The SGB/NP Hydration Free Energy Model Based on the Surface Generalized Born Solvent Reaction Field and Novel Nonpolar Hydration Free Energy Estimators

EMILIO GALLICCHIO, LINDA YU ZHANG, RONALD M. LEVY
Department of Chemistry and Chemical Biology, Rutgers University, Wright-Rieman Laboratories, 610 Taylor Rd, Piscataway, New Jersey 08854-8087

Received 6 July 2001; Accepted 6 November 2001

Abstract: The development and parameterization of a solvent potential of mean force designed to reproduce the hydration thermodynamics of small molecules and macromolecules aimed toward applications in conformation prediction and ligand binding free energy prediction is presented. The model, named SGB/NP, is based on a parameterization of the Surface Generalized Born continuum dielectric electrostatic model using explicit solvent free energy perturbation calculations and a newly developed nonpolar hydration free energy estimator motivated by the results of explicit solvent simulations of the thermodynamics of hydration of hydrocarbons. The nonpolar model contains, in addition to the more commonly used solvent accessible surface area term, a component corresponding to the attractive solute–solvent interactions. This term is found to be important to improve the accuracy of the model, particularly for cyclic and hydrogen bonding compounds. The model is parameterized against the experimental hydration free energies of a set of small organic molecules. The model reproduces the experimental hydration free energies of small organic molecules with an accuracy comparable or superior to similar models employing more computationally demanding estimators and/or a more extensive set of parameters.

 J Comput Chem 23: 517–529, 2002; DOI 10.1002/jcc.10045

Key words: solvation; Generalized Born model; hydrophobic hydration; hydration free energy prediction; solvent potential of mean force

Introduction

Hydration phenomena play an important role in virtually every process occurring in aqueous solution. Hydration has a particularly large effect on processes involving the breakage or formation of noncovalent bonds. The accurate description of the thermodynamics of hydration is, therefore, an essential element to model protein folding, binding, and conformational equilibria.[1–5]

Molecular simulations with explicit solvent models provide the most detailed and complete description of hydration phenomena.[6] They are, however, computationally demanding because of the large number of atoms involved, and the need to average over many solvent configurations to obtain meaningful thermodynamic parameters. Models that account implicitly for the presence of the solvent have a clear theoretical foundation, can be parameterized, and are not as computationally demanding. Implicit solvent models are based on the concept of the solvent potential of mean force, defined as the hydration free energy of a fixed conformation of the solute or of the binding pair.[7]

An accurate model of the water potential of mean force should reproduce the following main characteristics of hydration: (1) the affinity of ionic and polar groups for water, (2) the dielectric shielding provided by the water medium, and (3) hydrophobic interactions. All of these components, for instance, have been found to be important in characterizing the protein native state. Collectively, solvation effects favor the placement of polar residues on the protein surface and of nonpolar residues in the protein core, and disfavor intramolecular hydrogen bonds and salt bridges.[8,9]

In this article we present the development and parameterization of an implicit solvent model of hydration based on the Surface Generalized Born (SGB)[10] electrostatic model and a newly developed nonpolar hydration free energy estimator. The model is designed to be equally applicable to small molecules and macromolecules, a necessary feature for the model to be applicable to ligand binding studies. In this work we present the parameterization results of the model for small molecules.

The SGB model is based on the continuum dielectric representation of the solvent.[11,12] It provides a description the electrostatic

Correspondence to: E. Gallicchio; e-mail: emilio@hpcp.rutgers.edu

Contract/grant sponsor: National Institutes of Health; contract/grant numbers: GM30580 and RR06892

Contract/grant sponsor: High Permance Computing Project (Rutgers)

component of the solvent potential of mean force, including the electrostatic interactions between the solute atoms and the solvent and the solvent-mediated shielding of charge–charge interactions, being at the same time simple to implement and computationally efficient. The parameterization of the SGB model has been carried out using the results of explicit solvent free energy perturbation calculations.[13]

The nonpolar estimator is based on recent findings on the hydration thermodynamics of hydrocarbons.[14] It contains a component corresponding to the work of cavity formation and a component corresponding to the dispersion interactions between the solute atoms and the solvent. Solute–solvent dispersion interactions have been shown to be important even for regions of a macromolecule distant from the molecular surface.[15] The model is able to reproduce detailed characteristics of nonpolar hydration including the higher solubility of cyclic hydrocarbons relative to the corresponding linear species. The parameterization of the nonpolar model is carried out by fitting the sum of the SGB electrostatic and nonpolar estimators to the experimental hydration free energies of a database of small organic molecules.

Hydration Free Energy Model

The hydration free energy model we have developed is based in the electrostatic decomposition of the total hydration free energy ΔG_h,[12]

$$\Delta G_h = \Delta G_{np} + \Delta G_{el}, \tag{1}$$

where ΔG_{np} is the nonpolar free energy of hydration of the solute, defined as the hydration free energy of the solute when the atomic partial charges are set to zero, and ΔG_{el} is the electrostatic charging free energy of the solute, defined as the free energy change for reestablishing the solute atomic partial charges. Equation (1) follows from a thermodynamic cycle whereby the insertion of the solute in water from the gas phase is divided into three elementary processes: discharging the solute atoms in the gas phase, placing the uncharged solute in water, and finally, recharging the solute atoms in water. The solute is assumed to be rigid; this is a reasonable approximation for the parameterization of the SGB/NP model described here because most of the molecules that form the fitting database are small and relatively rigid.

Electrostatic Model

The electrostatic charging free energy ΔG_{el} of the solute in water is calculated using the SGB model,[10,13] the surface implementation of the generalized Born model.[16–19] The generalized Born equation

$$\Delta G_{el} = U_{\mathrm{SGB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_w}\right)\sum_{ij}\frac{q_i q_j}{f_{ij}(r_{ij})}, \tag{2}$$

where q_i is the charge of atom i and r_{ij} is the distance between atoms i and j, gives the electrostatic component of the free energy of transfer of a molecule with interior dielectric ϵ_{in} from vacuum to a continuum medium of dielectric constant ϵ_w, by interpolating between the two extreme cases that can be solved analytically: the one in which the atoms are infinitely separated and the other in which the atoms are completely overlapped. The interpolation function f_{ij} in eq. (2) is defined as

$$f_{ij} = [r_{ij}^2 + B_i B_j \exp(-r_{ij}^2/4B_iB_j)]^{1/2}, \tag{3}$$

where B_i is the Born radius of atom i defined as the effective radius that reproduces through the Born equation

$$U^i_{\mathrm{single}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_w}\right)\frac{q_i^2}{B_i}, \tag{4}$$

the electrostatic free energy, U^i_{single}, of the molecule when only the charge of atom i is turned on. The SGB method estimates U^i_{single} by integrating the interaction between atom i and the charge induced on the solute-solvent boundary, S, by the Coulomb field of the atom

$$U^i_{\mathrm{single}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_w}\right)\int_S \frac{q_i^2(\mathbf{r}-\mathbf{r}_i)}{|\mathbf{r}-\mathbf{r}_i|^4}\cdot\mathbf{n}(\mathbf{r})d^2\mathbf{r}, \tag{5}$$

where $\mathbf{n}(\mathbf{r})$ is the normal to the surface at $\mathbf{r}$. In this work we set $\epsilon_{in} = 1$.

The SGB method has been shown to compare well with the exact solution of the Poisson–Boltzmann (PB) equation. The SGB implementation used in this work includes further correction terms that improve the agreement between SGB and exact PB results.[10]

In the original parameterization of the SGB model[10] the atomic radii R_i that define the solute–solvent dielectric interface are set as the van der Waals radii. We find that this choice leads to a systematic over estimation of the magnitude of the electrostatic charging free energy when OPLS[20] charges are used, yielding differences between the experimental hydration free energies and the SGB electrostatic charging free energies in general too positive to be interpreted as pure nonpolar residuals. The SGB charging free energies were also found to be too negative with respect to explicit solvent free energy perturbation calculations.[13]

To correct for this, the atomic radii for the SGB model were optimized against the electrostatic charging free energies of 40 molecules obtained by explicit solvent free energy perturbation calculations.[13] We found that setting

$$R_i = a\frac{\sigma_i^{\mathrm{LJ}}}{2} + b, \tag{6}$$

where $a = 1.1$, $b = 0.05$ Å and σ_i^{LJ} is the OPLS-AA[20] Lennard–Jones diameter of the atom, gives results in good agreement with the free energy perturbation calculations. We have applied this parameterization of the SGB atomic radii to the calculation of the electrostatic component of the free energy of binding of an octapeptide to a domain of the MHC protein,[21] obtaining results in quantitative agreement with explicit solvent free energy perturbation calculations.

Nonpolar Model

Preliminary tests were performed to select a suitable expression for the nonpolar estimator. Explicit solvent simulations of the thermodynamics of hydration of small alkanes[14] showed that the hydration free energies of the alkanes can be decomposed into a surface area-dependent term and a surface area-independent term. The first term, representing the entropic and the solvent reorganization energy components associated with the formation of the solute cavity in water, is proportional to the solvent accessible surface area and the second term, representing the solute–solvent dispersion interaction energy component, depends only on the atomic composition of the solute.

This observation motivated a nonpolar fitting function of the following form,

$$\Delta G_{np} = \sum_{i=1}^{N} [\gamma(t_i) A_i + \alpha(t_i)], \tag{7}$$

where N is the number of atoms of the solute, A_i is the solvent accessible surface area of atom i and $\gamma(t_i)$ and $\alpha(t_i)$ are adjustable parameters dependent on the atom type t_i (the atom type assignment scheme is described below). In the calculation of the solvent-accessible surface area we set the atomic radii as half the Lennard–Jones σ_{LJ} parameter, the radius of hydrogens on heteroatoms is set to 0.86 Å. An offset of 1.4 Å, corresponding to the radius of a water molecule, is added to the atomic radii to locate the solvent-accessible surface.

In eq. (7) the first term, $\gamma(t_i)A_i$, represents the component of the hydration free energy proportional to the solvent accessible surface area of the solute. The γ parameter has the dimensions of a surface tension coefficient. The second term $\alpha(t_i)$ is introduced to represent effects that are not related to changes in surface area. This term is found to be particularly important to properly reproduce conformational free energy changes of hydrocarbons.[14,22]

According to eq. (1) the role of the nonpolar term ΔG_{np} is to reproduce the difference between the experimental hydration free energy ΔG_h and the electrostatic charging free energy ΔG_{el} estimated from model parameters. This residual will still retain some electrostatic features because the ΔG_{el} term is not expected to capture all the electrostatic effects contained in the experimental hydration free energies. For example, short range hydrogen bonding between solute atoms and first shell solvent molecules may not be represented well by eq. (2). Consequently, it is expected that the values of the optimized adjustable parameters in ΔG_{np} will also depend on the electrostatic properties of the solute (charge distribution, the ability to form hydrogen bonds, etc.) and will, in various degrees, differ from the values that could be predicted just on the basis of the nonpolar properties of the atom (such as size and the ability to form London dispersion interactions).

According to eq. (7) solute atoms that are not in direct contact with the solvent ($A_i = 0$) still contribute to the nonpolar hydration free energy through the constant surface area-independent term regardless of their position relative to the solvent. Explicit solvent results[14] indicate that this is physically correct for buried atoms in small molecules. However, deeply buried atoms in large molecules do not contribute significantly to the nonpolar hydration free energy. This issue is considered later by introducing a correction term dependent on the degree of "buriedness" of the target atom. In the next section we consider calculations on small molecules for which the correction term is not important.

Model Selection and Validation

The proposed hydration free energy model is tested against other possible alternatives by performing fitting/prediction exercises. Two databases[23] of the hydration free energy of, respectively, 99 and 93 small molecules are prepared containing many of the basic functional groups: alkanes, alkenes, alkynes, aromatics, alcohols, phenols, ethers, esters, acetals, amines, amides, ketones, aldehydes, carboxylic acids, nitriles, nitro compounds, aromatic heterocyclic, thiols, and sulfides. The 99 molecules set was used as a training set to fit the adjustable parameters, the second set (of 93 molecules) was used as a test set to probe the predictive abilities of the model.

The atom types assignment scheme is based on the OPLS[20] atom types. For the purpose of the fitting, the OPLS types present in the databases were reorganized into 17 types by similarity. This process was necessary to ensure a good representation of each type in the databases and was aided by singular value decomposition analysis to avoid overfitting. Tests indicated that there was no significant improvement in the predictive power of the model by increasing the number of types. The atom types used are listed in Table 1.

We have examined several scenarios to confirm the validity of the fitting function (7) and to identify an optimal atom type assignment scheme. In the first scheme, the fitting form in eq. (7) was used with the 17 atom types listed in Table 1. In the second scheme all atom types are constrained to have the same value of the γ parameter

$$\Delta G_{np} = \gamma A + \sum_{i=1}^{N} \alpha(t_i), \tag{8}$$

where $A = \sum_i A_i$ is the total SASA of the molecule. In the third scheme all atom types are constrained to have the same value of the α parameter

$$\Delta G_{np} = \sum_{i=1}^{N} \gamma(t_i) A_i + \alpha N. \tag{9}$$

The fourth scheme is the same as third except that α is set to zero

$$\Delta G_{np} = \sum_{i=1}^{N} \gamma(t_i) A_i. \tag{10}$$

In the fifth scheme all atom types are constrained to have the same value of both the γ and α parameters

$$\Delta G_{np} = \gamma A + \alpha N. \tag{11}$$

Table 1. Atom Types Used in the Preliminary Fitting Tests.

Type	OPLS Symbol(s)	Description
1	C	sp2 carbon in carbonyl and carboxylic group in ketones, aldehydes, amides, and carboxylic acids.
2	CM, C==	sp2 carbon in alkenes and dienes.
3	CA	sp2 carbon in aromatic rings.
4	CT, CY, CO	sp3 carbon.
5	CZ	sp carbon in nitriles and alkynes.[a]
6	H, HO, HS	hydrogen on heteroatoms.
7	HA	hydrogen on aromatic rings.
8	HC	hydrogen on sp3 carbons and on aldehyde carbonyl group.
9	N	sp2 nitrogen in amides.
10	NC	sp2 nitrogen in aromatic heterocyclic compounds.
11	NZ	sp nitrogen in nitriles.
12	NO, ON	nitrogen or oxygen in nitro (—NO_2) group.[b]
13	NT	sp3 nitrogen in ammonia and amines.
14	O	carbonyl and carboxy oxygen in ketones, aldehydes, amides, and carboxylic acids.
15	OH	oxygen on carbon carrying one hydrogen as in alcohols, diols, phenols, and carbonyl and carboxylic acids.
16	OS	oxygen in ethers and noncarbonyl oxygen in esters.
17	S, SH	sulfur in thiols and sulfides.
18	O2	oxygen in carboxylate group.
19	N3, N3A, NZ1	nitrogen in ammonium group.

[a]An all-atom OPLS type for sp aliphatic carbons is not available.
[b]As defined type 12 corresponds to the entire —NO_2 group.

Finally, in the sixth scheme we considered eq. (7) for a direct fitting to ΔG without the electrostatic term

$$\Delta G = \sum_{i=1}^{N} [\gamma(t_i) A_i + \alpha(t_i)]. \quad (12)$$

Equation (7) is the reference fitting function. The fitting forms in eqs. (8), (9), (10), and (11) are introduced to investigate the effect of constraining or eliminating some parameters. The fitting function (11) is closely related to the GB/SA[24] and FDPB/γ models.[25] These models implement a continuum electrostatic model coupled to a nonpolar term proportional to the surface area of the solute. Fitting function of the form as in eq. (12) with $\alpha_i = 0$ or $\alpha_i = \text{constant} = \alpha$,[26,27] or $\gamma_i = 0$[28] have been used to predict the hydration free energies of small molecules and macromolecules.

The fitting function of eq. (12) is different from all the other free energy estimators in that it attempts to reproduce directly the experimental hydration free energies rather than reproducing the nonpolar residuals resulting from subtracting the SGB electrostatic solvation free energy from the experimental hydration free energies. It cannot be assumed *a priori* that the use of the SGB electrostatic model necessarily improves the predictive ability of the solvation model, it may be the case, for example, that the accuracy of the SGB electrostatic solvation free energies model is so poor that the noise introduced in the nonpolar residues calculated from it degrades the overall predictive ability of the model. The analysis of the relative performance of the model derived from eq. (12) and eq. (7), which use the same number of adjustable parameters, will quantify the contribution of the SGB electrostatic model to the predictive ability of the solvation model.

The results of fitting the experimental hydration free energies of the training set of 99 molecules are summarized in Table 2; the corresponding prediction calculations on the test containing 93 molecules are shown in Table 3. The following conclusions can be drawn. The fitting function (7) is the most accurate in fitting the training set and in predicting the test set. The electrostatic term ΔG_{el} provides substantial information content; without it and even with the largest number available of adjustable nonpolar parameters [fitting function (12)] the errors are substantial especially in prediction. The average unsigned error increases from 0.588 kcal/mol using the nonpolar fitting function (7) to 1.392 kcal/mol when the experimental data is predicted directly using eq. (12) without using the SGB electrostatic model. Even more dramatically, the largest error in prediction increases from 2.37 kcal/mol to 11 kcal/mol when the SGB electrostatic term is excluded [see Table 3,

Table 2. Preliminary Fitting Results on 99 Molecules Training Set.

Fitting Function (Eq. Number)	Average Unsigned Error[a]	Average rms Error[a]	Largest Deviation[a]
(7)	0.350	0.48	1.67
(8)	0.589	0.76	2.50
(9)	0.640	0.86	3.57
(10)	0.650	0.84	3.84
(11)	0.882	1.12	4.05
(12)	0.548	0.72	2.42

[a]In kcal/mol.

compare fitting results (7) and (12)]. The electrostatic term aided by only two nonpolar parameters [fitting function (11)] is generally inadequate. We found, however, that the two-parameter fitting function is quite successful in predicting the hydration free energies of apolar compounds. Transfer of these nonpolar parameters to polar compounds introduces large errors. Increasing the number of parameters [fitting functions (10), (9), and (8)] gradually improves the quality of the fitting and of the prediction.

Extension to Macromolecules

Earlier it was observed that the fitting function in eq. (7) is not directly applicable to macromolecules because it neglects the loss of van der Waals interactions between the solvent molecules and a solute atom as the atom is removed from the solute surface and moved to the interior of the molecule. This effect is negligible for small molecules, but it becomes important for macromolecules. In this section we outline the extension of the model to macromolecules. Although in this work we apply the model only to small molecules, the model is parameterized using the features described in this section to support future development and applications of the model to macromolecules.

We have chosen to use the Born radius as a measure of how deeply an atom is buried in a molecule. This choice was guided by the following considerations. The Born radius is sensitive to the geometry of the solvent environment around each solute atom. The inverse of the Born radius of an atom measures the effective volume of solvent outside the solute weighted by r^{-4}, relative to the distance r of the solvent volume element from the atom. A depth parameter, based simply on the distance of the atom from the solute surface, would lack the ability to "sense" in the same way the effective distance and geometry of the solvent environment. The limiting values of the Born radius are well defined. The Born radius of an isolated atom is the radius of the atom itself. The Born radius of an atom in the interior of a globular shaped molecule is approximately the radius of the molecule. The Born radius depends in a complex manner on the position of the atom in the molecule and the size and shape of the molecule, but, in general, the deeper the atom is placed in the interior of the molecule the larger its Born radius. Finally, the Born radii are already computed during the evaluation of the SGB electrostatic term, and do not require an additional calculation step.

Table 3. Preliminary Prediction Results on 93 Molecules Test Set.

Fitting Function (Eq. Number)	Average Unsigned Error[a]	Average rms Error[a]	Largest Deviation[a]
(7)	0.588	0.76	2.37
(8)	0.664	0.87	2.80
(9)	0.769	1.03	3.76
(10)	0.725	0.99	3.27
(11)	4.376	4.93	11.51
(12)	1.392	1.99	11.00

[a]In kcal/mol.

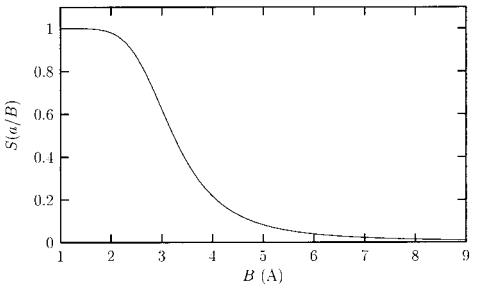

Figure 1. Switching function setting $a = 1.5$ Å, $c = 12$, and $b = 0.4$ as a function of the Born radius B.

The value of the Born radius is filtered through a switching function that multiplies the α parameters of the surface area-independent portion of the nonpolar estimator in such a way that as the Born radius increases the α contribution decreases and eventually vanishes for deeply buried atoms. The form of the switching function, as further explained below, was dictated by the constraint that the predicted hydration free energies of small molecules be changed as little as possible. Based on these considerations the following modified nonpolar fitting function was considered,

$$\Delta G_{np} = \sum_{i=1}^{N} [\gamma(t_i) A_i + \alpha(t_i) S(a/B_i)], \tag{13}$$

where B_i is the Born radius of the atom and $S(x)$ is a switching function

$$S(x) = \frac{1}{1 + (1/x)\exp[-c(x-b)]}, \tag{14}$$

where a, b, and c are positive adjustable parameters.

The switching function in eq. (14) is zero at $x = 0$ (infinitely large Born radius) and approaches 1 as $x \to \infty$ (infinitesimally small Born radius). It switches from 0 to 1 around $x = b$. The parameter c controls the rate at which the switch occurs. We selected c so that the switch occurs in an interval of roughly 0.5 units (see below). At $x = 1$ the switching function is designed to have reached more than 99% of its maximum value.

The a parameter is set so that a for a typical heavy atom completely exposed to the solvent $a/B_i \simeq 1$. The van der Waals radius of most heavy atoms is close to 1.5 Å so we set $a = 1.5$ Å. Given this value of a, for hydrogen atoms (Van der Waals radius $\simeq 1$ Å) completely exposed to the solvent $a/B_i > 1$. This is not a problem in practice, because the switching function is 1 in that region.

The b parameter is set to 0.4, this corresponds to reducing by 50% the α contribution from atoms having a Born radius two and a half times larger than their van der Waals radius. The switching function decreases rapidly for larger Born radius values.

The $S(a/B)$ function for $a = 1.5$ Å, $b = 0.4$ and $c = 12$ is shown in Figure 1 as a function of Born radius. An important property shown in Figure 1 is the plateau for small Born radii, ensuring that for the atoms in a small molecule $S(a/B_i) \simeq 1$. All

the atoms in the small molecule database used for the fitting have values of the switching function between 1 and 0.9. Consequently, the presence of the switching function has little effect on the γ and α parameters obtained from fitting the small molecules. This feature offers an important practical advantage because the parameterization of the γ and α parameters can be performed to a large degree independently from the switching function parameterization. The available experimental data on the gas solubilities of small molecules, however, is insufficient to optimize the switching function because the range of solute sizes is too limited within this database. The optimization of the switching function parameters for macromolecules will be carried out in future work.

Model Optimization

This section describes the optimization procedure used to parameterize the model outlined in the previous section. Starting with the reaction field atomic radii given by eq. (6) and the switching function parameters $b = 0.4$ and $c = 12$, the γ and α parameters in eq. (13) are obtained by fitting to a database of small molecules.[23] As further explained below, the reaction field atomic radii of some atom types were adjusted to yield reasonable experimental nonpolar free energies and the parameterization repeated.

The database of hydration free energies of small molecules used in the parameterization of the model is composed of 221 entries comprising alkanes, alkenes, alkynes, aromatics, alcohols, phenols, ethers, esters, acetals, amines, amides, ketones, aldehydes, carboxylic acids, nitriles, nitro compounds, aromatic heterocyclic, thiols, sulfides, and ammonium and carboxylate ions.[23] This database was constructed from the two databases used in an earlier section (less piperazine, see below) with the addition of 22 ionic compounds and 8 multifunctional compounds.

Fitting Procedure

Because the hydration free energies of the ionic compounds are much larger in magnitude than the hydration free energies of the neutral compounds, neutral and ionic compounds are parameterized separately. Doing otherwise would yield a parameterization dominated by the ionic functional groups. The parameters for the neutral compounds were obtained first, then the parameters relative to the ionic functional groups were obtained by fitting to the hydration free energies of the ionic compounds with fixed values for the parameters of the neutral functional groups.

The SGB electrostatic hydration free energies were calculated with the program IMPACT[29] using the atomic solvent exclusion radii developed previously[13] and further refined in this work (see below). OPLS atomic charges were assigned using the automatic atom-typing facility in IMPACT that assigns the closest OPLS atom type parameters based on an analysis of chemical functionalities and atomic hybridizations (a list of the assigned atomic charges for each molecule is available upon request).

The molecules were prepared in low internal energy conformations. The electrostatic free energies were calculated and subtracted from the experimental hydration free energies. The nonpolar function (13) was then parameterized by fitting the residuals by singular value decomposition, yielding a set of γ and α parameters.

For some classes of compounds (nitriles and ammonium ions) the experimental nonpolar free energies (the difference between the experimental hydration free energies and the SGB electrostatic charging free energies) appeared unreasonably large and positive. This indicated overestimation by the electrostatic model of the electrostatic charging free energy of these compounds. To correct for this, we increased by 14 and 34%, respectively, the atomic solvent exclusion radii of the NZ and N3 atom types given by eq. (6).

The predictive capabilities of the model were analyzed by performing a jackknife test. In 10 separate batches, 10% of the members of the database were randomly removed from the fitting set. The model was reparameterized using the remaining compounds and the hydration free energies of the removed compounds were predicted based on the resulting parameters. The predicted free energies were then compared to the experiments. The procedure was designed so that the hydration free energy of each molecule is predicted once. Separate jackknife tests were performed for the neutral and ionic compounds.

Results and Discussion

The fitting and jackknife results for the neutral and ionic compounds are shown in Tables 4 and 6. Tables 5 and 7 summarize the results of the fitting and jackknife tests. The corresponding γ and α parameters are listed in Table 8.

The predicted hydration free energies of the alkanes are very good. Given the small electrostatic contribution for these molecules, the results for the alkane set reflects the accuracy of the nonpolar estimator. The main characteristics on which the nonpolar estimator is based are reproduced. The hydration free energies of the normal alkanes increase linearly with chain length at the same rate as for the experiments. The hydration free energies of the cyclic alkanes are correctly predicted to be lower than the corresponding linear alkanes. The magnitude of the free energy gap between linear and cyclic alkanes is also correctly reproduced. Small discrepancies still exist for branched alkanes predicted to have generally slightly lower hydration free energies than the corresponding linear alkanes (rather than higher, as indicated by the experiments). The anomalously high hydration free energy of methane is not accurately reproduced; it is known that surface area models are not accurate for very small molecules such as methane.[14]

The model predicts the hydration free energies of unsaturated hydrocarbons very well. The trends in the experiments with respect to the number of double bonds, triple bonds, and conjugate double bonds are correctly reproduced. The hydration free energies of aromatic compounds are also correctly reproduced including complex fused ring systems. The surprisingly favorable experimental hydration free energies of multiple alkyl-substituted benzene molecules are not predicted as accurately, the model does its best in fitting the hydration free energies of these molecules, but in so doing it underestimates the favorable hydration free energy of benzene itself.

The hydration free energies of alcohols are reproduced with very good accuracy, including primary, secondary, and tertiary alcohols, cyclic alcohols, and diols. The hydration free energies of phenols are also accurately reproduced.

Table 4. Fitting and Prediction Results for the Hydration Free Energies of the Neutral Compounds.

Molecule	Experiment[a]	Fit[a,b]	Prediction[a,b]
		Alkanes	
Ethane	1.830	1.698 (−0.132)	1.670 (−0.160)
Propane	1.960	1.882 (−0.078)	1.796 (−0.164)
Butane	2.080	2.090 (0.010)	2.120 (0.040)
Isobutane	2.320	1.934 (−0.386)	1.914 (−0.406)
Pentane	2.330	2.227 (−0.103)	2.244 (−0.086)
2-Methylbutane	2.380	2.089 (−0.291)	2.080 (−0.300)
Neopentane	2.500	2.000 (−0.500)	1.971 (−0.529)
Hexane	2.490	2.518 (0.028)	2.446 (−0.044)
2-Methylpentane	2.520	2.313 (−0.207)	2.302 (−0.218)
3-Methylpentane	2.510	2.270 (−0.240)	2.269 (−0.241)
2,3-Dimethylbutane	2.590	2.121 (−0.469)	2.080 (−0.510)
Cyclopropane	0.750	0.852 (0.102)	0.875 (0.125)
Cyclopentane	1.200	1.073 (−0.127)	1.025 (−0.175)
Methylcyclopentane	1.600	1.267 (−0.333)	1.266 (−0.334)
Cyclohexane	1.230	1.262 (0.032)	1.333 (0.103)
Methylcyclohexane	1.710	1.438 (−0.272)	1.388 (−0.322)
Methane	1.910	1.438 (−0.472)	1.402 (−0.508)
Heptane	2.620	2.741 (0.121)	2.783 (0.163)
Octane	2.890	2.995 (0.105)	3.000 (0.110)
		Alkenes and Dienes	
Ethylene	1.270	1.489 (0.219)	1.656 (0.386)
1-Propene	1.270	1.215 (−0.055)	1.158 (−0.112)
1-Butene	1.380	1.372 (−0.008)	1.362 (−0.018)
2-Methyl-2-butene	1.310	1.138 (−0.172)	1.026 (−0.284)
1-Hexene	1.680	1.807 (0.127)	1.801 (0.121)
trans-2-Pentene	1.340	1.126 (−0.214)	1.013 (−0.327)
2-Methyl-1,3-butadiene	0.680	0.679 (−0.001)	0.692 (0.012)
1,4-Pentadiene	0.940	0.609 (−0.331)	0.531 (−0.409)
Cyclopentene	0.560	0.230 (−0.330)	0.167 (−0.393)
Butadiene	0.600	0.564 (−0.036)	0.555 (−0.045)
1-Pentene	1.400	1.100 (−0.300)	1.064 (−0.336)
		Alkynes	
Acetylene	−0.010	−0.066 (−0.056)	−0.511 (−0.501)
1-Propyne	−0.310	0.043 (0.353)	0.122 (0.432)
1-Butyne	−0.160	−0.005 (0.155)	0.023 (0.183)
1-Pentyne	0.010	0.119 (0.109)	0.168 (0.158)
1-Hexyne	0.290	0.214 (−0.076)	0.190 (−0.100)
1-Buten-3-yne	0.040	−0.428 (−0.468)	−0.577 (−0.617)
		Arenes	
Benzene	−0.870	−0.286 (0.584)	−0.120 (0.750)
Toluene	−0.890	−0.608 (0.282)	−0.578 (0.312)
o-Xylene	−0.900	−1.313 (−0.413)	−1.383 (−0.483)
m-Xylene	−0.840	−1.270 (−0.430)	−1.331 (−0.491)
p-Xylene	−0.810	−1.112 (−0.302)	−1.061 (−0.251)
Ethylbenzene	−0.800	−0.377 (0.423)	−0.292 (0.508)
1,2,4-Trimethylbenzene	−0.860	−1.602 (−0.742)	−1.699 (−0.839)
Biphenyl	−2.640	−2.426 (0.214)	−2.216 (0.424)
Naphthalene	−2.390	−1.680 (0.710)	−1.622 (0.768)
1-Methylnaphthalene	−2.370	−2.091 (0.279)	−1.944 (0.426)
1,3-Dimethylnaphthalene	−2.470	−2.514 (−0.044)	−2.486 (−0.016)
1,4-Dimethylnaphthalene	−2.820	−2.562 (0.258)	−2.539 (0.281)
2,3-Dimethylnaphthalene	−2.780	−2.226 (0.554)	−2.156 (0.624)
2,7-Dimethylnaphthalene	−2.630	−2.227 (0.403)	−2.179 (0.451)
Fluorene	−3.440	−3.670 (−0.230)	−3.547 (−0.107)
Phenanthrene	−3.950	−4.466 (−0.516)	−4.151 (−0.201)
Pyrene	−4.460	−5.230 (−0.770)	−5.338 (−0.878)

Table 4. (continued)

Molecule	Experiment[a]	Fit[a,b]	Prediction[a,b]
Acenaphthalene	−3.150	−2.673 (0.477)	−2.390 (0.760)
Anthracene	−4.230	−4.262 (−0.032)	−4.024 (0.206)
Isopropylbenzene	−0.300	−0.040 (0.260)	0.012 (0.312)
t-Butylbenzene	−0.440	−0.281 (0.159)	−0.233 (0.207)
		Alcohols and Phenols	
Methanol	−5.110	−5.047 (0.063)	−5.044 (0.066)
Ethanol	−5.010	−5.322 (−0.312)	−5.377 (−0.367)
1-Propanol	−4.830	−4.794 (0.036)	−4.788 (0.042)
2-Propanol	−4.760	−4.844 (−0.084)	−4.917 (−0.157)
1-Butanol	−4.720	−4.684 (0.036)	−4.704 (0.016)
2-Methyl-1-propanol	−4.520	−4.764 (−0.244)	−4.821 (−0.301)
2-Butanol	−4.580	−4.644 (−0.064)	−4.657 (−0.077)
t-Butanol	−4.510	−4.661 (−0.151)	−4.626 (−0.116)
1-Pentanol	−4.470	−4.337 (0.133)	−4.391 (0.079)
3-Methyl-1-butanol	−4.420	−4.471 (−0.051)	−4.505 (−0.085)
2-Pentanol	−4.390	−4.493 (−0.103)	−4.540 (−0.150)
3-Pentanol	−4.350	−3.718 (0.632)	−3.719 (0.631)
2-Methyl-2-butanol	−4.430	−4.273 (0.157)	−4.233 (0.197)
1-Hexanol	−4.360	−4.167 (0.193)	−4.145 (0.215)
2,3-Dimethyl-2-butanol	−3.910	−3.770 (0.140)	−3.770 (0.140)
3-Hexanol	−4.080	−3.262 (0.818)	−3.204 (0.876)
4-Methyl-2-pentanol	−3.740	−4.086 (−0.346)	−4.084 (−0.344)
2-Methyl-3-pentanol	−3.890	−4.044 (−0.154)	−3.998 (−0.108)
2-Methyl-2-pentanol	−3.930	−4.171 (−0.241)	−4.194 (−0.264)
Cyclopentanol	−5.490	−5.533 (−0.043)	−5.560 (−0.070)
Cyclohexanol	−5.480	−5.662 (−0.182)	−5.714 (−0.234)
Phenol	−6.620	−6.217 (0.403)	−6.113 (0.507)
2-Methylphenol	−5.870	−6.323 (−0.453)	−6.399 (−0.529)
4-Methylphenol	−6.140	−6.513 (−0.373)	−6.568 (−0.428)
4-*t*-Butylphenol	−5.920	−6.470 (−0.550)	−6.687 (−0.767)
Ethandiol	−9.600	−10.031 (−0.431)	−10.068 (−0.468)
2-Propene-1-ol	−4.800	−4.464 (0.336)	−4.445 (0.355)
		Ethers	
Dimethyl ether	−1.900	−1.498 (0.402)	−1.492 (0.408)
Diethyl ether	−1.630	−1.596 (0.034)	−1.544 (0.086)
Methyl n-propyl ether	−1.660	−1.155 (0.505)	−1.145 (0.515)
Methyl isopropyl ether	−2.010	−1.627 (0.383)	−1.669 (0.341)
Ethyl n-propyl ether	−1.810	−1.408 (0.402)	−1.373 (0.437)
Methyl t-butyl ether	−2.210	−1.490 (0.720)	−1.405 (0.805)
Di-*n*-propyl ether	−1.150	−1.026 (0.124)	−1.032 (0.118)
Di-sopropylether	−0.530	−1.450 (−0.920)	−1.507 (−0.977)
Di-*n*-butyl ether	−0.830	−0.459 (0.371)	−0.456 (0.374)
Tetrahydrofuran	−3.470	−2.172 (1.298)	−2.343 (1.127)
2-Methyltetrahydrofuran	−3.300	−2.353 (0.947)	−2.268 (1.032)
2,5-Dimethyltetrahydrofuran	−2.920	−2.323 (0.597)	−2.389 (0.531)
Phenyl methyl ether	−1.040	−2.139 (−1.099)	−2.214 (−1.174)
1,3-Dioxalane	−4.100	−4.142 (−0.042)	−4.254 (−0.154)
1,4-Dioxane	−5.050	−6.678 (−1.628)	−6.949 (−1.899)
Tetrahydropyran	−3.120	−2.138 (0.982)	−2.074 (1.046)
2-Methoxy-1-ethanol	−6.800	−6.681 (0.119)	−6.638 (0.162)
		Ketones and Aldehydes	
Acetone	−3.850	−3.685 (0.165)	−3.703 (0.147)
2-Butanone	−3.640	−3.362 (0.278)	−3.398 (0.242)
2-Pentanone	−3.530	−3.405 (0.125)	−3.418 (0.112)
3-Pentanone	−3.410	−3.437 (−0.027)	−3.446 (−0.036)
3-Methyl-2-butanone	−3.240	−3.554 (−0.314)	−3.598 (−0.358)
2-Hexanone	−3.290	−3.235 (0.055)	−3.217 (0.073)
4-Methyl-2-pentanone	−3.060	−2.923 (0.137)	−2.892 (0.168)

Table 4. (continued)

Molecule	Experiment[a]	Fit[a,b]	Prediction[a,b]
2-Heptanone	−3.040	−2.849 (0.191)	−2.833 (0.207)
4-Heptanone	−2.930	−2.893 (0.037)	−2.862 (0.068)
2,4-Dimethyl-3-pentanone	−2.740	−3.492 (−0.752)	−3.669 (−0.929)
Acetophenone	−4.580	−4.271 (0.309)	−4.228 (0.352)
Acetaldehyde	−3.500	−3.611 (−0.111)	−3.575 (−0.075)
Propanal	−3.440	−3.270 (0.170)	−3.252 (0.188)
Butanal	−3.180	−2.955 (0.225)	−2.947 (0.233)
Pentanal	−3.030	−2.877 (0.153)	−2.840 (0.190)
Hexanal	−2.810	−2.578 (0.232)	−2.668 (0.142)
Trans-2-Butenal	−4.230	−3.793 (0.437)	−3.752 (0.478)
trans,trans-2,4-Hexadienal	−4.630	−4.026 (0.604)	−3.744 (0.886)
Benzaldehyde	−4.020	−4.218 (−0.198)	−4.259 (−0.239)
p-Hydroxybenzaldehyde	−10.480	−10.557 (−0.077)	−10.599 (−0.119)
Ethanal	−3.500	−3.574 (−0.074)	−3.570 (−0.070)
meta-Hydroxy-benzaldehyde	−9.500	−9.824 (−0.324)	−9.842 (−0.342)
Carboxylic Acids			
Acetic acid	−6.700	−6.894 (−0.194)	−7.278 (−0.578)
Propionic acid	−6.480	−6.857 (−0.377)	−7.125 (−0.645)
Butyric acid	−6.360	−6.497 (−0.137)	−6.471 (−0.111)
Esters			
Methyl acetate	−3.320	−3.068 (0.252)	−3.043 (0.277)
Ethylacetate	−3.100	−3.264 (−0.164)	−3.267 (−0.167)
n-Propyl acetate	−2.860	−2.923 (−0.063)	−2.894 (−0.034)
Isopropyl acetate	−2.650	−3.094 (−0.444)	−3.173 (−0.523)
Methyl propanoate	−2.930	−2.861 (0.069)	−2.870 (0.060)
Ethyl propanoate	−2.800	−2.675 (0.125)	−2.677 (0.123)
n-Propyl propanoate	−2.450	−2.622 (−0.172)	−2.608 (−0.158)
Isopropyl propanoate	−2.220	−2.308 (−0.088)	−2.266 (−0.046)
Methyl butanoate	−2.830	−2.619 (0.211)	−2.598 (0.232)
Ethyl butanoate	−2.500	−2.547 (−0.047)	−2.555 (−0.055)
n-Propyl butanoate	−2.280	−2.323 (−0.043)	−2.296 (−0.016)
Methyl pentanoate	−2.570	−2.302 (0.268)	−2.321 (0.249)
Ethyl pentanoate	−2.520	−2.276 (0.244)	−2.274 (0.246)
Methyl benzoate	−4.280	−3.959 (0.321)	−4.011 (0.269)
Ethyl formate	−2.650	−3.653 (−1.003)	−3.941 (−1.291)
Amines			
Methyl amine	−4.560	−4.369 (0.191)	−4.376 (0.184)
Ethyl amine	−4.500	−4.779 (−0.279)	−4.853 (−0.353)
n-Propyl amine	−4.390	−4.297 (0.093)	−4.260 (0.130)
n-Butyl amine	−4.290	−4.247 (0.043)	−4.242 (0.048)
n-Pentyl amine	−4.100	−3.844 (0.256)	−3.834 (0.266)
n-Hexyl amine	−4.030	−3.759 (0.271)	−3.787 (0.243)
Dimethyl amine	−4.290	−4.192 (0.098)	−4.241 (0.049)
Diethyl amine	−4.070	−4.129 (−0.059)	−4.142 (−0.072)
Di-*n*-propyl amine	−3.660	−3.474 (0.186)	−3.425 (0.235)
Di-*n*-butyl amine	−3.330	−2.976 (0.354)	−2.830 (0.500)
Trimethyl amine	−3.240	−3.598 (−0.358)	−3.692 (−0.452)
Triethyl amine	−3.020	−2.027 (0.993)	−1.779 (1.241)
Aziridine	−5.420	−5.562 (−0.142)	−5.550 (−0.130)
Azetidine	−5.560	−5.254 (0.306)	−5.193 (0.367)
Pyrrolidine	−5.480	−5.515 (−0.035)	−5.553 (−0.073)
Piperidine	−5.110	−5.530 (−0.420)	−5.544 (−0.434)
N-Methylpyrrolidine	−3.980	−4.242 (−0.262)	−4.315 (−0.335)
N-Methylpiperidine	−3.890	−3.960 (−0.070)	−4.039 (−0.149)
Morpholine	−7.170	−8.003 (−0.833)	−8.080 (−0.910)
4-Methylmorpholine	−6.340	−6.227 (0.113)	−6.169 (0.171)
Ammonia	−4.310	−4.074 (0.236)	−3.730 (0.580)
Aniline	−4.900	−5.006 (−0.106)	−4.986 (−0.086)

Table 4. (continued)

Molecule	Experiment[a]	Fit[a,b]	Prediction[a,b]
1-Amino-2-methoxy-ethane	−6.600	−7.101 (−0.501)	−7.221 (−0.621)
		Amides	
Acetamide	−9.710	−9.555 (0.155)	−9.357 (0.353)
Propionamide	−9.410	−9.412 (−0.002)	−9.570 (−0.160)
N-Methylacetamide	−10.080	−10.336 (−0.256)	−10.472 (−0.392)
N-Methyl formamide	−10.000	−10.207 (−0.207)	−10.333 (−0.333)
N,N-Dimethyl-acetamide	−8.500	−8.179 (0.321)	−7.720 (0.780)
		Nitriles	
Acetonitrile	−3.890	−3.733 (0.157)	−1.607 (2.283)
Propiononitrile	−3.850	−4.517 (−0.667)	−5.196 (−1.346)
Butyronitrile	−3.640	−4.030 (−0.390)	−4.641 (−1.001)
3-Hydroxybenzonitrile	−9.670	−9.643 (0.027)	−9.612 (0.058)
4-Hydroxybenzonitrile	−10.170	−9.297 (0.873)	−9.097 (1.073)
		Nitrohydrocarbons	
Nitroethane	−3.710	−4.380 (−0.670)	−7.004 (−3.294)
1-Nitropropane	−3.340	−4.122 (−0.782)	−6.547 (−3.207)
2-Nitropropane	−3.140	−4.043 (−0.903)	−9.023 (−5.883)
Nitrobenzene	−4.120	−3.129 (0.991)	−2.831 (1.289)
3-Nitrophenol	−9.630	−9.564 (0.066)	−10.929 (−1.299)
4-Nitrophenol	−10.650	−9.349 (1.301)	−10.195 (0.455)
		Nitrogen Heterocyclic	
Pyridine	−4.700	−5.649 (−0.949)	−5.839 (−1.139)
2-Methylpyridine	−4.630	−4.798 (−0.168)	−4.786 (−0.156)
3-Methylpyridine	−4.770	−5.049 (−0.279)	−5.075 (−0.305)
2-Etylypyridine	−4.330	−4.293 (0.037)	−4.399 (−0.069)
3-Ethylpyridine	−4.600	−4.575 (0.025)	−4.641 (−0.041)
2,3-Dimethylpyridine	−4.830	−4.105 (0.725)	−3.988 (0.842)
3,5-Dimethylpyridine	−4.840	−4.125 (0.715)	−3.927 (0.913)
2-Methylpyrazine	−5.520	−5.779 (−0.259)	−5.874 (−0.354)
2-Ethylpyrazine	−5.460	−5.197 (0.263)	−5.112 (0.348)
3-Ethyl-2-methoxy-pyrazine	−4.400	−4.168 (0.232)	−3.965 (0.435)
2,6-Dimethylpyridine	−4.600	−4.685 (−0.085)	−4.633 (−0.033)
2,6-Di-*t*-butylpyridine	−0.410	−0.915 (−0.505)	−1.360 (−0.950)
N-Methyl-2-pyridone	−10.000	−10.028 (−0.028)	−20.251 (−10.251)
		Thiols	
Methanethiol	−1.240	−0.744 (0.496)	−0.325 (0.915)
Ethanethiol	−1.300	−0.689 (0.611)	−0.288 (1.012)
Benzenethiol	−2.550	−3.609 (−1.059)	−4.334 (−1.784)
		Sulfides	
Dimethyl sulfide	−1.540	−1.176 (0.364)	−1.019 (0.521)
Diethyl sulfide	−1.430	−1.164 (0.266)	−0.798 (0.632)
Methyl phenyl sulfide	−2.730	−3.404 (−0.674)	−3.808 (−1.078)

The predicted results are obtained by the jackknife procedure explained in the text.
[a]In kcal/mol.
[b]Difference from experiment in parenthesis.

The accuracy achieved for ethers is generally good but not as good as with the other molecular series. The effect of highly branched alkyl substituents is generally qualitatively reproduced. The hydration free energies of tetrahydrofurans is predicted to be less negative than the experiments. The hydration free energy of 1,3-dioxalane, a five-atom ring compound with two oxygen atoms, is well reproduced whereas the fitted hydration free energy of the corresponding six-atom ring compound, 1,4-dioxane, is too negative. The hydration free energy of the multifunctional compound 1-methoxy-1-ethanol is correctly reproduced. We suspect that some of the problematic cases are due to the conformational dependence of the hydration free energy, which is not explicitly included in our parameterization. The electrostatic interaction of the ether functional group with the water solvent will depend on the specific positioning of the surrounding alkyl groups. The dipole moment of flexible molecules containing two oxygen atoms will

Table 5. Mean Unsigned Errors for the Fitted and Predicted Hydration Free Energies of the Neutral Compounds.

Molecular Class	Number of Molecules	Fit[a]	Prediction[a]
Compounds containing only C and H			
Alkanes	19	0.21	0.24
Alkenes and dienes	11	0.16	0.22
Alkynes	6	0.20	0.33
Arenes	21	0.38	0.44
Subtotal	57	0.26	0.32
Compounds containing only C, H, and O			
Alcohols	27	0.25	0.28
Ethers	17	0.62	0.66
Ketones and aldehydes	22	0.23	0.26
Carboxylic acids	3	0.23	0.44
Esters	15	0.23	0.25
Subtotal	103	0.29	0.33
Compounds containing only C, H, O, and N			
Amines	23	0.27	0.33
Amides	5	0.19	0.40
Nitriles	5	0.42	1.15
Nitro compounds	6	0.78	2.57
Nitrogen heterocyclic	13	0.33	1.22
Subtotal	52	0.35	0.89
Compounds containing C, H, O, N, and S			
Thiols	3	0.72	1.24
Sulfides	3	0.43	0.74
Subtotal	6	0.57	0.99
Total	199	0.32	0.50

[a]In kcal/mol.

also depend on conformation. The dipole moment of 1,4-dioxane, for example, is small in the chair conformation and large in the boat conformation; this can introduce conformational effects difficult to capture with a model based on a single conformation.

The fitted and predicted hydration free energies of aldehydes and ketones are generally in very good agreement with the experiments. Minor discrepancies are observed for branched ketones and aldehydes unsaturated in the β position. Notice the good accuracy achieved for the aromatic and multifunctional aldehydes, benzaldehyde, *para*-hydroxybenzaldehyde and *meta*-hydroxybenzaldehyde. The hydration free energies of carboxylic acids and esters are very well reproduced. This is a remarkable result given that no additional atom types are introduced for these functional groups (the carboxylic groups is assumed composed of carbonyl carbon and oxygen and a hydroxy group, the ester group is assumed composed of carbonyl carbon and oxygen and a ether oxygen).

The hydration free energy of the amines are reproduced with very good accuracy. The ranking between ammonia and the methylated amines is in agreement with the experiments. The puzzling dependence of the hydration free energy of the amines on degree of methylation has received significant attention,[30–32] only after a recent OPLS-AA charge parameterization of the amines,[33] it has been possible to predict reliably the hydration free energies of alkyl-substituted amines. The hydration free energies of cyclic, aromatic, and polyfunctional amines are also well reproduced. We were not, however, able to correctly fit the hydration free energy of piperazine, a six-member saturated ring with two nitrogen atoms in 1,4 relative position. Other models fail in the same way to predict accurately the hydration free energy of this molecule.[32] It was necessary to remove the piperazine compounds from the training database to accurately fit the other amines. The experimental hydration free energy of piperazine is -7.37 kcal/mol, the predicted hydration free energy of piperazine is significantly more negative (-11.82 kcal/mol). The discrepancy is due to the over estimation of the favorable electrostatic term. We suspect that also in this case, as for 1,4-dioxane, this is due to the dependence of the molecular dipole moment on ring conformation.

The model reproduces the hydration free energies of the amides with very good accuracy. We found the calculated electrostatic hydration free energy of the amides to be not negative enough to explain the large and negative experimental hydration free energies of the amides. This is reflected by the anomalous negative nonpolar residuals that we have obtained in this case. The nonpolar estimator offsets this deficiency by introducing a large and negative α parameter for the amide nitrogen (see Table 8). The inconsistency between the electrostatic model and the experiments is likely due to out-of-plane hydrogen bonds between the amide nitrogen and water molecules. The magnitude of the amide nitrogen α parameter is indeed consistent with a typical hydrogen bond energy.

The fitting results for nitriles and nitro compounds are generally acceptable. Given the small number of entries in these groups, however, the jackknife results are poor. We found that the electrostatic model systematically overestimates the favorable electrostatic hydration free energy of these compounds resulting in large positive nonpolar residuals. In the case of nitriles, this effect was particularly large, and was partially compensated by increasing the radius of the nitrile nitrogen (type NZ) used in the reaction field calculation.

The fitting and jackknife results for the rigid nitrogen-containing heterocyclic compounds are very good. Contrary to the case of piperazine, the hydration free energy of pyrazine (a six-member aromatic ring containing two nitrogen atoms in 1,4 relative position) and derivatives are correctly reproduced. This further indicates that the poor results for piperazine may be related to the flexibility of the piperazine molecule.

The training set contains few thiols and sulfides that were used to obtain nonpolar parameters for the sulfur atom. The fitting results for these compounds are acceptable, although a larger number of database entries would be required to establish conclusively the reliability of the model for these compounds.

The fitting errors for ionic compounds (see Table 7) are small compared to the larger hydration free energies of these compounds. The fitted and predicted values are generally within 5% of the experiments. The nonpolar residuals for the carboxylate anions are found to be large and positive, indicating that the calculated electrostatic hydration free energies are too negative. This is also observed, but to less extent, for the ammonium cations. This is reflected in the anomalous values of the γ and α parameters obtained for the ionic atom types (see Table 8).

Tables 5 and 7 summarize the fitting results. The overall average unsigned error of the fit for the 199 neutral compounds in

Table 6. Fitting and Prediction Results for the Hydration Free Energies of the Ionic Compounds.

Molecule	Experiment[a]	Fit[a,b]	Prediction[a,b]
		Carboxylate Anions	
Acetate	−79.900	−80.664 (−0.764)	−85.949 (−6.049)
Propionate	−79.100	−77.192 (1.908)	−76.716 (2.384)
Benzoate	−76.000	−77.145 (−1.145)	−79.927 (−3.927)
		Ammonium Cations	
Ammonium	−86.000	−83.295 (2.705)	−80.897 (5.103)
Methylammonium	−71.300	−71.849 (−0.549)	−72.299 (−0.999)
Ethylammonium	−68.400	−70.931 (−2.531)	−71.200 (−2.800)
n-Propylammonium	−66.700	−69.729 (−3.029)	−69.012 (−2.312)
Isopropylammonium	−66.500	−68.058 (−1.558)	−68.177 (−1.677)
n-Butylammonium	−66.200	−69.747 (−3.547)	−70.170 (−3.970)
t-Butylammonium	−63.100	−65.451 (−2.351)	−65.589 (−2.489)
Dimethylammonium	−63.900	−62.835 (1.065)	−62.625 (1.275)
Diethylammonium	−58.900	−59.490 (−0.590)	−59.613 (−0.713)
Di-*n*-propylammonium	−57.700	−59.039 (−1.339)	−59.162 (−1.462)
Pyrrolidinium	−61.600	−61.747 (−0.147)	−61.852 (−0.252)
Piperidinium	−60.000	−60.934 (−0.934)	−61.050 (−1.050)
Trimethylammonium	−56.600	−55.055 (1.545)	−54.796 (1.804)
Triethylammonium	−50.200	−51.671 (−1.471)	−51.900 (−1.700)
1-Methylpyrrolidinium	−54.600	−52.189 (2.411)	−51.947 (2.653)
Tetramethylammonium	−52.300	−48.486 (3.814)	−47.902 (4.398)
Tetraethylammonium	−45.300	−45.153 (0.147)	−45.242 (0.058)
Anilinium	−66.000	−59.364 (6.636)	−58.688 (7.312)
N,N-Dimethylanilinium	−52.000	−52.263 (−0.263)	−52.340 (−0.340)

The predicted results are obtained by the jackknife procedure explained in the text.
[a]In kcal/mol.
[b]Difference from experiment in parenthesis.

Table 5 is 0.32 kcal/mol, the corresponding average unsigned error of the jackknife test is 0.50 kcal/mol. The average unsigned error of the fit and jackknife test for the 22 ionic compounds in Table 7 is 1.84 and 2.49 kcal/mol, respectively.

Other hydration free energy models based on electrostatic/nonpolar decompositions have been proposed.[32,34,35] A comparison between the performance of our model and the others is difficult given the differences in the number of adjustable parameters, in the molecular training sets employed, and the ways of estimating the reliability and transferability of the model. We provide here an overview of these models. Compared to the other models our model generally appears to be simpler and at least as accurate.

The model of Sitkoff et al.[34] estimates the electrostatic component of the hydration free energy by computing the exact solu-

Table 7. Mean Unsigned Errors for the Fitted and Predicted Hydration Free Energies of the Ionic Compounds.

Molecular Class	Number of Molecules	Fit[a]	Prediction[a]
Carboxylate anions	3	1.27	4.12
Ammonium cations	19	1.93	2.23
Total	22	1.84	2.49

[a]In kcal/mol.

Table 8. Values of Fitted γ (cal/mol Å^2) and α (in kcal/mol) Parameters for the Nonpolar Function (13).

Type	Symbol(s)	γ	α
1	C	−26.934	−0.722
2	CM, C═	26.787	−0.863
3	CA	−26.971	0.119
4	CT, CY, CO	0.086	−0.754
5	CZ	8.702	−1.021
6	H, HO, HS	79.818	−1.129
7	HA	48.476	−0.591
8	HC	8.395	0.285
9	N	168.600	−5.229
10	NC	−98.958	1.178
11	NZ	145.337	−18.838
12	NO, ON	16.956	−0.486
13	NT	52.251	−2.200
14	O	10.478	0.636
15	OH	9.361	0.302
16	OS	65.636	−1.097
17	S, SH	36.235	−1.946
18	O2	−350.477	15.005
19	N3, NZ1	−100.907	4.806

tion of the Poisson–Boltzmann equation. The nonpolar component is obtained by the solvent accessible surface area and a surface tension coefficient. The results obtained for a set of molecules mimicking amino acid side chains using OPLS or *ab initio* charges were considered not very accurate. Reparameterization of the charges for a set of about 50 molecules improved the results considerably albeit with the introduction of many additional adjustable parameters. The model was shown to be able to reproduce the hydration free energy of a set of about 70 small molecules not included in the training set with an average unsigned error of 0.44 kcal/mol.

The model of Marten et al.[32] estimates the electrostatic component of the hydration free energy using a self-consistent reaction field method, which couples an electronic structure calculation to a dielectric continuum model. The nonpolar component is estimated from the solvent accessible surface area. The parameters optimized by Marten et al. are the atomic radii that define the solute–solvent interface and solvent exposure correction factors. The model fits the hydration free energies of a set of small molecules similar to ours with an average unsigned error of 0.36 kcal/mol.

The model of Hawkins et al.[35] evaluates the electrostatic polarization component using the generalized Born model implemented using a pairwise descreening approximation. The model incorporates first solvation shell effects using a linear function of the solvent-exposed atomic surface areas. The model features several different versions[35] that differ in their parameterization, fitting functional forms and level of the semiempirical charge distribution calculations. A substantial number of parameters of the model are optimized against experimental hydration free energies: Coulomb atomic radii, van der Waals radii, surface tension coefficients, and others. The fitting to the experimental hydration free energies of roughly 220 small neutral molecules yielded a best average unsigned error of 0.44 kcal/mol.

Conclusions

We have presented a parameterization of the water solvent potential of mean force based on the SGB electrostatic model and a nonpolar hydration free energy estimator. The electrostatic hydration free energy model was parameterized against explicit solvent free energy perturbation calculations.[13] The form of the nonpolar hydration free energy estimator is motivated by the results of explicit solvent simulations of the thermodynamics of hydration of hydrocarbons.[14] It contains, in addition to the more commonly used solvent-accessible surface area term, a component corresponding to the van der Waals solute–solvent interactions that is not zero even for buried atoms up to a certain distance from the solute solvent-exposed surface. This term is found to be important to improve the accuracy of the model, particularly for cyclic and hydrogen bonding compounds. The nonpolar estimator is parameterized by fitting the hydration free energy model to a database of experimental hydration free energies of small organic molecules.

The model reproduces the experimental hydration free energies of small organic molecules with an accuracy comparable or superior to similar models employing more computationally demanding estimators and/or a more extensive set of parameters. The model is designed to be applicable to macromolecules. Future studies will address the optimization of the switching function parameters of the nonpolar estimator (a feature of the model specific to macromolecules) and the application of the model to protein folding and ligand binding.

Acknowledgments

We thank Dr. Tom Halgren for providing the molecule database.

References

1. Lazaridis, T.; Archontis, G.; Karplus, M. Adv Protein Chem 1995, 47, 262.
2. Honig, B.; Yang, A. Adv Protein Chem 1995, 46, 27.
3. Dill, K. A. Biochemistry 1990, 29, 7133.
4. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. Biophysical J 1997, 72, 1047.
5. Apostolakis, J.; Ferrara, P.; Calfish, A. J Chem Phys 1999, 110, 2099.
6. Levy, R. M.; Gallicchio, E. Annu Rev Phys Chem 1998, 49, 531.
7. Roux, B.; Simonson, T. Biophy Chem 1999, 78, 1.
8. Lazaridis, T.; Karplus, M. Curr Opin Struct Biol 2000, 10, 139.
9. Wallqvist, A.; Gallicchio, E.; Felts, A. K.; Levy, R. M. Adv Chem Phys 2001, to appear.
10. Ghosh, A.; Rapp, C. S.; Friesner, R. A. J Phys Chem B 1998, 102, 10983.
11. Zauhar, R. J.; Morgan, R. S. J Mol Biol 1985, 186, 815.
12. Honig, B.; Sharp, K.; Yang, A. J Phys Chem 1993, 97, 1101.
13. Zhang, L.; Gallicchio, E.; Friesner, R.; Levy, R. M. J Comp Chem 2001, 22, 591.
14. Gallicchio, E.; Kubo, M. M.; Levy, R. M. J Phys Chem B 2000, 104, 6271.
15. Pitera, J. W.; van Gunsteren, W. F. J Am Chem Soc 2001, 123, 3163.
16. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrikson, T. J Am Chem Soc 1990, 112, 6127.
17. Nina, M.; Beglov, D.; Roux, B. J Phys Chem B 1997, 101, 5239.
18. Dominy, B. N.; Brooks, C. L. I. J Phys Chem B 1999, 103, 3765.
19. Onufriev, A.; Bashford, D.; Case, D. A. J Phys Chem B 2000, 104, 3712.
20. Jorgensen, W. L.; Maxwell, D. S.; Tirado–Rives, J. J Am Chem Soc 1996, 118, 11225.
21. Zhang, L.; Gallicchio, E.; Levy, R. M. In AIP Conference Proceedings (Simulation and Theory of Electrostatic Interactions in Solutions), Vol. 492; 1999.
22. Ashbaugh, H. S.; Kaler, E. W.; Paulaitis, M. E. J Am Chem Soc 1999, 121, 9243.
23. Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. J Solut Chem 1981, 10, 563.
24. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, C. W. J Phys Chem A 1997, 101, 3005.
25. Sitkoff, D.; Sharp, K. A.; Honig, B. J Phys Chem 1994, 98, 1978.
26. Eisenberg, D.; McLachlan, A. D. Nature 1986, 319, 199.
27. Ooi, T.; Oobatake, M.; Nemethy, G.; Sheraga, A. Proc Natl Acad Sci USA 1987, 84, 3086.
28. Privalov, P. L.; Makhatadze, G. I. J Mol Biol 1993, 232, 660.
29. Kitchen, D. B.; Hirata, F.; Westbrook, J. D.; Levy, R. M.; Kofke, D.; Yarmush, M. J Comput Chem 1990, 11, 1169.
30. Ding, Y.; Bernardo, D. N.; Krogh–Jespersen, K.; Levy, R. M. J Phys Chem 1995, 99, 11575.
31. Kubo, M. F.; Gallicchio, E.; Levy, R. M. J Phys Chem 1997, 101, 10527.
32. 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.
33. Rizzo, R. C.; Jorgensen, W. L. J Am Chem Soc 1999, 121, 4827.
34. Sitkoff, D.; Sharp, K. A.; Honig, B. Biophys Chem 1994, 51, 397.
35. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. J Phys Chem 1996, 100, 19824. CW000104-TI