branching, while the second one accounts for the length of the alkyl group. Both terms are sensitive to the type of the heteroatom nearest to a hydrocarbon chain. The general expression for H is given in eq 20.

$$H = \sum_i n^*_{C_i} \frac{\chi_C}{\chi_{A_i}} + 2(N_C - 1)\frac{\chi_C}{\chi_A} \quad (20)$$

$$n^*_{C_i} = \frac{2\chi_H}{\sigma_i} - \chi_C = \frac{2 \cdot 7.18}{\sigma_i} - 8.06 \quad (21)$$

Here, χ_C, χ_H, and χ_A are the σ-electronegativities of a carbon atom, a hydrogen atom, and a heteroatom nearest to the i-th

carbon atom, respectively. Symmetry numbers σ_i for the *i*-th carbon atoms are defined in the same manner as it was done by Yalkowsky et al.[38] Thus, σ_i for di-, tri-, and tetrasubstituted carbons is 2, 3, and 12, respectively. N_C is the length of the longest unbranched hydrocarbon chain. The summation in the first term runs over all carbon atoms in a chain or in a ring containing only saturated carbon atoms. The $n^*_{C_i}$ values for monosubstituted carbon atoms and for that in methane are put equal to 1 and 3, respectively. The χ_A value is taken for the heteroatom nearest to the hydrocarbon group; but, if there are two or more equidistant heteroatoms, the most electronegative is taken.

2.5.4. Steric Correction. This descriptor was designed primarily for log $L_{w/g}$ correlation, but it was shown also to be useful for log $P_{o/w}$ predictions. However, the correlation results were highly overestimated for almost all ethers, esters, furans, and tertiary amines. Here a simple empirical correction is introduced that leads to lower solubility of fully substituted basic centers. This steric correction, *S*, is derived from σ-orbital electronegativities of basic centers, χ_{A_i}, and is computed as the sum of atomic contributions, s_i, from all totally substituted aliphatic basic centers in a molecule: $S = \sum s_i$; where

$$s_i = \chi_C - \chi_{A_i} = 8.06 - \chi_{A_i} \quad (22)$$

The *S* descriptor tends to combine two effects. First, it is a simplified way to account for the influence of steric hindrance in fully alkyl-substituted heteroatoms on their ability to accept hydrogen bonding. Second, it may be an indirect way of accounting for the lowering of the heteroatom basicity due to the involvement into resonance conjugation. The steric correction values for atom types of interest are as follows: −7.19 for O_{sp^3}; −3.43 for N_{sp^3}; −2.08 for S_{sp^3}; −8.67 for O_{furan}; −4.2 for $N_{pyrrole}$; −2.82 for $S_{thiophene}$.

2.5.5. π-Energy. Resonance involving π-electrons influences both molecular polarizability and the weak π-basicity because of H−π interactions. This can be reflected to some extent by the total energy of the π-system or by the resonance energy of π-electrons. The total π-energy, E_π, is calculated for specifically selected conjugated parts of molecules, where all sp^2- and sp-hybridized heteroatoms are treated as respective carbon atoms. Heteroatoms that have other types of hybridization are not counted. In cases where several isolated unsaturated fragments are present in a molecule, E_π is calculated simply by summing the π-energy values of the fragments.

3. COMPUTATIONAL DETAILS

The overall data set comprised 525 diverse small organic molecules including mono-, di-, and polyfunctional aliphatic and aromatic species, heterocycles, amino acids, nucleotides, and pharmaceuticals. Molecular structures were drawn using a Hyperchem 7 software[39] and subsequently optimized at the AM1[40] semiempirical level of theory by the eigenvector following algorithm. The MOPAC 6.0 semiempirical program was used.[41] A gradient norm of 0.01 kcal/Å was forced along with keyword PRECISE to calculate stationary points. Polarizability was calculated on the same AM1 wave function using the finite field method.[42] All π-electronic calculations in the Hückel approximation including energy and density matrix were done with a SHMO2 program.[30] Multiple linear regressions were derived with Statistica 6 software.[43]

Table 4. pK_{HB} Values Reproduced with *B* and E_π Descriptors

compound	*B*	E_π	pK_{HB} exptl	pK_{HB} calcd
fluoromethane	6.48	0	0[a]	−0.14
chloromethane	5.23	0	−0.3[a]	−0.25
bromomethane	4.59	0	−0.3[a]	−0.30
iodomethane	4.31	0	−0.4[a]	−0.33
water	14.87	0	0.64[a]	0.56
methanol	17.67	0	0.82[a]	0.79
dimethyl ether	20.47	0	1[a]	1.02
tetrahydropyrane	24.69	0	1.23[a]	1.38
tetrahydrofuran	24.05	0	1.28[a]	1.32
triethylamine	33.85	0	1.99[a]	2.14
dimethylsulfane	13.13	0	0.3[a]	0.41
ethene	0	−2	−0.8[a]	−0.62
cyclohexene	0	−2	−0.82[a]	−0.62
benzene	0	−8	−0.5[a]	−0.43
p-xylene	0	−8	−0.3[a]	−0.43
naphthalene	0	−13.68	−0.26[a]	−0.25
acetaldehyde	15.77	−2	0.65[b]	0.69
acetone	17.56	−2	1.2[a]	0.84
cyclododecanone	20.26	−2	1.23[b]	1.07
acetonitrile	16.3	−2	0.9[c]	0.74
benzaldehyde	13.98	−10	0.78[a]	0.80
benzophenone	13.98	−18	1.07[b]	1.05
benzonitrile	14.51	−10	0.8[c]	0.84
2-chloropyridine	15.57	−8	1.05[c]	0.87
pyridine *N*-oxide	41.39	−8	2.7[a]	3.02
triethylphosphane oxide	42.6	−2	3.16[a]	2.93

[a] Ref 5, Berthelot, M.; Besseau, F.; Laurence, C. *Eur. J. Org. Chem.* **1998**, 925−931. [b] Besseau, F.; Luçon, M.; Laurence, C.; Berthelot, M. *J. Chem. Soc., Perkin Trans. 2*, **1998**, 101−107. [c] Berthelot, M.; Laurence, C.; Safar, M.; Besseau, F. *J. Chem. Soc., Perkin Trans. 2*, **1998**, 283−290.

4. RESULTS

4.1. Hydrogen Bond Basicity. The variability range of the HB basicity scale extends from zero for hydrocarbons to 200 for [2,2,2]cryptand. The oxygen atoms of the deprotonated carboxyl groups in amino acids (58.7 for arginine) and the oxygen atom in DMSO (53.4) possess the highest local basicity values. Less basic are the oxygen atoms in the urea (37−39) and amide (30−34) fragments of such molecules as caffeine, uracil, cytosine and some alkyl amides. Equally strong hydrogen acceptors are the oxygen atoms in triphenyl phosphane oxide (35.3) and the nitrogen centers in quinuclidine and morpholine (34). Weaker HB acceptors can be ordered in the following manner:

$$\underset{22.9}{MeNH_2} > \underset{22.7}{Me_2SO_2} > \underset{17.7}{MeOH} > \underset{16.5}{PhSO_2NH_2} > \underset{16.3}{MeCN} > \underset{15.8}{MeCHO} > \underset{10.3}{MeSH} > \underset{9.8}{MeNO_2} > \underset{6.5-4.3}{MeHal}$$

This sequence corresponds to that found in the experimental study by Laurence and Berthelot.[4] These authors defined the site basicity as a log *K* value (pK_{HB}) for the process of 1:1 complexation of HB acceptors with 4-fluorophenol in carbon tetrachloride. Equation 23 produces an excellent fit ($R^2 = 0.982$) between the pK_{HB} and our two descriptors *B* and E_π for the 26 varied monofunctional molecules shown in Table 4.

$$pK_{HB} = -0.685 + 0.83B - 0.032E_\pi \quad (23)$$

This result supports the view that the total π-energy can be interpreted as the ability of a conjugated system to participate in hydrogen bonding as a weak electron density donor. However, some discrepancies occur. For example,

DMSO, aniline, phenol, pyrrole, and furan have higher basicity values, while pyridine is a weaker HB acceptor on our scale.

A satisfactory ($r = 0.851$) single-parameter correlation was found between B and the Abraham's hydrogen-bond basicity β_2^{H44} for a data set of 200 compounds.

4.2. Hydrogen Bond Acidity Scale. This scale varies from zero to almost 100 in amino acids. The highest acidity is achieved in the protonated nitrogen centers in amino acids (61−64), while OH-acids are three times less active hydrogen donors (20), and amines occupy the third position on the HB acidity scale (8−10).

Our hydrogen bond acidity descriptor A correlates with the Abraham's α_2^H with $r = 0.883$ for a data set consisting of 114 points. Correlation data both for β_2^H and α_2^H can be found in Table 3S of Supporting Information.

4.3. Application of the Universal Solvation Equation. *4.3.1. Octanol/Water Partitioning.* Log $P_{o/w}$ experimental data are collected for 516 organic molecules. The experimental data covers a range of 13 log units with a maximum value of 9.35 for 2,2′,3,3′,4,4′,5,5′-octachlorobiphenyl and a minimum value of −4.2 for arginine. Multiple regression analysis performed with six above-described USE descriptors yields a very good fit (eq 24) with $R^2 = 0.950$ and a root-mean-square deviation (rmsd) of 0.37 logarithmic units. These and all further correlation equations are given with autoscaled coefficients (calculated by subtracting the mean value and then dividing by the standard deviation).

$$\log P_{o/w} = 1.526 - 1.511B + 1.659\alpha + 0.428P + 0.273H - 0.150S + 0.502E_\pi \quad (24)$$

where $N = 516$, $R^2 = 0.950$, and rmsd = 0.37.

As can be seen from eq 24, the negative contribution of HB basicity B and the positive contribution of molecular polarizability α are most important. The contribution of E_π is three times smaller than that of B (the same ratio will be observed further in eq 26 for log $L_{w/g}$). This suggests that aromatic systems are about three times weaker HB acceptors than molecules with lone pairs-bearing heteroatoms. This is in accord with the theoretical calculations of electrostatic energies for benzene−water,[45] acetylene−water, and water−water[46] complexes, which are in the order −2.64, −4.63, and −9.18 kcal/mol, respectively. The polarity descriptor P favors the aqueous solvation and improves the correlation considerably. At the same time the hydrophobicity term H and the steric correction S exert less impact on the fit: leaving them out reduces R^2 from 0.95 to 0.93. The acidity descriptor A is statistically insignificant for octanol−water partitioning as its addition/removal does not alter the correlation. Compounds with predicted unsigned errors exceeding 1 log unit are pyridine *N*-oxide (−1.03), salicylic acid (1.21), codeine (−1.18), and corticosterone (−1.35). Removal of the main outliers does not change R^2 by much, which implies that the model is rather stable.

To provide a comparison with previous efforts, we selected from our general data set a subset of 110 organic molecules, which were challenging even for the Monte Carlo simulations-based descriptors developed by Duffy and Jorgensen.[23] A correlation equation was generated for this data set, and the fit of log $P_{o/w}$ values predicted versus observed was good with $R^2 = 0.94$, rmsd = 0.35, the strongest outliers being anthracene, pyridazine, and cytosine.

A challenge to the predictive ability of a model is an external test set. We compiled 57 diverse molecules (including six pharmaceuticals) and calculated the USE descriptors for each. Then the log $P_{o/w}$ values were computed for each test set compound using the calculated descriptors' values and the coefficients of eq 18. Finally, correlation of the predicted log $P_{o/w}$ with experimental values showed $R^2 = 0.93$ and rmsd = 0.42. Thus, the explanatory and predictive power of eq 24 is confirmed. The numerical data of all the experimental and predicted octanol−water partition coefficients along with the calculated descriptors are given in Table 5. All experimental and predicted log $P_{o/w}$ values for the training set and all subsets as well as the USE descriptors are given in Tables 3S and 4S of Supporting Information, respectively.

4.3.2. Gas/Water Partitioning (Aqueous Solvation Free Energy). A multiple regression model built on all seven descriptors for experimental log $L_{w/g}$ data for 219 compounds covering a range of 10.7 log units with the maximum value related to diclofenac (8.35) yields (Eq. 25) a satisfactory correlation.

$$\log L_{w/g} = 4.337 + 3.076B + 1.066A - 0.237\alpha + 0.139P - 0.251H + 0.840S - 1.030E_\pi \quad (25)$$

where $N = 219$, $R^2 = 0.95$, and rmsd = 0.50.

B, A, S, and E_π are the most important descriptors. Contributions from polarizability α and polarity P are much smaller, and their presence improves the fit only slightly and their removal leads to the five-parameter eq 26, which has almost the same statistics.

$$\log L_{w/g} = 4.305 + 3.020B + 1.058A - 0.321H + 0.868S - 0.827E_\pi \quad (26)$$

where $N = 219$, $R^2 = 0.951$, and rmsd = 0.51.

The strong outliers of eq 26 (prediction error more than 1.0 log unit) are *tert*-butyl chloride (−1.25), bromoform (1.43), ethylene glycol (−1.43), 4-phenylbutyric acid (−1.66), pyrazine (−1.23), and diclofenac (−1.18). Again, as in the case of log $P_{o/w}$, we tested our hydration energy model on 85 molecules taken from the study of Duffy and Jorgensen.[23] The fit was even better than in the previous study:[23] $R^2 = 0.942$ and rmsd = 0.50. The experimental and predicted gas−water partition coefficients and the corresponding descriptors for the training and all subsets are given in Tables 3S and 5S of the Supporting Information, respectively.

The solvation model represented by eq 26 was tested on a test set of 17 small molecules first proposed by Nicholls et al.[47] and subsequently used by Truhlar et al. for testing their SM8 model.[48] This test set was specifically designed for testing computational models of solvation. The USE descriptors were calculated for these 17 molecules (actually for 6 of them, because 11 of these molecules were already in our training set) and log $L_{w/g}$ values were computed by substituting the descriptors values into eq 26. The resultant logarithmic units were then recalculated into kcal/mol using the standard state definition (pressure 1 atm and temperature 298 K), which gave theoretical estimates of aqueous solvation free energy. All data, including experimental and calculated ΔG_{solv}, USE descriptor values, mean signed errors, mean unsigned errors, and rmsd, are given in Table 6S of the Supporting Information. Imidazole and *m*-bis(trifluorometh-

Table 5. Experimental and Predicted log $P_{o/w}$ Values and USE Descriptors for the Test Set

molecule	log P_{exptl}	log P_{calcd}	diff.	B	α	P	H	S	E_{π}
1,1,2-trichloroethane[a]	1.89	1.74	0.15	13.02	5.03	9.6	0	0	0
chlorotrifluoromethane[b]	1.65	1.62	0.03	0	2.71	6.95	0	0	0
tetrachloroethene[a]	3.4	3.42	−0.02	0	6.47	16.8	0	0	−2
2-methyl-2-butanol[a]	0.89	1.18	−0.29	23.69	6.45	0.8	4.12	0	0
hexafluoropropan-2-ol[a]	1.66	1.47	0.19	16.06	5.11	8.3	0	0	0
methyl *n*-propyl ether[a]	1.21	1.07	0.14	22.9	5.57	1.6	2.17	−7.19	0
1,3,5-trioxane[c]	−0.43	0.22	−0.65	43	5.15	4.8	0	−21.57	0
teflurane[a]	2.01	2.12	−0.11	5.18	4.57	10	0	0	0
2,2,2-trifluoroethyl-*N,N*-dimethylamine[d]	1.06	0.78	0.28	33.48	6.49	5.46	0.98	−3.43	0
bis-(2-hydroxyethyl)amine[c]	−1.43	−1.49	0.06	64.96	6.89	2.74	0	0	0
allylamine[a]	0.03	0.19	−0.16	22.88	4.66	0.57	0.98	0	−2
pentane-3-one[a]	0.82	1.20	−0.38	18.85	6.42	2.3	1.73	0	−2
3,3-dimethylbutane-2-one[a]	1.2	1.58	−0.38	19.49	7.43	2.3	2.87	0	−2
di-*tert*-butyl ketone[a]	3	2.71	0.29	21.43	10.87	2.3	4.85	0	−2
2-methyl-2-nitropropane[a]	1.17	1.36	−0.19	20.83	6.94	3.3	2.17	0	−2
3-methylbutanoic acid[a]	1.16	0.87	0.29	27.85	6.79	3.7	1.82	0	−2
methyl acrylate[a]	0.8	1.51	−0.71	25.77	7.71	4.5	0.95	−7.19	−2
methyl oxalate[e]	−0.17	0.13	−0.30	37.09	5.62	8.2	−0.28	−7.19	−4.47
N-methylurea[f]	−1.4	−1.50	0.10	56.06	4.65	5.87	0.49	0	−2
tetramethylurea[h]	0.19	−0.48	0.67	68.33	8.76	7.58	1.97	−7	−2
n-hexylbenzene[a]	5.52	5.74	−0.22	0	14.75	0	17.4	0	−8
1,2-diphenylethane[a]	4.79	4.30	0.49	0	17.65	0	0	0	−16
1-bromo-4-iodobenzene[e]	4.05	3.97	0.08	2.87	10.84	15.5	0	0	−8
2,2′,3,4,4′-PCB[h]	6.61	6.42	0.19	5.51	21.37	16	0	0	−16
bromomethylbenzene[a]	2.92	2.77	0.15	4.59	10.11	5	0	0	−8
N,N-dimethylbenzylamine[a]	1.98	1.90	0.08	28.48	12.31	1.71	0.98	−3.43	−8
2,4,6-trimethylphenol[a]	2.97	2.55	0.42	16.36	12.27	1.4	1.4	0	−8
trifluoromethylbenzene[a]	3.01	2.48	0.53	1.8	8.93	3.75	0	0	−8
4-chloro-2-methylphenol[a]	2.78	2.21	0.57	17.64	10.51	5.6	0.84	0	−8
3-cyanophenol[a]	1.7	1.35	0.35	27.37	10.20	7.15	0	0	−10.42
phenyl *n*-propyl ether[a]	3.18	2.64	0.54	20.11	12.19	2.2	1.89	−7.19	−8
3-methoxyaniline[e]	0.93	0.63	0.30	36.86	9.37	3.7	−0.28	−7.19	−8
3-nitrotoluene[a]	2.42	2.17	0.25	17.95	11.10	5.5	0.55	0	−10
3-nitrophenol[a]	2	1.29	0.71	30.63	10.61	6.9	0	0	−10
2-nitroacetanilide[f]	1	1.08	−0.08	57.14	14.60	10.8	0.67	0	−12
3,4-dinitrobenzoic acid[e]	1.51	1.84	−0.33	49.54	14.65	16.85	0	0	−14
3-chloroacetophenone[a]	2.51	2.34	0.17	17.71	11.45	6.5	0.22	0	−10
2-ethylphenyl acetate[a]	2.75	2.57	0.18	26.55	13.84	2.2	1.87	−7.19	−10
4-bromopyridine[a]	1.54	1.52	0.02	18.81	8.21	7.72	0	0	−8
2-methylpyrazine[a]	0.23	0.05	0.18	36.2	7.90	2.84	1.19	0	−8
5-quinolinol[e]	2.08	1.49	0.59	31.63	13.57	2.82	0	0	−13.68
2-aminoquinoline[e]	1.87	1.53	0.34	36.04	14.55	2.92	0	0	−13.68
acridine[a]	3.4	3.91	−0.51	17.91	20.17	2.84	0	0	−19.31
3-aminoacridine	2.77	3.40	−0.63	37	22.34	2.92	0	0	−19.31
4-isopropylbenzenesulfonamide[a]	1.75	2.09	−0.34	41.8	14.64	10.06	1.7	0	−12
benzoin[f]	2.13	2.36	−0.23	35.44	18.28	3.1	0	0	−18
N-amyl carbamate[f]	1.35	0.70	0.65	49.07	9.24	3.47	6.14	0	−2
tri-*n*-propyl phosphate[a]	1.87	2.08	−0.21	52.12	14.19	6.98	3.43	0	−2
citric acid[f]	−1.72	−1.50	−0.22	84.81	10.08	11.9	0	0	−6
Leu[h]	−1.52	−1.63	0.11	77.33	9.09	4.27	0	0	−2
thioguanine[h]	−0.07	0.87	−0.94	67.16	15.27	11.44	0	0	−9.45
nifuroxime[g]	1.28	1.62	−0.34	40.21	12.17	7.27	0	−8.67	−8.47
nitrofurazone[g]	0.23	0.77	−0.54	81.26	17.23	11.13	0	−8.67	−10.47
testosterone[h]	3.32	4.66	−1.34	42.97	23.17	3.1	0.95	0	−4.47
fluoxetine *	4.05	4.08	−0.03	49.12	24.03	7.09	0.49	−7.19	−16
hydrocortisone[g]	1.55	2.14	−0.59	103.36	26.84	7	0.56	0	−6.47
cimetidine[h]	0.4	0.43	−0.03	117.94	21.69	18.42	0.49	−2.08	−8.94

[a] Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C. *J. Pharm. Sci*. **1994**, *83*, 1085−1100. [b] Basak, S. C.; Gute, B. D.; Grunwald G. D. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054−1060. [c] J. Sangster. *Octanol−Water Partition Coefficients: Fundamental and Physical Chemistry*; John Wiley & Sons: Chichester, U.K., 1997. [d] Ran, Y.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 354−357. [e] Howard, Ph. H.; Meylan, W. M. *Handbook of Physical Properties of Organic Chemicals*; 1997. [f] Ran, Y.; Jain, N.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1208−1217. [g] Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F.; Abraham, M. H. *J. Med. Chem.* **2000**, *43*, 2922−2928. [h] C. Hansch, A. Leo, D. Hoekman. *Exploring QSAR. Hydrophobic, Electronic, and Steric Constants*; ACS Professional Reference Book; Amercian Chemical Society: Washington, DC, 1995. * www.chemfinder.com.

yl)benzene, molecules in which complex conjugation is coupled with strong polar influences, have largest errors ranging between two and three kilocalories per mole. Compounds with calculation errors more than 1 kcal/mol but less than 2 kcal/mol are *N*,*N*,4-trimethylbenzamide, 1,1-diacetoxyethane, and 1,4-dioxane. However, the whole data set's rmsd of the USE model (eq 26) is 1.11 kcal/mol, which is virtually the same as the best SM8 result[48] and even smaller than those of other implicit and explicit solvation models discussed in refs 47 and 48.

4.3.3. Aqueous Solubility. The aqueous solubility, log S_w, of a chemical is an important characteristic for the compound to be absorbed in the human intestine and one of the assays in pharmacokinetic profiling of drug candidates in pharmaceutical industry. Various theoretical approaches have been employed over the past decade: artificial neural networks,[49−51] the group-contribution approach,[52] and multiple linear regression approaches based on experimentally derived,[4] experimentally trained,[53] and theoretically calculated[54−56] descriptors. In the present attempt we apply our USE set of descriptors to predict the aqueous solubility log S_w. At the first step, we tried our calculated log $P_{o/w}$ as a single predictor of log S_w because a linear relationship between the aqueous solubility and the octanol−water partition coefficient was first noted by Hansch et al.[57] and then successfully used by Yalkowsky.[7] Because the overlap of log $P_{o/w}$ and log S_w data points in our general data set consists of 280 points, a correlation equation log S_w versus log $P_{o/w}$ was designed based on this amount of data, and the correlation results (eq 27) were

$$\log S_w = -1.047 \log P_{o/w} + 0.390 \quad (27)$$

where $N = 280$, $R = 0.901$, and rmsd = 0.91.

This correlation is satisfactory overall, but prediction errors may accumulate unacceptably in the two regression steps. Generally speaking, log S_w requires more complex theoretical treatment than log $P_{o/w}$, because the aqueous solubility depends on the crystal lattice energy of a solute, which is not easy to model or to account for in a simple way. Moreover, aqueous solubility differs for different crystal forms and amorphous states of the solute. Therefore, correlations of log S_w are commonly inferior to those of more straightforward free energy-related properties such as partition coefficients log $P_{o/w}$ and log $L_{w/g}$. Predictions of the aqueous solubility of complex organics and multifunctional drug-like molecules, which exist as solids under normal conditions, is a challenging and very important problem. In our attempt to model log S_w, we used a diverse training set consisting of 288 compounds. This data set covers the range of 12 log units and combines molecules of different structural classes, including hydrophobic and, hence, poorly soluble compounds (polychlorobiphenyls), zwitterionic amino acids, complex polyfunctional molecules, and 27 pharmaceuticals. A four-parameter regression (eq 28) was produced for this data set

$$\log S_w = -1.264 + 1.315B - 1.582\alpha - 0.607P - 0.387H \quad (28)$$

where $N = 288$, $R^2 = 0.892$, rmsd = 0.73.

The HB basicity term B and the cavity formation term α are the best predictors of log S_w as the fit shown has rather high statistical features ($R^2 = 0.793$ and rmsd = 1.02). Just as polarizability, the polarity term P contributes negatively to aqueous solubility, so does the hydrophobicity term H that reflects the presence of branched alkyl groups in the solute. There are six compounds with prediction errors more than 1.8 log units: trichloroacetic acid (2.07), 2-methylpyridine (1.92), and four nucleotide bases (the greatest error for guanine was 3.87). Also overestimated were solubilities of molecules containing the amide and the urea groups probably because they form strong intermolecular H-bonds that stabilize the crystal lattice. At the same time, predicted log S_w values for amines and pyridines were considerably lower than experimental ones. Removal of the four nucleotide bases improved the model by increasing the R^2 value from 0.892 to 0.923 and by decreasing the rmsd from 0.73 to 0.65. Leaving out two more strong outliers preserves the quality of the fit. Thus, the final USE model (eq 29) for log S_w is the following:

$$\log S_w = -1.197 + 1.427B - 1.609\alpha - 0.580P - 0.381H \quad (29)$$

where $N = 284$, $R^2 = 0.92$, and rmsd = 0.65.

Experimental and calculated log S_w values are given in Table 3S of the Supporting Information.

4.3.4. Aqueous Solvation of Ionic Species. Determination of aqueous solvation free energies ΔG_s^* of ions is a difficult problem partially because of the lack of experimental data and partially because of the relatively high estimated uncertainty of determination (2−5 kcal/mol). In addition, solvation free energies are 1 order of magnitude higher for ions. There are few theoretical works devoted to the prediction of the Gibbs energy of hydration of ions and many of them use rather high levels of theory such as Density Functional Theory (DFT) or Hartree−Fock.[58−60] In the literature, we found no information about attempted QSPR modeling of ionic solvation or any unified treatment of ionic and neutral species.

As we showed above, the intrinsic ionic basicities of the OH^-, NH_2^-, and SH^- ions were estimated from respective experimental ΔG_S^* values and a back-calculation procedure. Basicity and acidity values for other ionic species were calculated according to eqs 6 and 10. A regression model (eq 30) including these two descriptors was built on a data set of 50 experimentally available solvation energies of anions and cations, which showed an encouraging R^2 of 0.902 and rmsd of 4.26 kcal/mol (which is within the limits of experimental error for charged molecules).

$$\Delta G_S^* = -73.18 - 37.17B - 34.86A \quad (30)$$

An attempt of a unified neutral/charged treatment is represented by eq 33. This is a five-parameter equation which includes 269 log $L_{w/g}$ data points from both a neutral data set of 219 compounds and a data set of 50 ionic molecules. The separate neutral and ionic correlations (eqs 31 and 32) were derived from the existing data. For ions the log $L_{w/g}$ values were calculated from the corresponding data ΔG_S^* values. These correlations are shown in eq 33 to demonstrate numerical similarity and consistency of the variables (H and S descriptors are zeroed for ions). The high statistical quality of eq 33 for the combined data set ($R^2 = 0.996$ and rmsd = 1.36 kcal/mol) indicates that a unified QSPR treatment of neutral and ionic species is feasible.

$$N=219\text{(neutrals)} \quad \log L_{w/g} = 50.1 + 28.6B + 22.53A - 0.33H + 0.87S - 0.82E_\pi \quad (31)$$

$$N=50\text{(ions)} \quad \log L_{w/g} = 53.6 + 27.25B + 25.55A \quad (32)$$

$$N=269\text{(neutrals and ions)} \quad \log L_{w/g} = 53.6 + 28.26B + 26.70A - 0.30H + 0.80S - 0.90E_\pi \quad (33)$$

The smooth distribution of both neutral and ionic clusters can be seen in Figure 2, where a plot of experimental versus

calculated data is presented (an extended view of neutrals is given in the inset). The data on ionic species both experimental and calculated by eq 30 are given in Table 6 along with the calculated HB acidity and basicity values.

5. DISCUSSION

The correlations in the previous section clearly attest the viability and usefulness of the hydrogen bond and supplementary descriptors comprising the Universal Solvation Equation. They show that these USE descriptors enable the derivation of accurate explanatory and predictive models of aqueous solubility and interphase transfer. The universal character of our USE solvation model manifests itself in a consistent treatment of a structurally highly diverse data set, which also includes polyfunctional drug molecules, ionic species, and other difficult cases for any modeling. In principle, the USE descriptors cover the whole chemical space relevant to organic and medicinal chemistry while being easily calculated for any small- or medium-size molecule. Although just four solvation-related properties (log P, log L, log S_w, and ΔG^*_S) are discussed in the present work, similar equations can be derived for any solvation-related property given sufficient experimental data. Another feature which lends credit to the definition 'universal' is the consistency and tightness of the solvation equation: just seven descriptors, catch most of molecular complexity, regardless of the property being modeled. Variations in the coefficients of the equation implicitly reflect properties of the medium under study. This distinguishes our USE methodology from the common QSAR/QSPR techniques, where different correlation equations are generated from large descriptor pools for each particular task and hinder direct comparison and consistent interpretation.

Large descriptor pools are still critically needed in the quest of new knowledge at the beginning of any "structure−property" study or ligand-based design. They are used for outlining new chemical spaces and for revealing important structural and binding features.[61] However, once the key characteristics of a particular phenomenon are defined, the next logical step forward should be to create a stable model consisting of a few energy-related contributions, with each having a clear physical meaning.

In the current study, such a step resulted in the development of the set of advanced solvation descriptors, which while in the spirit of the linear free energy/solvation energy relationships (LFER/LSER), has a crucial difference in the way the descriptors are derived. Unlike LFER/LSER descriptors available only from experiment, the hydrogen bond scales A and B are derived from quantum chemical considerations and parametrized for better reproduction of experimental data. Although the derivation of A and B (eqs 13 and 3) may seem completely empirical, a closer look at these equations shows that the values of the intrinsic atomic property A_{in} (B_{in}) and of the molecular environment correction A_σ (B_σ) calculated for each atom in a molecule form a square matrix A_{ij} (B_{ij}), which can be viewed as a simplified task-specific Hamiltonian matrix. Analogies with the simple Hückel Hamiltonian are apparent. As shown in the Methodology section, the matrix elements are calculated for all relevant heteroatoms, π-atoms, acidic hydrogen atoms, and skeleton carbon atoms. Diagonal elements of such model Hamiltonians **A** and **B** represented by matrices A_{ij} and B_{ij} are filled by the intrinsic A_{in} and B_{in}, whereas off-diagonal elements consist of the distance-dependent terms A_σ and B_σ. In terms of quantum chemistry, the corresponding elements now designated as A_{ii} (B_{ii}) and A_{ij} (B_{ij}) can be thought of as Coulomb integrals H_{ii} and resonance integrals H_{ij}, respectively. Such analogies are justified because these matrix elements are functions of valence state ionization energies and internuclear separations. A respective eigenproblem could be solved, but this is out of the scope of this study. Currently a simpler formalism is employed, where the desired atomic acidity (basicity) values are calculated by summing up the rows of matrices A_{ij} (B_{ij}).

One could argue that such "smart" descriptors could easily be replaced by simpler characteristics such as the counts of hydrogen bond acceptors (HBA) and hydrogen bond donors (HBD). But we can show that in this case "simpler" is not

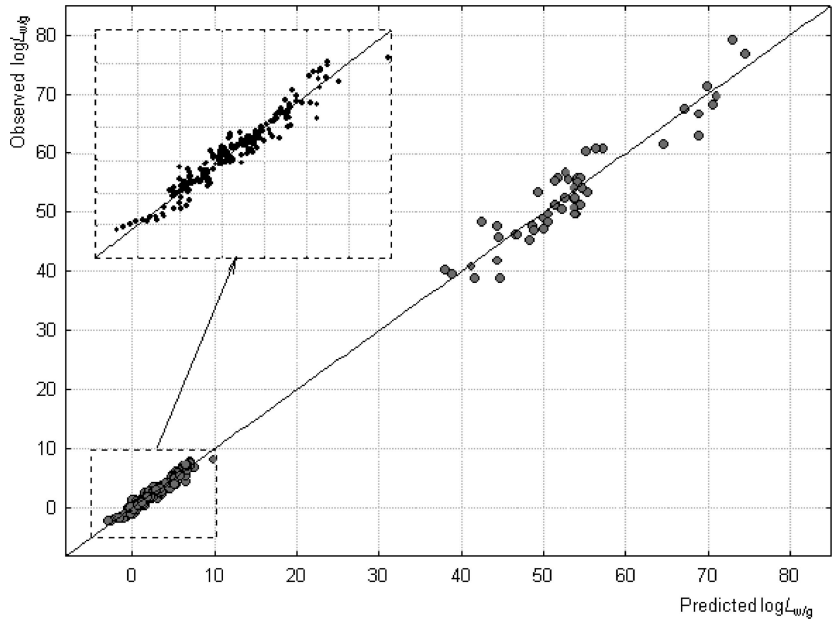

Figure 2. Experimental versus Predicted log $L_{w/g}$ values for a combined set of neutral and charged molecules.

Table 6. HB Basicity and Acidity Values of Calculation Results for Ionic Species (in kcal/mol)

neutral form	ionic form	*B*	*A*	ΔG^*_sexptl	ΔG^*_s calcd	diff
water[a]	OH^-	610.00	0	−105.0	−101.4	−3.6
hydrogen peroxide[a]	HO_2^-	571.92	0	−97.3	−95.3	−2.0
methanol[a]	CH_3O^-	582.01	0	−95.2	−96.9	1.7
ethanol[a]	$C_2H_5O^-$	564.1	0	−91.1	−94.0	2.9
t-butanol[b]	$(CH_3)_3CO^-$	528.28	0	−84.0	−88.3	4.3
phenol[a]	$C_6H_5O^-$	431.34	0	−71.3	−72.6	1.3
4-chlorophenol[b]	$ClC_6H_4O^-$	431.93	0	−68.0	−72.7	4.7
ammonia[a]	NH_2^-	550.00	0	−92.2	−91.8	−0.4
aniline[b]	$C_6H_5NH^-$	388.91	0	−65.0	−65.8	0.8
diphenylamine[b]	$(C_6H_5)_2N^-$	317.54	0	−56.0	−54.3	−1.7
sulfane[a]	SH^-	430.00	0	−71.6	−72.4	0.8
formic acid[a]	HCO_2^-	441.22	0	−76.2	−74.2	−2.0
acetic acid[a]	$CH_3CO_2^-$	429.82	0	−77.3	−72.4	−4.9
hexanoic acid[b]	$CH_3(CH_2)_4CO_2^-$	421.51	0	−76.0	−71.0	−5.0
benzoic acid[b]	$C_6H_5CO_2^-$	441.22	0	−73.0	−74.2	1.2
acrylic acid[b]	$H_2C{=}CHCO_2^-$	441.22	0	−76.0	−74.2	−1.8
pyruvic acid[b]	$CH_3COCO_2^-$	440.91	0	−70.0	−74.2	4.2
acetaldehyde[a]	$CHO.CH_2^-$	431.34	0	−75.7	−72.6	−3.1
dimethyl ketone[a]	$CH_3COCH_2^-$	418.67	0	−75.6	−70.6	−5.0
dimethyl sulfoxide[b]	$CH_3S(O)CH_2^-$	418.67	0	−70.0	−70.6	0.6
methansulfonic acid[a]	$CH_3SO_3^-$	444.65	0	−71.7	−73.5	1.8
nitrous acid[c]	NO_2^-	438.07	0	−69.3	−73.7	4.4
nitric acid[c]	NO_3^-	359.33	0	−65.0	−61.0	−4.0
chloric acid[c]	ClO_3^-	359.73	0	−62.5	−61.1	−1.4
perchloric acid[c]	ClO_4^-	309.12	0	−54.1	−52.9	−1.2
phosphoric acid[d]	$H_2PO_4^-$	392.89	62.4	−68.0	−73.8	5.8
water[b]	H_3O^+	0	805.7	−108.0	−98.7	−9.3
methanol[a]	$CH_3OH_2^+$	0	777.7	−93.1	−95.3	2.2
ethanol[b]	$C_2H_5OH_2^+$	0	759.79	−86.0	−93.2	7.2
acetone[b]	$CH_3C(OH)CH_3^+$	0	595.65	−75.0	−73.8	−1.2
acetophenone[b]	$CH_3C(OH)C_6H_5^+$	0	496.68	−63.0	−62.0	−1.0
ammonia[b]	NH_4^+	0	632.64	−83.0	−78.1	−4.9
methylamine[b]	$CH_3NH_3^+$	0	604.65	−74.0	−74.8	0.8
t-butylamine[b]	$C(CH_3)_3NH_3^+$	0	550.92	−65.0	−68.1	3.07
triethylamine[b]	$(C_2H_5)_3NH^+$	0	494.95	−53.0	−61.8	8.8
cyclohexylamine[b]	$c\text{-}C_6H_{11}NH_3^+$	0	552.64	−67.0	−68.7	1.7
aziridine[b]	$C_2H_4NH_2^+$	0	576.67	−69.0	−71.5	2.5
azetidine[b]	$C_3H_6NH_2^+$	0	558.75	−66.0	−69.4	3.4
pyrrolidine[b]	$C_4H_8NH_3^+$	0	540.84	−64.0	−67.3	3.3
piperidine[b]	$C_5H_{10}NH_3^+$	0	534.4	−62.0	−66.5	4.5
morpholine[b]	$C_4H_8ONH_2^+$	22.06	528.64	−68.0	−69.4	1.4
N,N-dimethylaniline[b]	$C_6H_5NH(CH_3)_2^+$	0	407.76	−55.0	−51.5	−3.5
1-aminonaphthalene[b]	$C_{10}H_7NH_3^+$	0	447.34	−66.0	−56.2	−9.8
formamide[a]	$HCONH_3^+$	8.46	594.7	−82.5	−75.0	−7.5
acetamide[a]	$CH_3CONH_3^+$	10.25	576.82	−73.8	−73.2	−0.6
phosphane[d]	PH_4^+	0	546.00	−73.0	−67.9	−5.1
methylphosphane[d]	$CH_3PH_3^+$	0	518.03	−63.0	−64.5	1.5
dimethylphosphine[d]	$(CH_3)_2PH_2^+$	0	490.04	−57.0	−61.2	4.2
trimethylphosphane[d]	$(CH_3)_3PH^+$	0	462.06	−53.0	−57.9	4.9
dimethyl sulfane[a]	$(CH_3)_2SH^+$	0	552.0	−64.5	−68.6	4.1

[a] Pliego Jr, J. R.; Riveros, J. M. Gibbs energy of solvation of organic ions in aqueous and dimethyl sulfoxide solutions. *Phys. Chem. Chem. Phys.* **2002**, *4*, 1622−1627. [b] Ref 58. [c] Ref 60. [d] Ref 59.

necessarily "better". The HBA and HBD counts are frequently used in QSAR/QSPR studies and, as other molecular descriptors, may be selected from a large descriptor pool. Normally they serve as a substitute when there is a need to account for hydrogen bonding, but more elaborate hydrogen bond descriptors are not available. The HBA and HBD counts are discrete variables not distinguished by the nature of the acceptor and donor atoms. Continual variables such as hydrogen bond acidity *A* and hydrogen bond basicity *B* clearly outperform the simple counts, because the former take into account electronic structure of the donor and acceptor centers as well as their molecular environment. For example, if we retain only *B* in the pK_{HB} correlation (Eq 23), the R^2 drops slightly from 0.982 to 0.950. But replacing *B* by the HBA count results in a very poor correlation ($R^2 = 0.645$). Similar results are produced if one substitutes the HBA and HBD counts into the correlations with the Abraham's basicity and acidity descriptors β_2^H and α_2^H as discussed in sections 4.1 and 4.2: the resultant correlation coefficients are now 0.682 and 0.637 compared to 0.851 and 0.883 if the original *B* and *A* hydrogen bond scales are used. The few examples above of single-parameter correlations are sufficient to illustrate the difference between the purposely designed descriptors *A* and *B* and simple HBA and HBD counts. Thus, any attempt to substitute these counts into the USE multilinear correlations (eq 24−33) would make little or no sense. As we stated above, the Universal Solvation Equation (similarly to the Abraham's Solvation Equation[4]) is a closed

system of a few thoroughly designed descriptors, and therefore, researchers are assured of meaningful results only if the equation is not violated. At the same time the present authors realize that the descriptors' definitions and parametrization may be refined as more liquid media and solvation properties are studied.

6. CONCLUSIONS

We have described new developments in the hydrogen bonding acidity, *A*, and basicity, *B*, theoretical scales that now extend the treatment of aliphatic systems to cover aromatic, heterocyclic, anionic, cationic and zwitter-ionic molecular fragments. Our new *A* and *B* descriptors, together with several other adjuvant descriptors, allow us to derive a new theoretical solvation equation, which we call the Universal Solvation Equation or USE. This equation is applicable to small organic molecules of different charge type and diverse complexity, including complex multifunctional molecules, pharmaceuticals, lead-like and drug-like molecules. We believe that all the important intra- and intermolecular effects taking place on solvation and interphase distribution are reflected in USE. Of the seven descriptors comprising USE, five are derived from simple quantum chemical and phenomenological considerations and can be calculated easily and rapidly: the hydrogen bond acidity and basicity scales, *A* and *B*, polarity *P*, hydrophobicity *H*, and a steric correction *S*. The other two descriptors, molecular polarizability α and the π-electron energy E_π can be calculated with commonly used quantum chemical software.

We report predictive computational models of high accuracy generated for the octanol−water partition coefficient, log *P*, and for the gas−water partition coefficients, log *L*, (aqueous solvation free energies). Aqueous solubility (log S_w) can also be modeled satisfactorily with the USE methodology. A unified solvation free energy model for neutral, zwitter-ionic and charged molecules was developed and tested on an extensive and diverse data set. This is the first QSPR model to consistently treat differently charged species. The accuracy of these predictions made by the USE methodology is comparable, and sometimes superior, to that of a number of solvation models which calculate explicit solute geometry. The topology-based USE model does not require 3D-geometry optimization or molecular dynamics simulations; it is thus extremely fast and well suited for use in virtual screening applications such as high-throughput protein−ligand docking and the accurate predictions of physical and ADMET properties that are paramount in drug development.

Note Added in Proof. While this manuscript was under review, an independent validation of our aqueous solubility model became available through the completion of the Solubility Challenge recently organized by this Journal (announced in *J. Chem. Inf. Model*. **2008**, *48*, 1289–1303, results made available in *J. Chem. Inf. Model*. published ASAP on December 31, 2008). This challenged the community to predict the aqueous solubilities of 32 drugs based on a training set consisting of 100 drug and drug-like molecules. We rebuilt our eq 29 based on this training set and subsequently used the result to predict the 32 unknown solubilities. Resultant correlations of our predictions against the actual values are as follows. For solubility *S* (mg/mL), the squared correlation coefficient, R^2, is equal to 0.631 and 0.575 before and after removal of four largest outliers, respectively. When logarithmic units (log *S*) are correlated, the corresponding R^2 values are 0.608 and 0.811, respectively. These results suggest that the USE equation is able to reproduce general trends of drug solubility. Since the goal of the Solubility Challenge is "... to advance our general understanding of how to better perform solubility estimations", our prediction results allow the general conclusion that explicit quantitative accounting for hydrogen bonding basicity and acidity can assist aqueous solubility predictions.

Supporting Information Available: Table 1S, containing approximate Coulomb α and resonance β integrals for some atoms used for the calculation of π-bond orders, Table 2S with calculated resonance criteria and corrected basicities for various functional groups and atoms, Tables 3S−6S with experimental and predicted log $P_{o/w}$, log $L_{w/g}$, log S_w, and ΔG_{solv} values and numerical data of descriptors calculated for 525 molecules, and two examples of HB basicity calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 3rd updated and revised ed.; Wiley-VCH: Weinheim, Germany, 1988.
(2) Van de Waterbeemd, H.; Gifford, E. ADMET In Silico modeling: Towards prediction paradise. *Nature Rev. Drug Discovery* **2003**, *2*, 192–204.
(3) Kamlet, M. J.; Abraham, M. H.; Doherty, R. M.; Taft, R. W. Solubility properties in polymers and biological media. 4. Correlation of octanol/water partition coefficients with solvatochromic parameters. *J. Am. Chem. Soc.* **1984**, *106*, 464–466.
(4) Abraham, M. H. Physicochemical and biological processes. *Chem. Soc. Rev.* **1993**, *22*, 73–83.
(5) Laurence, C.; Berthelot, M. Observations on the strength of hydrogen bonding. *Perspect. Drug Discovery Des.* **2000**, *18*, 39–60.
(6) Raevsky, O. A. Quantification of non-covalent interactions on the basis of the thermodynamic hydrogen bond parameters. *J. Phys. Org. Chem.* **1997**, *10*, 405–413.
(7) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208–1217.
(8) Weckwerth, J. D.; Vitha, M. F.; Carr, P. W. The development and determination of chemically distinct solute parameters for use in linear solvation energy relationships. *Fluid Phase Equilib.* **2001**, *183−184*, 143–157.
(9) Cronce, D. T.; Famini, G. R.; De Soto, J. A.; Wilson, L. Using theoretical descriptors in quantitative structrure−property relationships: Some distribution equilibria. *J. Chem. Soc., Perkin Trans.* **1998**, *2*, 1293–1301.
(10) Eisfeld, W.; Maurer, G. Study on the correlation and prediction of octanol/water partition coefficients by quantum chemical calculations. *J. Phys. Chem. B* **1999**, *103*, 5716–5729.
(11) Lamarche, O.; Platts, J. A. Theoretical prediction of the hydrogen-bond basicity p*K*(HB). *Chem.−Eur. J.* **2002**, *8*, 457–466.
(12) Politzer, P.; Murray, J. S.; Peralta-Inga, Z. Molecular surface electrostatic potential in relation to noncovalent interactions in biological systems. *J. Comput. Chem.* **2001**, *85*, 676–684.
(13) Haeberlein, M.; Brinck, T. Prediction of water−octanol partition coefficients using theoretical descriptors derived from the molecular surface area and the electrostatic potential. *J. Chem. Soc., Perkin Trans.* **1997**, *2*, 289–294.
(14) Viswanadhan, V. N; Ghose, A. K; Singh, U. C; Wendoloski, J. J. Prediction of solvation free energies of small organic molecules: Additive−constitutive models based on molecular fingerprints and atomic constants. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 405–412.
(15) Tetko, I. V.; Tanchuk, V. Yu.; Kasheva, T. N. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.
(16) Tetko, I. V.; Poda, G. I. Application of ALOGPS 2.1 to predict log *D* distribution coefficient for Pfizer proprietary compounds. *J. Med. Chem.* **2004**, *47*, 5601–5604.

(17) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State of the art and comparison of log *P* methods on more than 96,000 compounds. *J. Pharm. Sci.* In press.
(18) McElroy, N. R.; Jurs, P. C. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237–1247.
(19) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E, Jr. A general treatment of solubility. 1. The QSPR correlation of solvatation free energies of single solutes in series of solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794–1805.
(20) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E., Jr. A general treatment of solubility. 2. QSPR prediction of free energies of solvation of specified solutes in ranges of solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806–1814.
(21) *Codessa-Pro, Comprehensive Descriptors for Structural and Statistical Analysis*, version 1.0 RC2; University of Florida: Gainesville, FL, 2001. STN on the Web: www.codessa-pro.com (accessed Oct 21, 2008).
(22) DRAGON, Milano Chemometrics and QSAR Research Group. STN on the Web www.talete.mi.it/dragon_net.htm.
(23) Duffy, E. M.; Jorgensen, W. L. Prediction of properties from simulations: free energies of solvation in hexadecane, octanol, and water. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.
(24) Klamt, A.; Eckert, F.; Hornig, M. COSMO-RS: A novel view to physiological solvation and partition questions. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 355–365.
(25) Oliferenko, A. A.; Oliferenko, P. V.; Hudleston, J. G.; Rogers, R. D.; Palyulin, V. A.; Zefirov, N. S.; Katritzky, A. R. Theoretical scales of hydrogen bond acidity and basicity for application in QSAR/QSPR studies and drug design. Partitioning of aliphatic compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1042–1055.
(26) Parr, R. G.; Szentpaly, L. v.; Liu, S. Electrophilicity index. *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924.
(27) Hinze, J.; Whitehead, M. A.; Jaffé, H. H. Electronegativity. II. Bond and orbital electronegativities. *J. Am. Chem. Soc.* **1963**, *85*, 148–154.
(28) Bergmann, D.; Hinze, J. Electronegativity and molecular properties. *Angew. Chem. Angew. Chem., Int. Ed.* **1996**, *35*, 150–163; **1996**, *108*, 162−−176.
(29) Purcell, W. P.; Singer, J. A. Electronic and molecular structure of selected unsubstituted and dimethyl amides from measurements of electric moments and nuclear magnetic resonance. *J. Phys. Chem.* **1967**, *71*, 4316–4319.
(30) www.chem.ucalgary.ca/SHMO/.
(31) Abraham, M. H.; Grellier, P. L.; Prior, D. V.; Morris, J. J.; Taylor, P. J. Hydrogen bonding. Part 10. A scale of solute hydrogen-bond basicity using log *K* values for complexation in tetrachloromethane. *J. Chem. Soc., Perkin Trans.* **1990**, *2.*, 521–529.
(32) Leahy, D. E.; Morris, J. J.; Taylor, P. J.; Wait, A. R. Model solvent systems for QSAR. Part 3. An LSER analysis of the critical quartet. New light on hydrogen bond strength and directionality. *J. Chem. Soc., Perkin Trans. 2.* **1992**, *70*, 5–722.
(33) Hinze, J.; Jaffé, H. H. Electronegativity. IV. Orbital electronegativities of neutral atoms of the periods 3A and 4A and of positive ions of periods 1 and 2. *J. Am. Chem. Soc.* **1963**, *67*, 1501–1506.
(34) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley: New York, 2000.
(35) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally diverse quantitative structure−property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1–18.
(36) Abraham, M. H.; McGowan, J. C. *Chromatographia* **1987**, *23*, 243–247.
(37) Verechshagin, A N. *Molecular Polarizability*; Nauka: Moscow, 1987.
(38) Jain, A.; Yang, G.; Yalkowsky, S. H. Estimation of total entropy of melting of organic compounds. *Ind. Eng. Chem. Res.* **2004**, *43*, 4376–4379.
(39) HyperChem 7.0, HyperCube, 2002. http://www.hypercube.com
(40) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(41) Stewart, J. J. P. *MOPAC Program Package*; QCPE: 1989, No 455.
(42) Kurtz, H. A.; Stewart, J. J. P.; Dieter, K. M. Calculation of the nonlinear optical properties of molecules. *J. Comput. Chem.* **1990**, *11*, 82–87.
(43) *STATISTICA*, version 6; StatSoft, Inc.: 2001. STN on the Web www.statsoft.com.
(44) Abraham, M. H.; Chadha, H. S.; Whiting, G. S.; Mitchell, R. C. Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the Δlog *P* parameter of seiler. *J. Pharm. Sci.* **1994**, *83*, 1085–1100.
(45) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. Origin of the attraction and directionality of the NH/π interaction: comparison with OH/π and CH/π interactions. *J. Am. Chem. Soc.* **2000**, *122*, 11450–11458.
(46) Novoa, J. J.; Mota, F. The C−H ••• π bonds: Strength, identification, and hydrogen-bonded nature: A theoretical study. *Chem. Phys. Lett.* **2000**, *318*, 345–354.
(47) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *J. Med. Chem.* **2008**, *51*, 769–779.
(48) Chamberlin, A. C.; Cramer, C. J.; Truhlar, D. G. Performance of SM8 on a test set to predict small-molecule solvation free energies. *J. Phys. Chem. B* **2008**, *112*, 8651–8655.
(49) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A fuzzy ARTMAP based on quantitative structure-property relationships (QSPRs) fo predicting aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177–1207.
(50) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773–777.
(51) Bruneau, P. Search for predictive generic model of aqueous solubility using bayesian neural nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605–1616.
(52) Klopman, G.; Zhu, H. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439–445.
(53) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **1999**, *88*, 868–880.
(54) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T. QSPR studies on vapor pressure, aqueous solubility, and the prediction of water−air partition coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720–725.
(55) Chen, X.; Cho, S, J.; Li, Y.; Venkatesh, S. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J. Pharm. Sci.* **2002**, *91*, 1838–1852.
(56) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155–1158.
(57) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. The linear free energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33*, 347–350.
(58) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. SM6: A density functional theory continuum solvation model for calculating aqueous solvation free energies of neutrals, ions, and solute−water clusters. *J. Chem. Theory Comput.* **2005**, *1*, 1133–1152.
(59) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. Extention of the platform of applicability of the SM5.42R universal solvation model. *Theor. Chem. Acc.* **1999**, *103*, 9–63.
(60) Camaioni, D. M.; Dupuis, M.; Bentley, J. Theoretical characterization of oxoanion, XO_m^{n-}, solvation. *J. Phys. Chem. A* **2003**, *107*, 5778–5788.
(61) Katritzky, A. R.; Dobchev, D. A.; Slavov, S.; Karelson, M. Legitimate utilization of large descriptor pools for QSPR/QSAR models. *J. Chem. Inf. Model.* **2008**, *48*, 2207–2213.

CI800323Q