Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water

Erin M. Duffy*,† and William L. Jorgensen*,‡

Contribution from the Central Research Division, Pfizer Inc., Groton, Connecticut 06340, and Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107

Received October 12, 1999. Revised Manuscript Received January 12, 2000

Abstract: Monte Carlo (MC) statistical mechanics simulations have been carried out for more than 200 organic solutes, including 125 drugs and related heterocycles, in aqueous solution. The calculations were highly automated and used the OPLS-AA force field augmented with CM1P partial charges. Configurationally averaged results were obtained for a variety of physically significant quantities including the solute−water Coulomb and Lennard-Jones interaction energies, solvent-accessible surface area (SASA), and numbers of donor and acceptor hydrogen bonds. Correlations were then obtained between these descriptors and gas to liquid free energies of solvation in hexadecane, octanol, and water and octanol/water partition coefficients. Linear regressions with three or four descriptors yielded fits with correlation coefficients, r^2, of 0.9 in all cases. The regression equation for log *P*(octanol/water) only needs four descriptors to provide an rms error of 0.55 for 200 diverse compounds, which is competitive with the best fragment methods. For water, the expanded data set of 85 solutes and improved statistical analyses bring into question the significance of the Lennard-Jones and surface area terms that have been featured in prior linear-response treatments. The results are sensitive to the choice of partial charges for the solute atoms; poor representation of some functional groups can lead to the need for specific corrections in the regression equations. This is expected to also be true for force-field-based scoring functions for protein−ligand binding. In all cases, the present descriptors that emerge as most significant sensibly reveal the key physical factors that control solvation, especially solute size in organic solvents and electrostatic interactions in water. Furthermore, additional MC simulations for solutes in both water and ethanol clearly demonstrate that the key differential between water and alcohols is the greater hydrogen-bond-donating ability of water, which explains the significance of a solute's hydrogen-bond-accepting ability for log *P*(octanol/water).

Introduction

Quantum mechanics has been remarkably successful in predicting properties of molecules in the gas phase. Extension to the liquid state remains a significant challenge that is needed for understanding solvent effects on important processes, including reactions, conformational equilibria, electronic transitions, and receptor−ligand binding. One needs to address the corresponding changes in free energy of solvation between initial and transition states, alternate conformers, ground and excited states, and complexes versus separated host and guest. The principal approaches for computing free energies of solvation feature either a continuum description of the solvent or the use of a large number of discrete solvent molecules.[1−3] The continuum methods have evolved from classical electrostatics to provide complete free energies of solvation through addition of terms for solute cavity formation and solute−solvent van der Waals interactions, as in the generalized Born model,[4] and through better estimates of the electrostatic contribution via Poisson−Boltzmann calculations[5] or the quantum mechanical treatment of the solute in the reaction field of the polarized solvent.[3,6,7] These procedures generally require a charge distribution for the solute and various parameters for describing the solute's cavity, the nonelectrostatic terms, and dielectric constants for the solvent and solute domains. The discrete models feature Monte Carlo (MC) statistical mechanics or molecular dynamics (MD) simulations for the solutes and solvent molecules coupled with a procedure for computing free energy changes, most commonly, free energy perturbation (FEP) theory or thermodynamic integration (TI).[8] Standard force fields are used, and from that point there are no more adjustable parameters. Importantly, the calculations are rooted properly in classical statistical mechanics.

The most attractive feature of the non quantum mechanical continuum approaches is that they are computationally fast. Addition of the quantum mechanical calculation allows for

† Pfizer Inc.

‡ Yale University.

(1) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027−2094.

(2) Jorgensen, W. L.; Tirado-Rives, J. *Perspect. Drug Discovery Des.* **1995**, *3*, 123−138.

(3) (a) Hawkins, G. D.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Org. Chem.* **1998**, *63*, 4305−4313. (b) Li, J.; Zhu, T.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chim. Acta* **1999**, *103*, 9−63.

(4) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. F. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129. Best, S. A.; Merz, K. M., Jr.; Reynolds, C. H. *J. Phys. Chem. B* **1997**, *101*, 10479−10487.

(5) Jean-Charles, A.; Nicholls, A.; Sharp, K.; Honig, B.; Tempczyk, A.; Hendrickson, T. F.; Still, W. C. *J. Am. Chem. Soc.* **1991**, *113*, 1454−1455.

(6) Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 629−666.

(7) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775−11788.

(8) (a) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417. (b) Jorgensen, W. L. Free Energy Changes in Solution. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 2, p 1061−1070.

10.1021/ja993663t CCC: $19.00
Published on Web 03/10/2000

polarization of the solute by the solvent and a potentially more accurate description of the charge distribution and molecular shape. The principal drawbacks of most continuum models are as follows: (1) new parametrization is needed for a change in solvent or temperature and pressure, (2) a single solute structure is used with no configurational averaging, and (3) the lack of discrete solvent molecules allows limited insights into changes in solvation and is expected to be problematic for specific interactions such as hydrogen bonding. The discrete models are more general in that, given a general force field, results can be obtained without reparametrization for any solvent under wide ranges of temperatures and pressures. The configurational averaging is also advantageous, particularly for conformationally flexible systems, and details of the individual solute−solvent interactions are readily available. The drawbacks are as follows: (1) the computational demands are much higher than those with most continuum treatments, (2) the standard fixed-charge potential functions can only reflect polarization of solute and solvent in an average sense, and (3) the FEP and TI calculations most readily yield relative rather than absolute free energies of solvation, and mutations between very different structures can be impractical.

An interesting compromise has been explored stemming from work by Åqvist et al., who introduced a procedure based on linear response (LR) theory for estimating free energies of binding.[9] In this model, the free energy of interaction of a solute with its environment is a linear function of the electrostatic (Coulombic) energy plus the van der Waals (Lennard−Jones) energy. For a ligand binding to a protein, the differences in the interactions between the ligand in the unbound state and that bound in the complex then provide an estimate of the free energy of binding, ΔG_b, via eq 1. On the basis of classical electrostatics,

$$\Delta G_b = \beta\langle\Delta E_{elec}\rangle + \alpha\langle\Delta E_{vdW}\rangle \quad (1)$$

β was set to 0.5, while α was determined empirically by fitting to experimental data for a series of inhibitors.[9a] The required energy components were obtained from MD simulations for the inhibitors and the protein−inhibitor complexes in water. Key advantages over FEP methods are that absolute free energies of binding are estimated and that only simulations at the end points of a mutation are required, which allows much easier application to structurally diverse sets of molecules. Despite the approximations in eq 1, the approach has yielded promising results for several applications.[9,10]

We sought validation on a simpler problem, estimation of free energies of hydration, ΔG_{hyd},[11] and we subsequently considered free energies of solvation in chloroform and chloroform/water partition coefficients, log *P*.[12] This extension required a third term for cavity formation, which could be a constant or, to be more general, could be made proportional to the solute's solvent-accessible surface area (SASA), eq 2, and

$$\Delta G_{sol} = \beta\langle E_{elec}\rangle + \alpha\langle E_{vdW}\rangle + \gamma\langle SASA\rangle \quad (2)$$

β was allowed to vary from 0.5. The solvation studies were performed for 35 organic molecules with diverse functional groups in water and for 16 of the molecules in chloroform. The three averages came from MC simulations with the OPLS-AA force field and 6-31G* CHELPG charges in TIP4P water. In comparison to experimental data, the rms deviations for the predicted free energies of solvation were 1.0 and 0.5 kcal/mol in water and chloroform and the rms error was 0.35 for the log *P* values.[12] The results are easily competitive with FEP calculations, and the LR approach also retains the advantages of explicit-solvent simulations including conformational sampling and facile changes for *T*, *P*, and the solvent model.

As reported here, to further test the methodology, we have sought to treat much larger data sets that would include results for polyfunctional molecules. Extension of the hydration studies to 85 molecules revealed problems with eq 2, in particular, high correlation and questionable statistical significance for E_{vdW} and SASA. Consequently, other descriptors, q_i, were sought from the MC simulations in water that had physical significance and could be used in a general regression equation (eq 3) for

$$\log P = \sum_i a_i\langle q_i\rangle + \text{const} \quad (3)$$

predicting free energies of solvation or partition coefficients ($\Delta G_{sol} = -2.303\ RT \log P$). In the present study, results are provided for log *P*(hexadecane/gas), log *P*(octanol/gas), log *P*(water/gas), and log *P*(octanol/water). Correlations with r^2 values of 0.9 are obtained in each case with just a few descriptors, including for log *P*(octanol/water), for which 230 diverse organic molecules and drugs have been treated. The optimal choices of descriptors also illuminate the factors that control solvation in each case.

Computational Methods

Force Field. The potential energy function consists of harmonic bond-stretching and angle-bending terms, a Fourier series for torsional energetics, and Coulomb and Lennard-Jones terms for the nonbonded interactions, eqs 4−7.[13] The parameters are the force constants *k*, the

$$E_{bond} = \sum_i k_{b,i}(r_i - r_{0,i})^2 \quad (4)$$

$$E_{angle} = \sum_i k_{\vartheta,i}(\vartheta_i - \vartheta_{0,i})^2 \quad (5)$$

$$E_{torsion} = \sum_i \left[\frac{1}{2}V_{1,i}(1 + \cos\varphi_i) + \frac{1}{2}V_{2,i}(1 - \cos 2\varphi_i) + \frac{1}{2}V_{3,i}(1 + \cos 3\varphi_i)\right] \quad (6)$$

$$E_{nonbond} = \sum_i \sum_{j>i} \left\{\frac{q_i q_j e^2}{r_{ij}} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]\right\} \quad (7)$$

r_0 and ϑ_0 reference values, the Fourier coefficients *V*, the partial atomic charges *q*, and the Lennard-Jones radii and well depths, σ and ϵ. Standard combining rules are used such that $\sigma_{ij} = (\sigma_{ii}\sigma_{jj})^{1/2}$ and $\epsilon_{ij} = (\epsilon_{ii}\epsilon_{jj})^{1/2}$.[8] The nonbonded interactions are evaluated for intermolecular interactions and for intramolecular atom pairs separated by three or more bonds. The 1,4-intramolecular interactions are reduced by a factor of 2 in order to use the same parameters for both intra- and intermolecular interactions.[13]

The bond-stretching, angle-bending, and torsional parameters are from the OPLS-AA force field.[13,14] Since OPLS-AA charges are not

(9) (a) Åqvist, J.; Medina, C.; Sammuelsson, J.-E. *Protein Eng.* **1994**, *7*, 385−391. (b) Åqvist, J.; Hansson, T. *J. Phys. Chem.* **1996**, *100*, 9512−9521.

(10) For example, see: Paulsen, M. D.; Ornstein, R. L. *Protein Eng.* **1996**, *9*, 567−571. Hansson, T.; Marelius, J.; Åqvist, J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27−35. Lamb, M. L.; Tirado-Rives, J.; Jorgensen, W. L. *Bioorg. Med. Chem.* **1999**, *7*, 851−860.

(11) Carlson, H. A.; Jorgensen, W. L. *J. Phys. Chem.* **1995**, *99*, 10667−10673.

(12) McDonald, N. A.; Carlson, H. A.; Jorgensen, W. L. *J. Phys. Org. Chem.* **1997**, *10*, 563−576.

(13) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(14) Jorgensen, W. L. *BOSS, Version 4.1*; Yale University: New Haven, CT, 1999.

available for some of the functional groups in the drugs, all partial charges were obtained from PM3 calculations using the CM1P procedure.[15] Though we have previously used AM1-based CM1A charges,[16] we now prefer CM1P owing to better representation of the partial charges for nitrogen-containing functional groups, particularly amides. Both the CM1P and CM1A procedures provide charges that yield excellent gas-phase dipole moments. However, charges appropriate for solution-phase simulations need to be enhanced.[13] From comparisons of OPLS-AA and CM1P charges for many molecules, a scaling factor of 1.3 was determined as optimal for neutral molecules and has been applied here. The results using the scaled CM1P charges for the full set of 230 molecules are focused on in this work. However, results have also been obtained for 145 molecules with OPLS-AA charges and they receive some mention for comparison.

All calculations have been performed with BOSS 4.1 in an automated manner. The only input that is required is a coordinate file for the solute, e.g., PDB or mol2. The initial solute structures were built in low-energy conformations and submitted to an initial geometry optimization with the ChemEdit program.[17] The resulting coordinate file was then processed by BOSS in the following sequence: perform the PM3 single-point calculation → compute scaled CM1P charges → assign atom types → assign OPLS-AA parameters → energy-minimize the structure → recompute charges → place the structure in a water box → perform the MC simulation. Each atom has an associated atom type represented by a two-letter code that is used to look up the parameters for the individual atoms (Lennard-Jones), atom pairs (bond stretching), triplets (angle bending), and quartets (torsions). The two-letter codes used with OPLS-AA are expanded from the AMBER set[18] to include, for example, CO (acetal C), C= (C2 in dienes), CZ (sp C), C! (biphenyl C1), CY (sp^3 C in small ring), C$ (carbonyl C in small ring), NZ (sp N), N$ (amide N in small ring), NO (N in N═O), ON (O in N═O), OY (O in S═O), O$ (O in small ring), S= (S in C═S), and SY (S in S═O). If an OPLS-AA parameter is missing, it is estimated in BOSS from the nearest matches and/or generic entries with ?? wildcards, e.g., ??-CA-CA-?? for any torsion about an aromatic C−C bond.

Monte Carlo Simulations. The MC calculations were performed for a single solute in a periodic cube with 500 TIP4P water[19] molecules at 25 °C and 1 atm in the NPT ensemble.[20] The solute was initially placed in an equilibrated box containing 512 water molecules, and the 12 water molecules with the highest energy interactions with the solute were discarded. The water−water cutoff was at 9 Å on the basis of the O−O distance, and the solute−water interactions were included if any non-hydrogen atom of the solute was within 9 Å of the water O. The interactions were quadratically smoothed to zero within 0.5 Å of the cutoff. Each simulation consisted of 0.2 million configurations of equilibration without volume changes, followed by 3 million configurations of full equilibration and 10 million configurations of averaging. The TIP4P water molecules underwent only rigid-body translations and rotations, while the sampling of the solutes included all internal degrees of freedom, as well as the total translations and rotations. Solute and volume moves were attempted every 120 and 3125 configurations, respectively. Acceptance rates of 30−50% for new configurations were normally maintained by automatic adjustment of the ranges for the motions and, for large solutes, by only varying random subsets of internal coordinates on each attempted move. The simulations could be performed for about 25 solutes per day on a dual-processor 500 MHz Pentium III computer. Collections of hundreds of compounds have been readily processed on larger multiprocessor systems.

Table 1. Descriptors Averaged during the Monte Carlo Simulations in Water

symbol	description
ESXC	solute−solvent Coulomb energy
ESXL	solute−solvent Lennard-Jones energy
SASA	total solvent-accessible surface area
FOSA	hydrophobic SASA
FISA	hydrophilic SASA
ARSA	aromatic SASA
DIPL	solute dipole moment
INST	no. of solute−solvent interactions < -3.75 kcal mol^{-1}
INME	no. of solute−solvent interactions < -2.75 kcal mol^{-1}
HBDN	no. of solute as donor hydrogen bonds
HBAC	no. of solute as acceptor hydrogen bonds

The 11 descriptors that were considered are listed in Table 1. Each quantity was averaged over the entire MC run and was output by the BOSS program. ESXC and ESXL are the electrostatic and van der Waals solute−solvent interaction energies, which are featured in eqs 1 and 2. The SASA was determined with a probe 1.4 Å in radius. The atomic radii were computed from the OPLS-AA Lennard-Jones σ's via $r = 2^{1/6}\sigma/2$. The SASA was decomposed into hydrophobic, hydrophilic, and aromatic contributions on the basis of the atom types. Heteroatoms and attached hydrogens are hydrophilic, carbons and attached hydrogens in aromatic rings are aromatic, and the remainder are hydrophobic. This decomposition could be refined in the future, e.g., to separate more hydrophilic atoms such as N and O from less hydrophilic ones. Counts were also averaged for the number of strong and medium interactions between the entire solute and water molecules. Generally, hydrogen bonds have interaction energies below about −3 kcal mol^{-1}, so medium and strong interactions were defined with energy cutoffs of −2.75 and −3.75 kcal mol^{-1}. The average numbers of solute as donor and acceptor hydrogen bonds were also averaged using a geometric cutoff of 2.5 Å for solute H/water O and solute N, O, or S/water H distances. The distance cutoff follows from numerous studies of organic solutes in water; hydrogen bonds are reflected in sharp first peaks with minima near 2.5 Å in X−H radial distribution functions.[21] Future refinement here could also use an energetic criterion to separate weaker hydrogen bonds from stronger ones. A technical advantage for the use of the medium and strong interactions or hydrogen-bond counts is that they involve interactions near the surface of the solute; they are not prone to influences of the number of water molecules used in the simulations that may affect ESXC, in particular, for large solutes. Tests for several of the larger solutes in a periodic box with 725 water molecules instead of 500 revealed insignificant effects for all of the descriptors in Table 1 except ESXC, which showed variations of 10−15%. Some other descriptors including the solute's molecular weight and internal energy in both water and the gas phase and various ratios such as ARSA/SASA (fraction of aromatic surface area) were considered but did not prove generally useful.

Statistical Analyses. The resulting database for the solutes was maintained and analyzed with the JMP program.[22] Linear regression analyses were performed, and the optimal descriptor sets were chosen to maximize the correlation coefficient r^2 and to minimize the rms error with as few descriptors as possible. The statistical significance of the descriptors was confirmed from the analysis of variance using the F ratios (regression model mean square/error mean square) and requiring that the probability of a greater F value occurring by chance (Prob > F) is less than 0.001. A cross-validated r^2 value, q^2, was obtained for log P(octanol/water) by a leave-one batch-out procedure using 20 batches of 10 randomly chosen compounds. Division of these original 200 compounds, which were used for the reported correlation, into training and test sets was not performed, since this is only statistically meaningful for significantly larger data sets. However, the original 200 compounds were augmented by another 30 molecules while this paper was under review. These compounds were selected for their availablility

(15) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87−110.

(16) Kaminski, G. A.; Jorgensen, W. L. *J. Phys. Chem. B* **1998**, *102*, 1787−1796.

(17) Lim, D.; Jorgensen, W. L. ChemEdit. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 5, pp 3295−3302.

(18) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(19) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(20) Jorgensen, W. L. Monte Carlo Simulations of Liquids. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 3, pp 1754−1763.

(21) See, for example: Jorgensen, W. L.; Nguyen, T. B. *J. Comput. Chem.* **1993**, *14*, 195−205.

(22) *JMP*, Version 3; SAS Institute Inc.: Cary, NC, 1995.

Table 2. Selected Results for Numbers of Donated and Accepted Hydrogen Bonds

molecule	HBDN	HBAC	molecule	HBDN	HBAC
methanol	1.0	2.0	acetamide	2.0	2.8
2-propanol	0.9	1.7	*N*-methylacetamide	0.8	2.7
2-methyl-2-propanol	1.0	1.1	propionitrile	0.0	1.9
2,2,2-trifluoroethanol	1.0	1.0	nitromethane	0.0	3.1
1,2-ethanediol	1.3	2.8	ethylamine	1.3	1.0
phenol	1.0	1.9	pyrrole	1.0	0.4
diethyl ether	0.0	1.9	1,4-dimethylpiperazine	0.0	1.8
furan	0.0	0.7	aniline	1.7	0.6
18-crown-6	0.0	8.5	methanethiol	0.8	0.4
acetaldehyde	0.0	1.6	dimethyl sulfoxide	0.0	3.1
acetone	0.0	1.9	dimethyl sulfone	0.0	4.1
acetophenone	0.0	2.3	uracil	2.0	4.7
methyl benzoate	0.0	2.0	cytosine	3.0	4.6
benzoic acid	1.0	2.7	adenine	3.1	5.2
salicylic acid	1.1	2.5	guanine	4.2	6.1
sucrose	8.1	12.7	cimetidine	3.0	6.5
diazepam	0.0	3.0	estradiol	1.9	3.5
lovastatin	1.0	6.1	nifedipine	1.1	7.8
acetaminophen	2.0	3.1	zidovudine	2.0	7.3
penicillin G	2.0	7.6	verapamil	0.0	7.6

of experimental data on their aqueous solubilities, which were needed in a related project. The predictions for these additional molecules are also reported.

A cross-correlation was performed for the 11 descriptors and revealed the following pairs have correlation coefficients greater than 0.75: ESXC with INST (0.99), INME (0.99), HBDN (0.78), and HBAC (0.92); ESXL with SASA (0.93); INST with INME (0.99), HBDN (0.80), and HBAC (0.93); and INME with HBDN (0.78) and HBAC (0.92). ESXC, INST, and INME measure electrostatic interactions. INST, INME, HBDN, and HBAC are correlated because INST $\approx$ HBDN + HBAC and INST contains all of INME. Finally, the short-range nature of ESXL is reflected in its correlation with SASA.

Results

Donor and Acceptor Hydrogen Bonds. The results for the 11 descriptors are provided in the Supporting Information for the 230 compounds. Some representative results for the numbers of donor and acceptor hydrogen bonds are listed in Table 2. The results generally correspond well to chemical intuition. Thus, methanol donates one hydrogen bond and accepts two; however, branching or addition of an electron-withdrawing group reduces the accepting ability of alcohols. Simple ethers accept nearly two hydrogen bonds with the CM1P charges, while ca. 1.4 hydrogen bonds is the norm with OPLS-AA charges. Aldehydes, ketones, and esters have two hydrogen-bonded water molecules on the carbonyl oxygen, while carboxylic acids and amides are somewhat better acceptors. Saturated amines accept the expected one hydrogen bond. Though nitriles might be anticipated to accept only one hydrogen bond, two is the rule with both CM1P and OPLS-AA charges. Sulfoxides, sulfones, and nitro compounds are particularly good acceptors with three or four hydrogen bonds using the CM1P charges, which is one more than with OPLS-AA charges. Naturally, human predictions for polyfunctional systems are less clear. Reasonable additivity can be expected when the functional groups are well separated and not sterically shielded. For example, the donor and acceptor results of one and six are reasonable for lovastatin with its two ester groups and a secondary alcohol. However, the competition between internal and external hydrogen bonding for sucrose makes a confident prediction impossible. Most results in Table 2 can be rationalized after the fact, but the value of a reliable, automated procedure is evident.

Hexadecane. Experimental free energies of solvation in hexadecane are available for 68 of the compounds.[23] The experimental data cover a range of 8 log units with a maximum value of 7.6 for anthracene. These refer to transfer from the gas phase into hexadecane. The standard state for all free energies of transfer in this paper is 1 M in both phases. A change in standard state would yield a change in the constant in eq 3. The regression analyses yielded a good fit using just three descriptors with $r^2 = 0.90$ and an rms of 0.43 (eq 8 and Table 3). Transfer to hexadecane becomes more favorable for solutes as their size increases with added boosts for aromatic fragments and increased polarity. The aromatic and polarity terms can be attributed to polarization of the solute and solvent, respectively. Not surprisingly, the dipole measure of polarity can be replaced with little deficit by the Coulomb energy ($r^2 = 0.90$, rms = 0.45) or the number of medium interactions, INME ($r^2 = 0.89$, rms = 0.46). The damage is greater for replacing SASA by the Lennard-Jones energy, ESXL ($r^2 = 0.84$, rms = 0.55). Addition of any of the remaining descriptors to eq 8 makes no statistically significant improvement to the fit. The dipole term is the least significant; however, leaving it out yields a fit with $r^2 = 0.86$ and rms = 0.52. The only compounds for which the error with eq 8 is greater than 1 log unit are anthracene (1.15) and dimethyl disulfide (1.15). Very similar results are obtained for a subset of 63 compounds with the OPLS-AA charges. The terms in eq 8 remain the most significant.

Table 3. Experimental[a] and Predicted (Eq 8) log *P*(hexadecane/gas) Values

molecule	calcd	exptl	molecule	calcd	exptl
methane	0.07	−0.32	water	−0.12	0.26
ethane	0.78	0.49	phenol	3.53	3.77
propane	1.37	1.05	*p*-cresol	4.00	4.31
butane	1.96	1.61	dimethyl ether	1.10	1.09
pentane	2.54	2.16	diethyl ether	2.63	2.06
hexane	2.99	2.67	tetrahydrofuran	1.75	2.64
cyclohexane	2.36	2.96	1,2-dimethoxyethane	2.62	2.66
propene	1.56	0.95	anisole	3.89	3.92
1,3-butadiene	2.15	1.54	acetaldehyde	1.22	1.23
1-pentyne	2.25	2.01	propanal	1.86	1.82
ethyl fluoride	1.24	0.56	benzaldehyde	4.32	3.99
ethyl chloride	1.70	1.54	acetone	1.61	1.69
1,1,1-trichloroethane	2.34	2.69	butanone	2.31	2.29
1-chloropropane	2.22	2.20	acetophenone	4.55	4.50
1,2-dichloroethane	2.42	2.57	acetic acid	1.14	1.75
cis−1,2-dichloroethene	2.22	2.45	methyl acetate	2.08	1.96
trans−1,2-dichloroethene	1.81	2.35	methyl butyrate	3.29	2.94
trichloroethene	2.09	3.00	methyl benzoate	5.07	4.63
benzene	3.27	2.79	ethyl acetate	2.78	2.38
toluene	3.59	3.33	ethylamine	1.44	1.68
naphthalene	4.85	5.34	dimethylamine	1.42	1.60
anthracene	6.42	7.57	trimethylamine	1.90	1.62
biphenyl	6.01	6.13	aniline	3.72	3.99
fluorobenzene	3.59	2.84	pyridine	3.31	3.01
chlorobenzene	3.86	3.64	acetonitrile	1.67	1.74
bromobenzene	3.81	4.04	propionitrile	2.33	2.08
methanol	0.76	0.97	benzonitrile	4.69	4.04
ethanol	1.43	1.49	acetamide	1.88	2.44
1-propanol	1.94	2.10	nitromethane	1.79	1.89
2-propanol	1.93	1.82	nitroethane	2.39	2.37
allyl alcohol	1.87	2.00	nitrobenzene	4.83	4.56
2,2,2-trifluoroethanol	1.79	1.22	dimethyl sulfide	1.73	2.24
2-methyl-2-propanol	2.43	2.01	dimethyl disulfide	2.40	3.55
1,2-ethanediol	1.93	2.06	thiophene	2.73	2.94

[a] References 3b and 23.

$$\log P(\text{hexadecane/gas}) = 0.01767(\text{SASA}) + 0.005163(\text{ARSA}) + 0.1747(\text{DIPL}) - 2.801 \quad (8)$$

Octanol. The data refer to water-saturated octanol and are obtained by combining the experimental data for free energies

(23) Abraham, M. H.; Whiting, G. S.; Fuchs, R.; Chambers, E. J. *J. Chem. Soc., Perkin Trans. 2* **1990**, 291−300.

Table 4. Experimental[a] and Predicted (Eqs 9 and 10) log *P*(octanol/gas) and log *P*(water/gas) Values

	log *P*(octanol/gas)		log *P*(water/gas)			log *P*(octanol/gas)		log *P*(water/gas)	
molecule	calcd	exptl	calcd	exptl	molecule	calcd	exptl	calcd	exptl
methane	−0.21	−0.36	−0.55	−1.45	tetrahydrofuran	2.26	3.03	1.77	2.57
ethane	0.55	0.47	−0.68	−1.34	dimethoxymethane	2.34	2.36	1.99	2.18
propane	1.18	0.92	−0.78	−1.44	1,2-dimethoxyethane	3.44	3.34	2.42	3.55
butane	1.80	1.37	−0.89	−1.52	anisole	4.07	2.88	1.73	0.77
pentane	2.43	1.69	−0.98	−1.70	acetaldehyde	1.98	2.66	3.06	2.60
hexane	2.91	2.08	−1.06	−1.82	propanal	2.50	3.14	2.61	2.55
cyclohexane	2.24	1.96	−0.97	−0.90	benzaldehyde	5.06	4.47	3.77	2.99
propene	1.27	0.83	−0.37	−0.94	acetone	2.22	2.58	2.69	2.82
1,3-butadiene	1.92	1.57	−0.01	−0.42	butanone	3.11	3.01	2.98	2.72
1-pentyne	2.58	1.97	0.53	−0.01	acetophenone	5.11	4.94	3.59	3.36
cyclohexene	2.46	2.59	−0.31	−0.27	acetic acid	3.96	4.80	5.38	4.97
ethyl chloride	1.77	1.90	0.51	0.47	methyl acetate	3.24	2.61	3.33	2.43
1,1,1-trichloroethane	2.72	2.63	0.22	0.14	methyl butyrate	4.46	3.39	2.96	2.10
1-chloropropane	2.25	2.24	0.25	0.20	methyl benzoate	6.05	5.26	3.68	3.14
1,2-dichloroethane	2.58	2.76	0.74	1.28	ethyl acetate	3.93	2.89	3.17	2.16
cis-1,2-dichloroethene	2.38	2.37	0.65	0.51	methylamine	3.07	2.78	2.90	3.35
trans-1,2-dichloroethene	2.29	2.56	0.11	0.57	ethylamine	3.23	3.17	2.41	3.30
trichloroethene	2.58	2.74	0.05	0.32	dimethylamine	2.70	2.77	3.76	3.15
benzene	3.34	2.79	1.50	0.66	piperidine	3.46	4.58	3.44	3.74
toluene	3.67	3.29	1.06	0.56	trimethylamine	1.89	2.53	2.23	2.37
naphthalene	5.15	5.15	2.16	1.80	1-methylpyrollidine	2.70	3.87	2.39	2.95
anthracene	6.85	7.63	2.49	3.18	1,4-dimethylpiperazine	3.94	5.16	5.50	5.56
phenanthrene	6.77	7.48	2.63	3.02	morpholine	3.68	4.40	5.78	5.26
fluorene	6.59	6.64	2.98	2.46	aniline	6.31	4.49	4.37	3.59
pyrene	7.38	8.43	3.23	3.35	acetonitrile	2.32	2.55	3.23	2.89
biphenyl	6.38	6.02	2.32	1.93	propionitrile	2.96	3.02	3.01	2.86
fluorobenzene	3.31	2.84	0.97	0.57	benzonitrile	4.79	4.57	3.05	3.01
chlorobenzene	3.71	3.64	0.91	0.75	acetamide	6.09	5.86	8.31	7.12
bromobenzene	3.80	4.06	1.03	1.07	*N*-methylacetamide	5.52	6.34	7.46	7.39
methanol	2.52	2.98	3.22	3.75	*N*,*N*-dimethylacetamide	4.24	5.50	4.60	6.27
ethanol	3.18	3.36	3.05	3.67	nitromethane	3.33	2.60	2.63	2.95
1-propanol	3.81	3.81	3.02	3.56	nitroethane	3.86	2.90	2.31	2.72
2-propanol	3.70	3.53	2.97	3.48	nitrobenzene	6.17	4.91	3.79	3.06
allyl alcohol	3.76	3.91	3.34	3.74	methanethiol	1.53	1.70	0.08	0.91
2,2,2-trifluoroethanol	4.07	3.56	3.29	3.15	dimethyl sulfide	1.68	1.73	0.07	0.63
2-methyl-2-propanol	4.36	3.66	3.25	3.31	dimethyl disulfide	2.45	3.12	0.10	1.35
1,2-ethanediol	4.30	4.32	4.46	5.68	thioanisole	4.65	4.76	1.63	2.02
water	3.08	3.24	5.15	4.62	thiophene	2.69	2.85	0.90	1.04
phenol	5.21	6.31	3.60	4.85	pyridine	3.41	4.09	2.16	3.44
p-cresol	6.02	6.44	4.29	4.50	pyrazine	3.31	3.80	2.33	4.03
p-*tert*-butylphenol	6.80	7.70	3.09	4.39	pyrrole	4.46	4.26	4.05	3.51
dimethyl ether	1.40	1.49	1.47	1.39	3-methylindole	6.88	6.93	4.84	4.33
diethyl ether	3.26	2.06	1.56	1.17					

[a] References 3b and 23−25.

of hydration[23,24] and log *P*(octanol/water).[25,26] There are 85 data points in this case with an experimental range of 9 log units. A reasonable fit is obtained with four descriptors, as given by eq 9 ($r^2 = 0.87$, rms = 0.64). The results are listed in Table 4.

$$\log P(\text{octanol/gas}) = 0.02265(\text{SASA}) - 0.07030(\text{ESXC}) - 0.003704(\text{FOSA}) + 0.7553(\text{HBDN}) - 3.282 \quad (9)$$

SASA is again the most significant, which is probably true for most organic solvents. Nevertheless, it can be replaced by the Lennard-Jones energy, ESXL, with no change in r^2 or rms. A general measure of polarity, ESXC, is the next most important, followed by a correction to make hydrophobic compounds less soluble and a term for the number of donor hydrogen bonds. The Coulomb term can be replaced by the number of strong interactions, INST ($r^2 = 0.84$, rms = 0.71) INME ($r^2 = 0.83$, rms = 0.73), or the number of acceptor hydrogen bonds, HBAC ($r^2 = 0.81$, rms = 0.77), with modest degradations in the fit.

Though the four descriptors in eq 9 all have Prob > *F* values less than 0.0001, a three-descriptor fit can be obtained with $r^2 = 0.86$ and rms = 0.66 using ESXL, ESXC, and HBDN. The choice of descriptors is sensible. Solute polarity is expected to be more important in octanol than in hexadecane. Furthermore, the emergence of HBDN is reasonable, since an alcohol solvent has an excess of hydrogen-bond acceptor sites. In Table 4, the errors greater than 1.2 log units are for 1,4-dimethylpiperazine (1.22), aniline (−1.82), nitrobenzene (−1.28), and *N*,*N*-dimethylacetamide (1.26). If one adds a term for the number of nonconjugated tertiary amines and amides to eq 9, a fit with $r^2 = 0.89$ and rms = 0.59 is obtained.

Similar results are obtained for 76 of the compounds with the OPLS-AA charges. The same descriptors are the most significant and a three-descriptor fit with ESXL, ESXC, and HBDN yields $r^2 = 0.88$ and rms = 0.54. The largest errors are now for *N*-methylacetamide (1.25), 1,2-dimethoxyethane (1.24), phenol (1.15), and 1,2-ethanediol (1.12). Notably, amines are not problematic. Otherwise, it might have been tempting to argue that the discrepancies for amines with the CM1P charges stem

(24) Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 405−412.

(25) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR—Hydrophobic, Electronic, and Steric Constants*; American Chemical Society: Washington, DC, 1995.

(26) Sangster, J. *Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry*; Wiley: Chichester, U.K., 1997.

from uncertainties in the experimental data for the protonation state of the amines, though the experimental data for log *P*(octanol/water) of amines are normally either obtained at high pH or corrected for protonation in water. The better performance for amines with the OPLS-AA charges can be traced to lower Coulomb energies, ESXC, than those with the CM1P charges.

Water. The experimental data for log *P*(water/gas) for the 85 compounds cover a 9.2 unit range in Table 4.[23,24] A fit for eq 3 with the three descriptors from eq 2 yields $r^2 = 0.74$ and rms = 1.04. Furthermore, the SASA and Lennard-Jones terms are found not to be statistically significant; leaving them out does not alter the fit. ESXC is, in fact, the best single predictor of log *P*(water/gas), and the addition of any of the remaining descriptors has negligible impact on the fit. With just ESXC, its coefficient is −0.15 and the intercept is −0.16. Multiplying the coefficient by $-2.3RT$ yields a β value (eqs 1 and 2) of 0.20, which is lower than the theoretical 0.5 or our prior result of 0.314 with 6-31G* CHELPG charges.[12] With just ESXC, it is apparent that there is a problem with secondary and tertiary amines, which are not predicted to be hydrophilic enough. In view of the results for octanol, this is likely a reflection of insufficiently polar charge distributions for amines with the CM1P charges; e.g., the charges on nitrogen in methyl-, dimethyl-, and trimethylamine are −0.90, −0.78, and −0.63 with OPLS-AA[27] and −0.87, −0.65, and −0.42 with the scaled CM1P charges. On the other hand, nitro groups and to a lesser extent esters are predicted to be too hydrophilic with the scaled CM1P charges. The charges on nitrogen and oxygen in nitroalkanes are +0.54 and −0.37 with OPLS-AA and +1.30 and −0.68 with scaled CM1P. These opposing discrepancies lead to diminution of the β value. To significantly improve the fit, corrections are needed for the numbers of nitro groups and secondary and tertiary saturated amines. The hydrophobic surface area then also becomes a statistically significant descriptor, and eq 10 yields $r^2 = 0.89$ and rms = 0.67. If the FOSA term is left out, $r^2 = 0.87$ and rms = 0.73.

$$\log P(\text{water/gas}) = -0.1752(\text{ESXC}) + 2.211(\text{no. of amines-2,3}) - 2.486(\text{no. of nitro groups}) - 0.003256(\text{FOSA}) - 0.02234 \quad (10)$$

In this case, the results with the OPLS-AA charges are notably better. For 76 compounds, the best three-descriptor fit is obtained with the traditional terms of eq 2. ESXC, ESXL, and SASA are all significant, and eq 11 yields $r^2 = 0.90$ and rms = 0.64.

$$\log P(\text{water/gas}) = -0.3399(\text{ESXC}) - 0.5060(\text{ESXL}) - 0.02852(\text{SASA}) + 1.993 \quad (11)$$

The corresponding rms for the free energies of hydration of 0.87 kcal mol^{-1} shows improvement over the 0.99 kcal mol^{-1} that was obtained previously with the smaller set of 35 molecules using 6-31G* CHELPG charges.[12] The coefficient for ESXC multiplied by $-2.3RT$ becomes 0.46, which is close to the theoretical value for β of 0.5 in eq 1. This and the quality of the fit suggest that the OPLS-AA charges provide a more consistently accurate description of solutes' electrostatic potential fields than that obtained in the CM1P procedure. ESXL and SASA provide alternate measures of size that contribute in opposing manners. If SASA is dropped from eq 11, the coefficient for ESXL is reduced to −0.1227 ($r^2 = 0.83$ and rms = 0.82). SASA and ESXL can both be replaced in eq 11 by the solute's molecular weight and r^2 is still 0.83 and the rms is 0.83. There is also flexibility in representing the electrostatic term by ESXC, INME, or HBDA + HBAC. The measures of size in eq 11 are important; with just ESXC and a constant, the coefficient for ESXC becomes −0.25, $r^2 = 0.78$, and rms = 0.93. Thus, there is a cost for cavity formation in water that is usually more than offset by the gain in electrostatic interactions.

The dominance of the electrostatic term is apparent in aqueous solution. In turn, this emphasizes the need for accurate charge distributions. Though the OPLS-AA charges appear preferable to the CM1P ones, they are not available for all combinations of functional groups. Furthermore, even 85 compounds is a limited data set that contained, for example, only three amides and five compounds with more than one functional group. The results for the full data set of over 200 compounds are a far more stringent test; however, the test can only be performed for the octanol/water partition coefficients owing to the very limited experimental data for free energies of solvation for polyfunctional molecules.

Octanol/Water. The experimental data for log *P*(octanol/water) have been taken almost entirely from the recommended "∗" values in the compilation by Hansch et al.[25] The exceptions are the values for fluconazole, nifedipine, and nifuroxime, which were determined at Pfizer Inc.,[28] and the value for [2.2.2]-cryptand is that recommended by Sangster.[26] In all cases, the solutes are considered to be in their neutral, un-ionized forms. The data cover a range of 9.0 log units from ca. −4 to +5.

On the basis of the results above for octanol/gas and water/gas, it is clear that SASA and electrostatic terms are going to be most important, respectively. A fit for the CM1P results with just SASA and ESXC yields $r^2 = 0.68$ and rms = 0.97. Among the alternative measures for the electrostatics, the best replacement for ESXC is the number of accepted hydrogen bonds, HBAC ($r^2 = 0.77$ and rms = 0.83).

Analysis of the compounds with the larger errors quickly shows that significant improvement requires corrections for nitro groups and unconjugated amines, as found above for water and attributed to imperfections in the CM1P charges. With those additions $r^2 = 0.87$ and rms = 0.63. Furthermore, molecules with carboxylic acid groups are uniformly predicted to be too hydrophilic, which is again not surprising when CM1P and OPLS-AA charges are compared; e.g., the carbonyl oxygen and carbon charges in acetic acid are −0.53 and +0.52 with OPLS-AA and −0.62 and +0.65 with CM1P. There was only one acid, acetic acid, in the data set for free energies of hydration (Table 4). Both CO_2 and NO_2 fragments are overly polarized with CM1P. If they are corrected for together, eq 12 results, which has an r^2 of 0.90, an rms of 0.55, and a mean unsigned error of 0.44. The largest error among the 200 predicted values

$$\log P(\text{o/w}) = 0.01448(\text{SASA}) - 0.7311(\text{HBAC}) - 1.064(\text{no. of amines}) + 1.1718(\text{no. of nitro + acid groups}) - 1.772 \quad (12)$$

is 1.45 log units, and 93% of the predicted values are within 1.0 log unit of the experimental results. The results with eq 12 are compared with the experimental data in Tables 5−7 and Figure 1.

The inclusion of the 89 drugs does not substantially degrade the fit; a fit for only the other 111 compounds using the terms in eq 12 yields $r^2 = 0.90$ and rms = 0.48. For the 89 drugs in

(27) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827−4836.

(28) (a) Lombardo, F.; Shalaeva, M.; Tupper, K. A. *Symposium on Strategies for Optimizing Oral Drug Delivery: Scientific and Regulatory Approaches*; Kobe, Japan, April 19−21, 1999. (b) Jezequel, S. G. *J. Pharm. Pharmacol.* **1994**, *46*, 196−199.

Table 5. Experimental[a] and Predicted (Eq 12) log *P*(octanol/water) Values for Reference Organic Molecules

molecule	calcd	exptl	molecule	calcd	exptl
methane	0.57	1.09	dimethoxymethane	0.19	0.18
ethane	1.15	1.81	1,2-dimethoxyethane	0.16	−0.21
propane	1.63	2.36	anisole	1.91	2.11
butane	2.11	2.89	acetaldehyde	−0.23	0.06
pentane	2.59	3.39	propanal	0.12	0.59
hexane	2.95	3.90	benzaldehyde	1.40	1.48
cyclohexane	2.45	2.86	(*Z*)-3-penten-2-one	1.22	0.52
propene	1.47	1.77	acrolein	−0.18	−0.01
1,3-butadiene	1.73	1.99	acetone	−0.08	−0.24
1-pentyne	2.26	1.98	butanone	0.26	0.29
cyclohexene	2.33	2.86	acetophenone	1.10	1.58
ethyl chloride	1.55	1.43	acetic acid	−0.34	−0.17
1,1,1-trichloroethane	2.12	2.49	benzoic acid	1.92	1.87
1-chloropropane	1.98	2.04	methyl acetate	−0.07	0.18
1,2-dichloroethane	1.94	1.48	methyl butyrate	1.15	1.29
cis-1,2-dichloroethene	1.74	1.86	methyl benzoate	1.91	2.12
trans-1,2-dichloroethene	1.81	2.09	ethyl acetate	0.56	0.73
trichloroethene	2.03	2.42	methylamine	−1.03	−0.57
benzene	2.07	2.13	ethylamine	−0.40	−0.13
toluene	2.53	2.73	dimethylamine	−0.54	−0.38
naphthalene	3.06	3.35	piperidine	0.51	0.84
anthracene	4.06	4.45	trimethylamine	0.51	0.16
biphenyl	3.81	4.09	1-methylpyrrolidine	0.66	0.92
[2.2]paracyclophane	3.97	3.70	1,4-dimethyl-piperazine	−0.40	−0.40
fluorobenzene	2.19	2.27	aniline	1.59	0.90
chlorobenzene	2.46	2.89	acetonitrile	−0.49	−0.34
bromobenzene	2.48	2.99	propionitrile	0.29	0.16
hexafluorobenzene	2.72	2.55	benzonitrile	1.49	1.56
methanol	−0.68	−0.77	acetamide	−0.83	−1.26
ethanol	−0.12	−0.31	*N*-methylacetamide	−0.16	−1.05
1-propanol	0.66	0.25	urea	−1.32	−2.11
2-propanol	0.52	0.05	AcAlaNHMe C5	−0.29	−1.21
allyl alcohol	0.59	0.17	AcAlaNHMe C7eq	−0.36	−1.21
2,2,2-trifluoroethanol	0.93	0.41	*N*,*O*-dimethyl-carbamate	0.37	−0.06
2-methyl-2-propanol	1.33	0.35	nitromethane	0.18	−0.35
1,2-ethanediol	−0.59	−1.36	nitroethane	0.69	0.18
trans-1,2-cyclohexanediol	0.79	0.08	nitrobenzene	0.87	1.85
water	−1.36	−1.38	methanethiol	0.80	0.79
phenol	0.87	1.46	dimethyl sulfide	1.22	1.10
dimethyl ether	0.28	0.10	*N*-methylbenzene-sulfonamide	0.53	0.92
diethyl ether	1.08	0.89	dimethyl sulfoxide	−0.45	−1.35
tetrahydrofuran	0.39	0.46	dimethyl sulfone	−1.00	−1.34
morpholine	−0.92	−0.86	dimethyl disulfide	1.84	1.77
18-crown-6	−0.16	−0.68	thioanisole	2.76	2.74
[2.2.2]cryptand	−1.00	−0.10			

[a] References 25 and 26.

Table 6. Experimental[a] and Predicted (Eq 12) log *P*(octanol/water) Values for Aromatic Heterocycles

molecule	calcd	exptl	molecule	calcd	exptl
pyridine	0.85	0.65	adenine	−1.17	−0.09
pyrazine	0.50	−0.23	guanine	−1.69	−0.96
pyrimidine	0.18	−0.40	cytosine	−1.13	−1.73
pyridazine	−0.47	−0.72	uracil	−1.46	−1.07
1,3,5-triazine	0.37	−0.73	uridine	−0.97	−1.98
pyrrole	1.32	0.75	furan	0.78	1.34
3-methylindole	2.83	2.60	oxazole	0.08	0.12
imidazole	−0.86	−0.08	isoxazole	−0.67	0.08
9*H*-purine	−0.05	−0.37	thiazole	0.93	0.44
quinoline	2.51	2.09	thiophene	1.73	1.81

[a] Reference 25.

Table 7, the standard deviation is 0.62 log unit, the mean unsigned error is 0.49 log unit, and the largest error is 1.45 log units for prednisone. The only other drugs with errors above 1.1 log units are cimetidine (−1.14), diphenhydramine (−1.31), ethyl *p*-hydroxybenzoate (1.17), indoprofen (−1.16), phenobarbital (1.17), pindolol (−1.22), and progesterone (1.21). Except for the two steroids, there is no obvious pattern for these compounds, which are illustrated in Figure 2. For the seven steroids in the dataset, the average error is +0.64, so they are predicted to be too hydrophilic. The same is true for alkanes.

Table 7. Experimental[a] and Predicted (Eq 12) log *P*(octanol/water) Values for Drugs

molecule	calcd	exptl	molecule	calcd	exptl
acebutolol	1.82	1.71	lorazepam	2.67	2.39
acetaminophen	1.09	0.51	lovastatin	4.15	4.26
acyclovir	−1.32	−1.56	meloxicam	0.92	0.09
allopurinol	−1.56	−0.55	mepyramine	2.94	3.27
alprenolol	3.06	3.10	6-mercaptopurine	0.48	0.01
amantadine	1.49	2.44	methadone	3.87	3.93
amphetamine	1.76	1.76	moricizine	2.73	2.98
aspirin	1.17	1.19	morphine	1.13	0.76
atenolol	1.19	0.16	naproxen	3.07	3.34
atropine	2.13	1.83	nicotine	1.41	1.17
benzocaine	2.06	1.86	nifedipine	2.32	3.17
bifonazole	5.03	4.77	nifuroxime	0.41	1.28
bromazepam	2.54	1.69	omeprazole	2.38	2.23
bromopride	2.92	2.83	oxazepam	2.60	2.24
caffeine	0.12	−0.07	oxyprenolol	2.08	2.18
carbamazepine	2.73	2.45	penicillin G	1.61	1.83
chloramphenicol	1.93	1.14	perphenazine	3.57	4.20
chlorothiazide	−1.32	−0.24	phenacetin	1.88	1.58
chlorpheniramine	3.59	3.39	phencyclidine	4.40	3.63
chlorpromazine	4.77	5.19	phenobarbital	0.30	1.47
cimetidine	1.54	0.40	phenytoin	2.09	2.47
clonidine	2.45	1.57	pindolol	2.97	1.75
corticosterone	1.63	1.94	pirenzipine	0.83	0.10
desipramine	4.48	4.90	piroxicam	0.68	0.26
dexamethasone	1.63	1.83	prednisone	0.01	1.46
diazepam	3.51	2.99	procainamide	1.69	0.88
diethylstilbestrol	4.30	5.07	procarbazine	−0.04	0.06
diflumidone	2.85	2.86	progesterone	2.66	3.87
diphenhydramine	4.58	3.27	propranolol	3.13	3.09
estradiol	3.03	4.01	prostaglandin E2	3.20	2.82
ethyl *p*-hydroxybenzoate	1.30	2.47	proxicromil	3.82	4.40
fenfluramine	3.45	3.36	salicylic acid	2.16	2.26
fluconazole	0.34	0.50	scopolamine	1.09	1.24
flufenamic acid	4.84	5.25	serotonin	1.04	0.21
5-fluorouracil	−1.30	−0.89	sucrose	−3.41	−3.70
fluphenazine	4.10	4.36	sultopride	1.67	1.06
griseofulvin	1.42	2.18	testosterone	3.42	3.32
haloperidol	3.45	3.23	tetracaine	3.13	3.73
hydrocortisone	1.15	1.61	timolol	0.96	1.83
ibuprofen	3.98	3.50	trifluoperazine	4.49	5.03
imipramine	4.78	4.80	trimethoprim	0.12	0.91
indomethacin	4.23	4.27	verapamil	3.85	3.79
indoprofen	3.93	2.77	warfarin	2.68	2.70
ketoprofen	3.73	3.12	zidovudine (AZT)	−0.46	0.05
lidocaine	2.74	2.26			

[a] References 25 and 28.

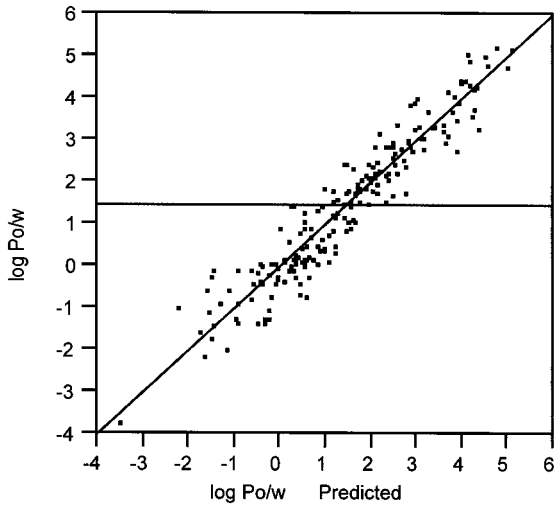


Figure 1. Experimental vs predicted (eq 12) log *P*(octanol/water) values for all 200 molecules.

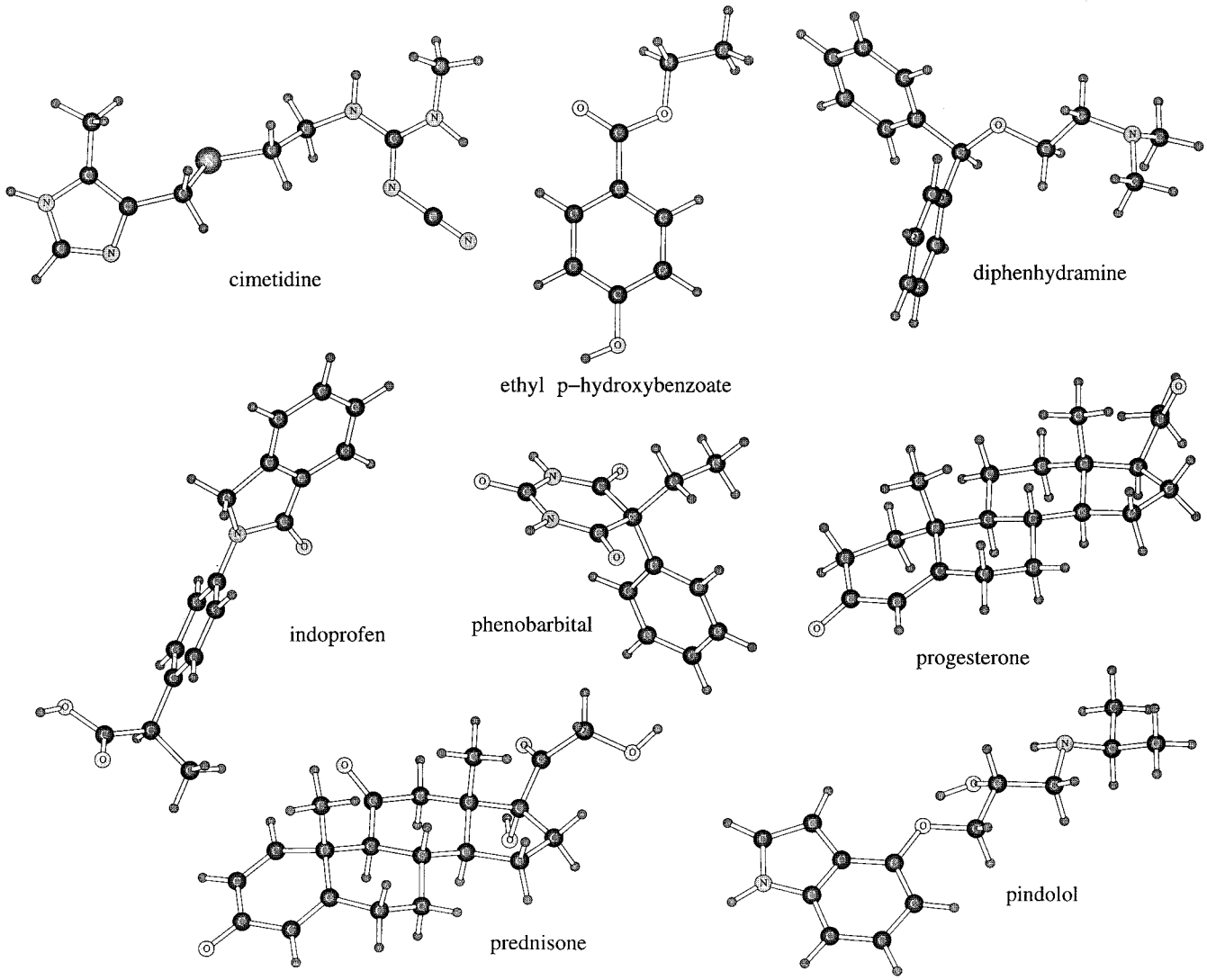


Figure 2. Structures of drugs with the largest errors in the predicted log *P*(octanol/water) values.

Nevertheless, if the fraction of hydrophobic surface area (FOSA/SASA) is considered as a descriptor, it does not improve the overall fit. Since the only nonzero descriptor for alkanes in eq 12 is SASA, a fit with just SASA and a constant for the seven alkanes yields log *P*(o/w) = 0.0160(SASA) − 1.478 with an r^2 of 0.97. Thus, log *P*(o/w) for alkanes increases more rapidly with size than for more polar compounds. One possible interpretation is that the lack of attractive interactions increases the effective size of alkanes in water and is associated with the formation of clathrate-like structures. The opposite tendency is obtained for amides, aliphatic alcohols, and azines, which are not predicted to be hydrophilic enough. The errors are not severe and can be associated with details of the CM1P charge distributions.

Overall, the diversity of the 200 molecules and the small number of terms in eq 12 are notable. The small number of variables helps ensure that the predictive ability of the model for new molecules should be good. Indeed, the cross-validated r^2, q^2, equals 0.89 for eq 12 using the 200 compounds in 20 random batches of 10. Furthermore, the predicted results for 30 additional compounds, which were not included in the training set, are compared with the experimental data in Table 8. This set includes 15 reference molecules and 15 drugs, pesticides, and herbicides. The mean unsigned error of 0.52 for the 30 compounds shows little variation from the 0.49 for the training set. The compounds with errors above 1 log unit are DDT (1.04), dibenzofuran (1.08), *N*,*N*-dimethylacetamide (1.27), sulindac (1.27), and 2,3,4,5,6-pentachlorobiphenyl (1.39). The model performs well for drug-like molecules, while it underestimates the log *P*(o/w) of highly lipophilic compounds such as DDT and polychlorinated biphenyls.

Table 8. Experimental[a] and Predicted (Eq 12) log *P*(octanol/water) Values for Additional Molecules Not in the Training Set

molecule	calcd	exptl	molecule	calcd	exptl
fluorene	3.85	4.18	cocaine	1.91	2.30
pyrene	4.21	4.88	desmedipham	3.85	3.39
hexamethylbenzene	4.21	4.61	fenbufen	3.66	3.20
p-cresol	1.76	1.94	fenoxycarb	4.24	4.30
p-*tert*-butylphenol	2.75	3.31	flurbiprofen	4.45	4.16
2,3-dichlorophenol	2.03	2.84	nitrofurantoin	0.08	−0.47
2-naphthol	2.22	2.70	sulindac	4.69	3.42
m-nitrobenzoic acid	1.97	1.83	terbutaline	0.45	0.08
p-chloroaniline	2.02	1.88	theophylline	−0.72	−0.02
benzamide	0.70	0.64	triflupromazine	5.37	5.19
N,*N*-dimethylacetamide	0.50	−0.77	2,3,4,5,6-PCB	5.13	6.52
acetanilide	1.53	1.16	DDT	5.87	6.91
p-toluenesulfonamide	−0.06	0.13	diuron	3.11	2.68
indole	2.46	2.14	atrazine	1.81	2.61
dibenzofuran	3.04	4.12	lindane	3.91	3.72

[a] Reference 25.

At this point, to go beyond eq 12, consideration of the remaining descriptors shows that a general measure of the

Table 9. Comparison of MC Results for Numbers of Donated and Accepted Hydrogen Bonds in Water and Ethanol Solutions at 25 °C and 1 atm

	HBDN		HBAC	
molecule	water	ethanol	water	ethanol
N-methylacetamide	0.8	1.0	2.7	2.0
diethyl ether	0.0	0.0	1.9	1.5
phenol	1.0	1.0	1.9	0.0
uracil	2.0	2.0	4.7	2.4
acetaminophen	2.0	2.0	3.1	1.6

electrostatics, INME or DIPL, is also fully significant (Prob > *F* is less than 0.0001) and brings the regression for eq 13 with

$$\log P(\text{o/w}) = 0.01469(\text{SASA}) - 0.5835(\text{HBAC}) - 1.089(\text{no. of amines}) + 1.097(\text{no. of nitro + acid groups}) - 0.1019(\text{INME}) - 1.809 \quad (13)$$

the 200 compounds to $r^2 = 0.91$ and rms = 0.53. However, the largest error rises to 1.63 log units (prednisone), while 93% of the predicted values still err by less than 1.0 log unit. Addition of any of the remaining descriptors from Table 1 to eq 13 provides no statistically significant improvement. Also, breaking down the SASA term into its three components does not improve the fit; the coefficients are essentially the same for the three components.

Treatment of the available OPLS-AA results for 146 of the compounds, which excludes about 50 of the drugs, confirmed the dominance of the SASA and HBAC terms; $r^2 = 0.86$ and rms = 0.62 for just a two-descriptor fit. A correction for nonconjugated amines is now statistically significant and gives $r^2 = 0.90$ and rms = 0.53. The principal outliers are amides and, if a correction is made for them, $r^2 = 0.91$ and rms = 0.49 for eq 14. On the basis of the experience with the CM1P

$$\log P(\text{o/w}) = 0.0150(\text{SASA}) - 0.8491(\text{HBAC}) - 0.9527(\text{no. of amines}) - 0.6515(\text{no. of amides}) - 1.794 \quad (14)$$

data set, the addition of the final 50 drugs can be anticipated to make the quality of the correlations with the CM1P or OPLS-AA charges very similar.

Analysis for Octanol/Water. The physical interpretation of the results is straightforward. A larger SASA favors solvation in octanol. It reflects the importance of van der Waals interactions in organic solvents, as also demonstrated by the dominance of this term for the hexadecane/gas and octanol/gas correlations. Greater electrostatics favor solvation in water. This can be represented in several ways, but the indices of solute hydrogen bonds are the best descriptors. The hydrogen-bond donor term cancels for octanol and water; both media have adequate hydrogen-bond-accepting ability to fully accommodate the generally small number of acidic hydrogens in a solute. The differential between water and alcohols is the greater hydrogen-bond-donating ability of water. This stems from the 2:1 ratio of protons to oxygens in water versus 1:1 in alcohols and the greater bulk of alcohols, which limits their ability to pack around a heteroatom that can accept more than one hydrogen bond. To illustrate the latter point, Monte Carlo simulations with CM1P charges were also performed for several of the compounds in ethanol solution.[29] The HBDN and HBAC results in Table 9 were obtained. Clearly, the striking difference between water and ethanol as solvents is in the greater number of hydrogen bonds donated by water.

The dominance of a size term and of a measure of hydrogen bonding has also been emphasized by Abraham et al.[23] They used the linear solvation energy relationship (LSER), eq 15,

$$\log P = d\delta_2 + s\pi_2 + a\alpha_2^{\text{H}} + b\beta_2^{\text{H}} + vV_x + \text{const} \quad (15)$$

which has been developed by Abraham, Kamlet, Taft, and coworkers,[30] and obtained good correlations for log *P*(hexadecane/water) and log *P*(octanol/water) for ca. 300 small molecules. The α and β parameters are measures of the solute's hydrogen-bond acidity and basicity, δ_2 is a polarizability correction for aromatics and polyhalo aliphatics, π_2^* reflects the solute's dipolarity, and V_x is the solute's volume. For octanol/water, the small a (−0.28) and much larger in magnitude b (−3.32) led to their conclusions that "the basicity of wet octanol must be almost the same as that of water" and "the hydrogen-bond acidity of wet octanol is appreciably less than that of water". The similar coefficient for the volume term for hexadecane/water and octanol/water also led them to conclude that "the cavity effect (or probably a combined cavity effect plus dispersion interactions) for wet octanol is not far away from that of hexadecane". The present results are completely consistent with this analysis, and the π, β, and V descriptors in eq 15 are closely related to INME, HBAC, and SASA in eqs 12 and 13. Though traditionally the LSER descriptors required experimental determination, a group contribution approach was recently reported for their estimation.[31] Buchwald and Bodor also recently noted the importance of hydrogen-bonding and volume terms in the development of their empirical method for predicting log *P*(octanol/water).[32] They also found that log *P*(o/w) increases more rapidly with size for alkanes than for other compounds; however, they needed a negative correction for steroids, which is opposite to the present tendency. A key advantage of the present approach is that the hydrogen-bond counts are fully automated and can be performed for any functionality.

Comparison with Other Methods for Predicting log *P*(octanol/water). Predictions were also obtained using several common, empirical procedures, CLOGP, LOGKOW, ALOGP, and MLOGP.[33] CLOGP is the fragment-based method of Hansch and Leo, which includes more than 200 fragment and correction terms.[34] LOGKOW is an expanded version of their approach with ca. 400 atom, fragment, and correction parameters; it is also known as KOWWIN and has a reduced incidence of molecules that cannot be processed.[35] ALOGP is the atom-based method of Ghose and Crippin that uses 110 descriptors,[36] and MLOGP is the notably simple approach of Moriguchi that uses only 13 parameters.[37] The present results (eq 12) are

(29) The OPLS-UA model was used with 260 ethanol molecules in a periodic cube at 25 °C and 1 atm: Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276−1284.

(30) Kamlet, M. J.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877−2887.

(31) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835−845.

(32) Buchwald, P.; Bodor, N. *Curr. Med. Chem.* **1998**, *5*, 353−380.

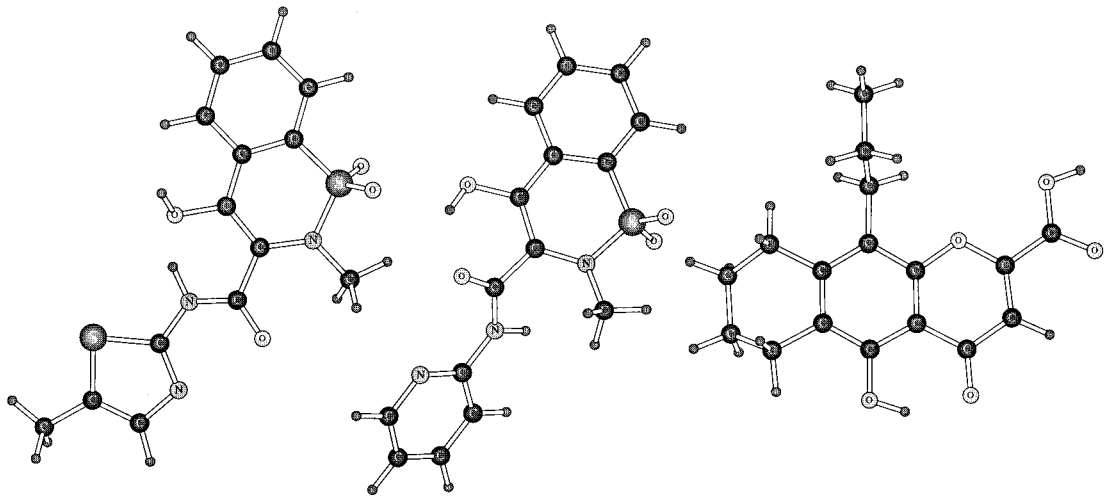
(33) CLOGP and LOGKOW values were computed with MedChem 3.55, Version 210 (Pomona College), and with KOWWIN, Version 1.57 (Syracuse Research Corp.), respectively. Calculations for ALOGP were performed with TSAR, Version 3.0 (Oxford Molecular, Inc.). The MLOGP results come from an in-house implementation of the algorithm in ref 37.

(34) Hansch, C.; Leo, A. *Exploring QSAR—Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(35) Meyland, W. M.; Howard, P. H. *J. Pharm. Sci.* **1995**, *84*, 83−92.

(36) Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 80−90.

(37) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. *Chem. Pharm. Bull.* **1992**, *40*, 127−130.



meloxicam	piroxicam	proxicromil	
0.9	0.7	3.8	eq 12
3.1	2.7	3.5	CLOGP
3.5	2.6	5.5	LOGKOW
–0.2	0.1	2.5	ALOGP
0.0	0.4	1.8	MLOGP
0.1	0.3	4.4	Exptl.

Figure 3. Comparison of predictions for log *P*(octanol/water) for three drugs.

Table 10. Comparison of Errors for Predictions of log *P*(octanol/water)[a]

method	r^2	rms	max error
MClogP (eq 12)	0.90	0.55	1.45
CLOGP	0.89	0.57	3.63
LOGKOW	0.90	0.53	3.41
ALOGP	0.87	0.62	2.93
MLOGP	0.78	0.80	4.43

[a] RSS = $\sum_i$(exptl log P − calcd log P)^2; $r^2 = 1 -$ RSS/TSS where TSS = $\sum_i$(exptl log P − (av exptl log P))^2; rms = (RSS/N)$^{1/2}$.

compared with those from these methods in Table 10. Only 198 of the 200 original compounds were included because [2.2.2]cryptand and procarbazine, which contains a hydrazine, could not be processed by all of the methods. The statistics used for the comparisons are the rms deviation and a correlation coefficient r^2, which is given by 1 − RSS/TSS, where RSS is the sum of the squared residuals and TSS is the sum of the squared deviations of the experimental values from their mean. It is clearly desirable for the rms to be small and for r^2 to approach 1.

The present procedure (eq 12) is seen to be competitive with the best alternative methods, CLOGP and LOGKOW. However, all the alternative procedures show significantly larger maximum errors. With CLOGP, the largest errors are for [2.2]paracyclophane (3.63) and meloxicam (3.01), while, with LOGKOW, they are for meloxicam (3.41) and piroxicam (2.32). With ALOGP, they occur for nifedipine (2.93) and griseofulvin (2.11), and with MLOGP, zidovudine (4.43) and [2.2]paracyclophane (3.35) are the most problematic. Most alternative methods have particular difficulties with meloxicam, piroxicam, and proxicromil, as summarized in Figure 3. All three molecules can form an internal hydrogen bond involving an enol fragment, which may be part of the trouble. These molecules and molecules such as the paracyclophane, which can easily be envisioned to be problematic for fragment-based approaches, do not cause unusual difficulties with the present method (eq 12).

Conclusion

This study has provided links between statistical mechanics simulations for solutes in solution, traditional physical−organic analyses, quantitative structure−property relationships (QSPR), and linear-response approaches for estimating free energies of solvation. It is found that a few physically significant descriptors from a simulation of a solute in water can be used to obtain correlations with r^2 values of 0.9 for free energies of solvation and transfer in a range of media extending from hexadecane to water. The very large numbers of descriptors that are featured in many fragment-based QSPRs and neural networks obscure the physical basis of solvation. The dominant terms are solute size in organic solvents and electrostatic interactions in water. The latter is well reflected in hydrogen-bond counts. Separate counts of donor and acceptor hydrogen bonds are required because water is exceptional in its ability to saturate the hydrogen-bond-accepting sites of a solute.

Owing to the importance of the electrostatic interactions, the simulation results are sensitive to the description of the charge distributions for the solutes. The practical benefits of a fully automated procedure using quantum mechanically derived charges such as CM1P are apparent, though improved procedures that avoid the need for corrections for some functional groups are desirable. In support of linear-response methods, the solute−water Coulomb energy was found to be the key descriptor for prediction of free energies of hydration. However, the significance of nonelectrostatic descriptors for this application is unclear. The high correlation between ESXL and SASA makes it generally unproductive to use both of them. Extension of linear-response approaches to protein−ligand binding with just ΔESXC and ΔESXL or ΔSASA in the scoring function is also unlikely to be optimal. Aside from the greater convergence

problems for ΔESXC, if one considers the protein environment to be akin to a moderately polar organic solvent like water-saturated octanol, changes in hydrogen-bond counts are expected to be especially valuable descriptors. Furthermore, depending on the choice of charges in the force field, it is likely that most force fields will have some problems with specific functional groups, e.g., amines, amides, and nitro compounds in particular.[27]

In addition to these general observations, the present study has yielded an automated tool that can be applied to predict readily a variety of solution-phase properties of organic solutes. Refinements through expansion of the descriptors and the collection of descriptors in other media can be considered. Application to prediction of additional properties of particular pharmaceutical relevance will also be described.

Acknowledgment. Gratitude is expressed to the National Science Foundation for support of this work, to Dr. Franco Lombardo for data and encouragement, to Dr. R. Daniel Meyer for statistical advice and assistance, and to Dr. Dongchul Lim for development of the autozmat utility for conversion of coordinate files to BOSS Z-matrix files.

Supporting Information Available: A table of the computed results for the 11 descriptors for the 230 compounds with the CM1P charges (PDF). This information is available free of charge via the Internet at http://pubs.acs.org.

JA993663T