JCTC Journal of Chemical Theory and Computation

Rapid Prediction of Solvation Free Energy. 2. The First-Shell Hydration (FiSH) Continuum Model

Christopher R. Corbeil, Traian Sulea, and Enrico O. Purisima*

Biotechnology Research Institute, National Research Council Canada, 6100 Royalmount Avenue, Montreal, Quebec H4P 2R2, Canada

Received November 17, 2009

Abstract: Local ordering of water in the first hydration shell around a solute is different from isotropic bulk water. This leads to effects that are captured by explicit solvation models and missed by continuum solvation models which replace the explicit waters with a continuous medium. In this paper, we introduce the First-Shell Hydration (FiSH) model as a first attempt to introduce first-shell effects within a continuum solvation framework. One such effect is charge asymmetry, which is captured by a modified electrostatic term within the FiSH model by introducing a nonlinear correction of atomic Born radii based on the induced surface charge density. A hybrid van der Waals formulation consisting of two continuum zones has been implemented. A shell of water restricted to and uniformly distributed over the solvent-accessible surface (SAS) represents the first solvation shell. A second region starting one solvent diameter away from the SAS is treated as bulk water with a uniform density function. Both the electrostatic and van der Waals terms of the FiSH model have been calibrated against linear interaction energy (LIE) data from molecular dynamics simulations. Extensive testing of the FiSH model was carried out on large hydration data sets including both simple compounds and drug-like molecules. The FiSH model accurately reproduces contributing terms, absolute predictions relative to experimental hydration free energies, and functional class trends of LIE MD simulations. Overall, the implementation of the FiSH model achieves a very acceptable performance and transferability improving over previously developed solvation models, while being complemented by a sound physical foundation.

Introduction

Molecular recognition in biological systems usually takes place in aqueous solution and is accompanied by the desolvation of the interacting surfaces and reorganization of the solvent around the ensuing complex. Hence, theoretical prediction of protein−ligand binding modes (i.e., docking) and binding affinities (i.e., scoring) require an accurate description of the change in hydration that accompanies solute binding.[1] With the advent of faster computers over the past decade, large-scale *in silico* docking-scoring (aka virtual screening) of small-molecule libraries has become appealing due to its speed and cost efficiency.[2] Unquestionably, the success (or failure) of virtual screening relies mostly on the quality of the underlying docking and scoring function(s). The challenge in virtual screening is augmented by the fact that, in order to provide a useful hit-enrichment level, accurate docking-scoring has to be achieved under the constraint of fast computing. To this end, a fast yet accurate solvation model is of paramount importance in the early stages of the drug discovery process.

Over the past years, much research has been dedicated to developing and parametrizing solvation models at various levels of theory.[3−9] Explicit-solvent models of hydration, including rigorous pathway methods such as free energy perturbation (FEP) and thermodynamic integration (TI),[1,10] or approximate end-point methods such as linear interaction energy (LIE),[11−14] address the discrete nature of water around the solute. This treatment results

* To whom correspondence should be addressed. Phone: 514-496-6343. Fax: 514-496-5143. E-mail: enrico.purisima@cnrc-nrc.gc.ca.

10.1021/ct9006037 Published 2010 by the American Chemical Society
Published on Web 04/02/2010

in transferability across a wide chemical space which is dependent on the underlying force-field. However, explicit models require molecular dynamics (MD) or Monte Carlo simulations and are therefore not practical for high-throughput applications. Implicit models of hydration (i.e., continuum models) have been precisely developed to address the speed issue, and they excel in this regard. However, this speed increase associated with continuum models has a cost, an impact on accuracy.[15−18] The current focus in the field of continuum solvation is for the development of models which can capture the underlying physics of solvation, while retaining the speed achieved by the current generation of continuum solvation models.

The local ordering of water in the first hydration shell around a solute is different from isotropic bulk water and varies depending on solute polarity. Around a hydrophobic solute surface, interactions within the first hydration shell itself are favored over interactions with the solute or with bulk solvent.[19] Around polar solute surfaces, water molecules interact strongly with the solute but orient differently around positively and negatively charged atoms, a phenomenon known as the charge asymmetry of water.[20] It is imperative for implicit solvation models to be able to capture the effects of first shell water ordering.

The philosophy adopted in this study is to develop a continuum solvation model that emulates physics-based explicit solvent models. In this way, the transferability should increase in comparison with empirical models that incorporate a large number of parameters fitted directly to experimental data.[21,22] The physical meaning of the tunable descriptors in empirical models is also often times lost. We propose here the First-Shell Hydration (FiSH) model, a continuum solvation model that reformulates the usual continuum electrostatics and van der Waals treatments in order to capture features present in the all important first shell of hydration.[23] The FiSH continuum model is designed to mimic an explicit solvent LIE model of hydration. As in the companion study using explicit solvent LIE simulation,[24] the FiSH model is applied on a large hydration data set encompassing 501 "traditional" compounds[25,26] and 63 neutral drug-like compounds from the more challenging SAMPL1 data set.[27] In the Theory and Implementation, we present improvements to the original continuum electrostatics-dispersion (CED) model of solvation[21] which have led to the development of the FiSH continuum solvation model. In the Results and Discussion section, we assess the main objective of the FiSH model, its ability to reproduce hydration free energies of the explicit-solvent LIE model. This is followed with a comparison to experimental hydration free energies and an assessment of its transferability compared to our previously developed solvation models.

Theory and Implementation

Continuum Electrostatics-Dispersion (CED) Solvation Model. Our previous attempts at formulating a continuum solvation model led to the development of the CED solvation model, which has the following functional form:[21]

$$\Delta G_{hyd}^{CED}(D_{in}, \rho, \gamma_{cav}, \{B_i\}) = \Delta G_{hyd}^{R}(D_{in}, \rho) + \gamma_{cav}\mathrm{MSA} + \sum_i U_i^{cvdW}(B_i) + C \quad (1)$$

where D_{in} is the solute dielectric constant, ρ is the block-scaling factor for the AMBER van der Waals radii, γ_{cav} is the cavity surface coefficient, and $\{B_i\}$ represents the set of atom-type-dependent continuum van der Waals coefficients. These coefficients were trained to fit experimental hydration free energies of a set of 129 neutral solutes. The electrostatic contribution, ΔG_{hyd}^{R}, was calculated using the BRI-BEM program,[28,29] which solves the Poisson equation using a boundary element method. The cavity contribution is proportional to the total molecular surface area, MSA, which was calculated using a variable surface probe.[30,31] The dispersion-repulsion term, U_i^{cvdW}, was calculated by integrating the 6−12 Lennard-Jones potential over the molecular surface[21,32,33] for a set of defined atom types, each with its own van der Waals coefficients trained to fit experimental hydration free energy. This model yielded very good results on a test set of traditional solutes similar to those used for its training. Application to the more challenging drug-like SAMPL1 data set[27] demonstrated limited transferability to more drug-like molecules. These results prompted us to change our strategy and develop a model trained primarily on explicit-water simulations instead of on experimental hydration free energies.

First-Shell Hydration (FiSH) Continuum Model Formulation. As with its CED solvation model predecessor, the FiSH model includes electrostatic, van der Waals and cavity contributions to solvation as formulated in eq 2:

$$\Delta G_{hyd}^{FiSH}(\{r_i^{Born}\}, \gamma_{cav}) = \Delta G_{hyd}^{R}(\{r_i^{Born}\}) + U^{vdw} + \gamma_{cav}\mathrm{MSA} + C \quad (2)$$

The electrostatic contribution is the change in the solute reaction field energy, ΔG_{hyd}^{R}, calculated by solving the Poisson equation in water and in vacuum dielectrics. The nonpolar hydration effects are described by the solute−solvent van der Waals interaction energy, U^{vdw}, and by the cost of cavity formation in water that is proportional to the solute molecular surface area, MSA. Even though the components of the FiSH and CED models are similar, they are obtained in different ways.

FiSH Born Radii. The FiSH continuum electrostatic term uses atomic Born radii $\{r_i^{Born}\}$, derived from general corrections to the van der Waals radii of atoms in a molecule that restore the asymmetric response of water to solutes of different polarities.[34−37] The charge asymmetry phenomenon is dominated by first solvation shell effects.[20] Water molecules orient differently around positively and negatively charged atoms, resulting in changes in the dielectric boundaries (Figure 1). This leads to different effective Born radii and an asymmetric dependence of the reaction field energies on solute charge. Charge asymmetries are captured by explicit water simulations, but the usual continuum electrostatics calculations fail miserably in capturing this phenomenon.[20,38] We have recently presented a proof of concept, in which charge asymmetry of solvation can be handled in

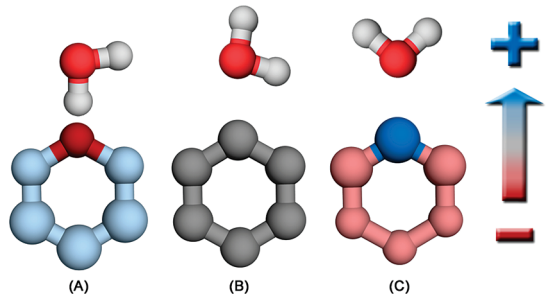

Figure 1. Schematic showing the dependence of the dielectric boundary on orientation of a water molecule around the dominant charge of a neutral hexagonal bracelet model.[20] (A) Bracelet with negative dominant charge (red = −1.0*e* charge, blue = +0.2*e* charge). (B) Uncharged bracelet (gray = 0.0*e* charge). (C) Bracelet with positive dominant charge (blue = +1.0*e* charge, red = −0.2*e* charge).

a simple, systematic, and transferable way within a purely continuum electrostatics framework.[23] In this approach, we used the average induced surface charge density (ISCD), σ_i, obtained from a boundary element solution of the Poisson equation to derive a simple linear correction to the van der Waals radius to obtain the Born radius, r_i^{Born}, for each atom, *i,* of a molecule.[23]

$$r_i^{\text{Born}} = \begin{cases} r_i^0 - c_+\sigma_i \text{ if } \sigma_i \geq 0 \\ r_i^0 - c_-\sigma_i \text{ if } \sigma_i < 0 \end{cases} \quad (3)$$

To obtain the σ_i for eq 3, all atoms are initially assigned Born radii equal to the General AMBER Force Field (GAFF)[39] van der Waals radii r_i^0. From the boundary element solution to the Poisson equation, the average ISCDs for each atom are then calculated by assigning the surface patches and their associated charge density to the nearest atom. Only atoms with solvent exposure have their Born radii modified from the initial van der Waals radii since only these atoms define the molecular surface. However, it should be noted that the correction embodied in eq 3 is based on a molecular property, the ISCD, and not just on a local effect. Thus, even buried atoms, whose Born radii remain unchanged, can affect the Born radii of surface atoms because of their influence on the molecule's ISCD. The two coefficients, c_+ and c_-, used for positive and negative σ_i were previously trained on the electrostatic free energy from FEP simulations for a set of model systems consisting of pairs of neutral hexagonal bracelets with mirrored charge distribution (Figure 1).[20] Tests on pairs of model systems with different geometries indicated the generality of the approach and the transferability of the calibrated coefficients.[23] However, the c_+ and c_- coefficients derived previously were for highly simplified model systems made of a single atom type. Thus, in this work, we retrained the continuum electrostatic coefficients, c_+ and c_-, to the explicit-solvent LIE electrostatic component for the training data set of 200 neutral molecules and obtained slightly different values of 16.222 Å^3/*e* and 11.843 Å^3/*e* for c_+ and c_-, respectively, compared to the previous values of 15.5 Å^3/*e* and 11.5 Å^3/*e*.[23]

The relatively small change in c_+ and c_- coefficients in going from simple model systems to a 200-molecule training set suggests that the coefficients are not overly sensitive to the atom types, at least as far as neutral molecules go. It also suggests that the linear correction in eq 3 may perform relatively well for the normal range of partial charges observed in neutral real molecules. However, that approximation has its limitations. We expect the linear dependence to level off at some point or even reverse in the case of a negative σ_i. At moderately large negative σ_i, the Born radius is larger than the Lennard-Jones radius because this reflects the orientation of the first solvation shell water molecule with the water hydrogen atoms pointing away from the surface. However, as the ISCD becomes even more negative (i.e., the solute electrostatic potential in that region becomes more positive), the water molecule will be drawn closer to the solute molecular surface, and the effective Born radius should decrease. For positive σ_i, increases in the value of σ_i are associated with a decrease in the Born radius as the hydrogen of the water molecule is pulled closer to the solute, effectively decreasing the Born radius. As these decreases become larger, a leveling off should occur since the van der Waals repulsion will start to become significant. Also, we expect the coefficients to depend upon the well depth of the Lennard-Jones potential. These limitations of the linear functional form in eq 3 motivated an exploration of a nonlinear dependence of the Born radii on the ISCD, as discussed below.

Nonlinear Dependence on ISCD. The dependence of the Born radii on the ISCD and van der Waals parameters can be examined using simple spherical solutes of varying partial charges, *q*, Lennard-Jones well depths, ε, and van der Waals radii, r^0. Spherical solutes are ideal to investigate the shape of the nonlinearity with respect to ISCD since effective Born radii can be obtained directly from the Born equation.[40] In Figure 2, we plot the difference between the effective Born radii and the original van der Waals radii of these model spherical solutes versus ISCD. Born radii were obtained by fitting the analytical expression for the reaction field energy of a spherical ion to that calculated with the LIE approach based on MD simulations in explicit water (see Materials and Methods). The results shown in Figure 2 indicate that the nonlinear dependence on the ISCD follows the expected behavior described above. Figure 2A shows the dependence of the radius correction on the Lennard-Jones well depth, ε, at a fixed van der Waals radius. The data points for each well depth define roughly parallel curves. For a given van der Waals radius, the radius correction becomes more negative for smaller well depths, ε, (Figure 2A) across the entire range of σ, which correspond to a series of partial charges from −1 to +1. This behavior is understandable due to the closer approach of water in the case of a "softer" solute sphere (i.e., smaller well depth). Figure 2B shows the dependence of the radius correction on the van der Waals radius at a fixed well depth. The variation among the different curves seems to be more pronounced for negative σ compared to positive ones. The radius correction is more negative for larger r^0, but only marginally so for positive σ. These changes in Born radii are size effects due to geometrical restrictions of accommodating discrete water molecules in the first solvation shell around very small solutes.

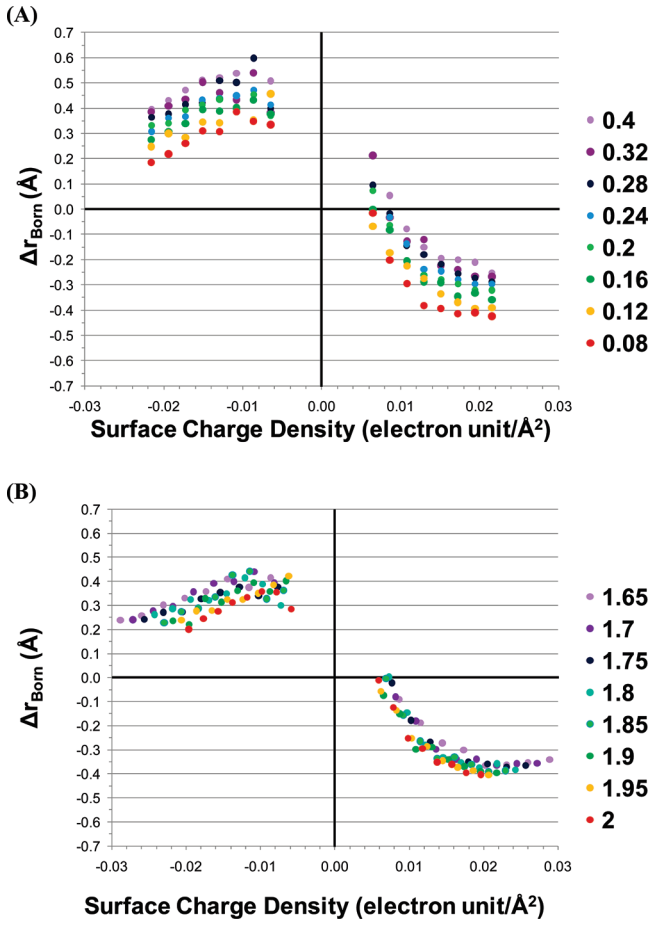

Figure 2. Calculated change in Born radius with induced surface charge density and its dependence on solute van der Waals parameters for singly charged spherical solutes. (A) Effect of the well depth of the 6−12 Lennard-Jones potential, ε (in kcal/mol), on the Born radius. (B) Effect of the equilibrium radius of the 6−12 Lennard-Jones potential, r^0 (in Å), on the Born radius. See the Theory and Implementation section for details.

This leads to a compensatory effect of maintaining a certain Born radius as the van der Waals radius decreases.

The nonlinear dependence on ISCD, the direct dependence on the well depth, and the inverse dependence on van der Waals radii led us to consider a functional form based on combining the arctan and Gaussian functions:

$$r_i^{\text{Born}} = r_i^0 + A \arctan(B\sigma_i + C) + \frac{D}{r_i^0}\exp\left[\frac{\left(\sigma_i - \frac{E}{r_i^0}\right)^2}{2\left(\frac{F}{r_i^0}\right)^2}\right] + G\varepsilon_i + \frac{H}{r_i^0} \quad (4)$$

where A, B, C, D, E, F, G, and H are fitted parameters. This allows us to easily relate the shape of the correction function to the underlying physical interactions. The arctan dependence of the Born radii on the ISCD relates to the water hydrogen orientation around a positively or negatively charged solute atom assuming the water oxygen atom is at a fixed contact distance to the solute atom. The shifted arctan is a more sophisticated version of the linear functions in eq 3. The Gaussian dependence of the Born radii on the ISCD emulates the attraction of the entire water molecule as the partial charge of the solute atom increases (irrespective of sign) and the limitation of the solute−solvent approach due to the Lennard-Jones repulsion. We noticed that the arctan component is fairly constant on the negative ISCD, allowing the nonlinear correction to be mostly controlled by the Gaussian component. The reverse is true for positive ISCD. These features should enable this function to capture all the aspects seen in Figure 2, including the hump at small negative σ. In terms of the dependence on Lennard-Jones parameters of the solute, the arctan−Gaussian function shifts the Born radius up with increasing well depth (ε) and decreasing size (r^0), arising mainly from the last two terms in eq 4. The inverse dependence on size is also included in the Gaussian component, critical at negative ISCD.

Even though the form of the arctan and Gaussian function allows for an interpretation in terms of the underlying physics, there is a danger of overfitting due to the large number of parameters. Hence, an alternate simpler functional form, a rational function, was also explored:

$$r_i^{\text{Born}} = r_i^0 + \frac{A\sigma_i + B\left(\frac{\varepsilon_i}{r_i^0}\right)}{D\sigma_i^2 + E\sigma_i + 1} \quad (5)$$

where A, B, D, and E are fitted parameters. Regarding the Born radius dependence on Lennard-Jones parameters, this rational function essentially shifts the Born radius correction up and down with the solute softness and size, respectively. The advantages of the rational function are the good quality of the fit with a lower number of fitting parameters compared to eq 4.

We note that both nonlinear correction functions report a Born radius for an uncharged spherical solute that is larger than its van der Waals radius, r_i^0. There is no physical reason why these radii should be identical in the case of uncharged solutes. In fact, for a convex solute, a slightly larger Born radius than the van der Waals radius may be expected because the hydrogen density from the first shell of waters would be located at a slightly greater distance from the surface than the oxygen density. This effect is especially pronounced for small spherical solutes, but still present around an uncharged protein using long MD simulations in SPC/E water[41] where hydrogen densities were 0.1 Å further away from the protein than peak oxygen densities for the first solvation shell. Also, interpolation of our data on varying the size of spherical solutes (Figure 2B) suggests slightly larger increases of Born radii for the uncharged spheres with smaller van der Waals radii (i.e., more convex).

The purpose of the calculations on the spherical model solutes was simply to guide the selection of the nonlinear functional form to use. All parameters in the two nonlinear functions were later retrained on real molecules from the training set against the electrostatic component of solute−solvent interaction energy from explicit water MD simulations using the LIE approximation.

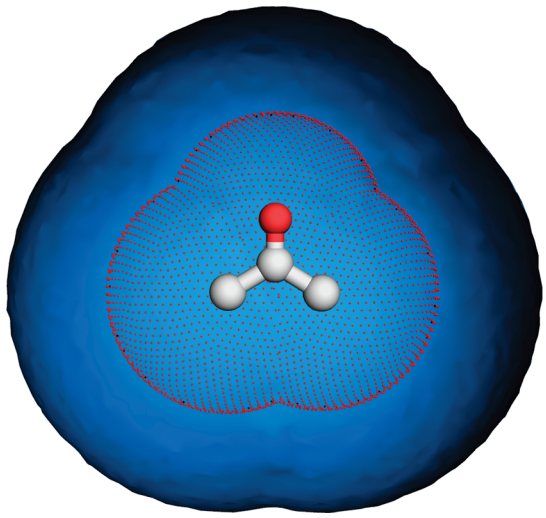

Figure 3. Illustration of the two regions defined for the hybrid van der Waals component of the FiSH continuum model, using acetone as an example. Red dots represent the first shell of water oxygen atoms uniformly dispersed over the solvent-accessible surface (SAS). The blue surface represents where the outer region of continuum solvent starts (SAS + 2.8 Å).

Continuum van der Waals Model. In the usual continuum van der Waals model, the solute−solvent van der Waals interactions are obtained from the integral of the solute-continuum interactions over all of space, with the volume integral transformed into a surface integral at the solute−solvent boundary represented by solvent-accessible surface or molecular surface.[21,32,33] In this approach, a uniform density function for the solvent is assumed. Clearly, this is a gross approximation for the first hydration shell. Partly due to this, scaling coefficients are typically required to adjust the continuum van der Waals energy to the magnitude of explicit solute−solvent interactions or experimental data.[42] Here, in order to avoid empirical scaling of atom-type-based van der Waals parameters as in the previous CED model and to mimic more closely the important first solvation shell interactions from explicit solvent simulations, we devised a two-region solvent model for the calculation of Lennard-Jones interactions with the solute (Figure 3).

$$U^{\mathrm{vdW}} = U_1^{\mathrm{vdW}} + U_2^{\mathrm{vdW}} \tag{6}$$

The basic idea is that, for the first shell, which is represented by the solvent-accessible surface (SAS), we assume that the water oxygen is completely restricted to lie on the SAS but is uniformly distributed along the surface. We take the second and succeeding shells to start 2.8 Å (a water diameter) away from the SAS and to be uniformly distributed from that point out to infinity. The contribution of the first shell, U_1^{vdw}, is then calculated as a discretized integral between the solute atoms i and the surface distribution of the TIP3P water oxygen atoms.

$$U_1^{\mathrm{vdW}} = \rho_S \sum_i^{\mathrm{atoms}} \sum_j^{\mathrm{patches}} \left(\frac{A_{iw}}{r_{ij}^{12}} - \frac{B_{iw}}{r_{ij}^{6}} \right) \mathrm{SA}_j \tag{7}$$

ρ_S is the number density of water along the surface. The A_{iw} and B_{iw} coefficients are from TIP3P and the general AMBER force field (GAFF).[39] SA_j is the area of the triangulated surface patch j.

The second region in our continuum van der Waals model is constructed by extending the SAS by 2.8 Å, i.e., one water diameter (Figure 3). The solvent density is assumed to be approximately uniform from this point onward, allowing the dispersion (attractive) component to be computed as a discretized surface integral in the usual way:

$$U_2^{\mathrm{vdW}} = -\rho_N \sum_i^{\mathrm{atoms}} \sum_j^{\mathrm{patches}} \frac{1}{3} \frac{B_{iw}}{r_{ij}^{6}} \mathbf{r}_{ij} \cdot \mathbf{n}_j SA_j \tag{8}$$

where $\mathbf{r}_{ij}$ is the vector from solute atom i to boundary surface patch j, $\mathbf{n}_j$ is the unit surface normal at j, SA_j is the area of patch j, and ρ_N is the solvent number density of bulk water. The atomic dispersion parameters B_{iw} are taken from the GAFF force field and TIP3P. Ignoring the repulsive r^{-12} contribution saves some computation time without introducing much error. It should be noted that no scaling or fitting of the U_2^{vdw} term is carried out.

A key component of this approach is that the SAS is constructed using solute atom-specific solvent probe radii. The starting point for defining the atom-specific solvent probe radii is the location of the first peak for various atom types in the radial distribution function (RDF) determined from MD simulations in explicit water for the training set. Specifically, we determine average first RDF peak distances between the water oxygen and the GAFF atom types and use that to define the atom-specific solvent probe radii. The SAS is then generated by operationally inflating the force field van der Waals atomic radii by the appropriate probe radii (Table 1). Additional manual fine-tuning of the radii is carried out in order to improve the agreement between U_1^{vdw} and the average solute−solvent van der Waals interaction energy with an effective first hydration shell defined as all water molecules with oxygen centers not farther than the SAS + 2.8 Å (a water diameter), calculated with the LIE approach based on MD simulations in explicit water.

Materials and Methods

Hydration Data Sets. A data set consisting of experimental hydration free energies for 501 neutral organic small molecules compiled from the published literature[25] was used as prepared in a previous study.[24] As in the previous study, the conformations used for implicit solvation predictions correspond to the conformation with the best potential energy in a vacuum, which have been shown to reproduce well hydration free energies in explicit-solvent models.[21,26] This data set was split into a training set of 200 compounds and a testing data set of 301 compounds. The training data set was used for calibrating the cost of cavity formation in water against experimental hydration free energy data, and for calibrating the electrostatic and van der Waals components of the FiSH continuum model against calculated explicit-solvent LIE data. In the training set, we included mostly rigid representatives of the various chemical classes, with the majority of compounds being monofunctional, and only a

Table 1. First RDF Peak for Various GAFF Atom Types[a]

atom type	RDF peak	r^o	solvent probe radius	atom type	RDF peak	r^o	solvent probe radius
c	3.20	1.908	1.292	f	3.15	1.750	1.400
c1	3.25	1.908	1.342	cl	3.40	1.948	1.452
c2	3.20	1.908	1.292	br	3.75	2.220	1.530
c3	3.30	1.908	1.392	i	3.85	2.350	1.500
ca	3.25	1.908	1.342	n	3.15	1.824	1.326
cp	3.25	1.908	1.342	n1	3.25	1.824	1.426
cq	3.25	1.908	1.342	n2	2.80	1.824	0.976
cc	3.25	1.908	1.342	n3	3.03	1.824	1.206
cd	3.25	1.908	1.342	na	2.95	1.824	1.126
ce	3.25	1.908	1.342	nb	2.95	1.824	1.126
cf	3.25	1.908	1.342	nc	2.95	1.824	1.126
cg	3.55	1.908	1.642	nd	2.95	1.824	1.126
ch	3.55	1.908	1.642	ne	2.80	1.824	0.976
cx	3.30	1.908	1.392	nf	2.80	1.824	0.976
cy	3.30	1.908	1.392	nh	3.10	1.824	1.276
cu	3.60	1.908	1.692	no	3.95	1.824	2.126
cv	3.60	1.908	1.692	o	2.92	1.661	1.259
h1	2.75	1.387	1.363	oh	2.92	1.721	1.199
h2	1.00	1.287	−0.287	os	2.95	1.684	1.266
h3	1.00	1.187	−0.187	ow	2.75	1.768	0.982
h4	2.75	1.409	1.341	p5	3.20	2.100	1.100
h5	2.75	1.359	1.391	s	3.65	2.000	1.650
ha	2.82	1.459	1.361	s4	3.30	2.000	1.300
hc	2.82	1.487	1.333	s6	3.30	2.000	1.300
hn	1.80	0.600	1.200	sh	3.40	2.000	1.400
ho	1.00	0.000	0.000	ss	3.40	2.000	1.400
hs	1.00	0.600	0.400	sx	3.65	2.000	1.650
hx	1.00	0.000	1.000	sy	4.05	2.000	2.050

[a] All values are in Ångstroms. The SAS is constructed using r^o + solvent probe radius. For hydrogens with small or even negative solvent probe radii, this simply means that the SAS is entirely determined by the heavy atom to which it is connected and the RDF value is a dummy one.

few polyfunctional compounds were included to increase coverage of some functional groups. The testing data set mirrors the training data set in terms of chemical class representation for monofunctional compounds but differs from the training analogs by having increased flexibility and containing a larger collection of polyfunctional compounds. We also used the more challenging SAMPL1 data set[27] consisting of 63 drug-like, diverse, polyfunctional, neutral polar compounds, which spans wider ranges of transfer free energies and molecular weights in comparison to the training and testing data sets. The SAMPL1 data set was also used as prepared in our previous study.[24] Details on the composition of the training and testing and SAMPL1 data sets are provided as Supporting Information (Table S1). A functional group analysis was carried out using the testing set. We used the definitions of groups used in the previous companion paper.[24]

LIE Data and MD Simulations. The LIE data for the 564 compounds in the training, testing, and SAMPL1 data sets were taken from the companion study,[24] in which the following implementation of the LIE approximation was used:

$$\Delta G_{\mathrm{hyd}}^{\mathrm{LIE}} = \underbrace{\alpha(\langle E_{\mathrm{S-W}}^{\mathrm{Coul}}\rangle_{\le 12\text{Å}} + \langle G_{\mathrm{S}}^{\mathrm{RF}}\rangle_{12\text{Å}-\infty})}_{electrostatic} + \underbrace{\beta(\langle E_{\mathrm{S-W}}^{\mathrm{vdW}}\rangle_{\le 12\text{Å}} + \langle E_{\mathrm{S}}^{\mathrm{cvdW}}\rangle_{12\text{Å}-\infty})}_{van\ der\ Waals} + \underbrace{\gamma_{\mathrm{cav}}\langle MSA\rangle + C}_{cavity} \quad (9)$$

From the various LIE models derived and described in the companion paper,[24] for this study, we have taken LIE data generated for rigid solute geometries at various partial charge models (primarily AM1BCC-SP, but also AM1BCC-OPT and RESP), with continuum corrections beyond the explicit water shell as described.[24] These data were considered most suitable for the calibration of a continuum solvation model described in this study. LIE simulations were favored over FEP-like methods for training due to their simpler decomposition of the electrostatic and van der Waals component along with a slightly improved accuracy on the testing and SAMPL1 data set.[24] It can be argued that the electrostatic component in the FEP method is possibly contaminated with a net positive change in the solute−solvent van der Waals interaction energy upon solute charging that is embedded into the FEP "electrostatic" component,[24,43,44] although this interpretation is a matter of some debate.[45] The LIE data based on rigid-solute geometries were selected for FiSH continuum model training since the rigid paradigm is often used by implicit solvation models. Hydration free energy predictions can be greatly affected by the choice of solute conformation and the degree of flexibility of the investigated molecules. Therefore, in principle, rigid-solute simulation data should streamline the training and the comparison of an implicit solvation model against a more rigorous explicit-water model.

We also generated LIE data for spherical model solutes that were used to elucidate the nonlinear dependence of atomic Born radii on the ISCD. Spherical model systems were created by varying their van der Waals radius, r^0, from 1.65 Å up to 2.00 Å with 0.05 Å increments while keeping the Lennard-Jones potential well depth, ε, at 0.1094 kcal/

mol (corresponding to the GAFF atom type c3), and by varying ε from 0.08 kcal/mol up to 0.32 kcal/mol with 0.04 kcal/mol increments and including 0.40 kcal/mol while keeping r^0 at 1.908 Å (for GAFF atom type c3). The spherical model solutes had a single atom-centered charge which systematically varied between $-1e$ and $+1e$ with $0.1e$ increments.

MD simulations were carried out with the AMBER 9 software[46] applying the systematically varied parameters described above for the spherical solute. The spherical model systems were solvated in an octahedron of TIP3P water[47,48] extending 13 Å around the solute. The system was energy-minimized first, followed by heating from 100 K to 300 K over 25 ps in the canonical ensemble (NVT), and equilibrating to adjust the solvent density under 1 atm of pressure over 25 ps in the isothermal−isobaric ensemble (NPT) simulation. A 1 ns production NPT run was obtained with snapshots collected every 10 ps, using a 2 fs time-step and 9 Å nonbonded cutoff. The Particle Mesh Ewald (PME) method[49] was used to treat long-range electrostatic interactions, and bond lengths involving bonds to hydrogen atoms were constrained by SHAKE.[50]

Continuum Electrostatic Calculations. Reaction field energies were calculated for a single conformer of each solute molecule using the BRI BEM program, which solves the Poisson equation using a boundary element method.[28,29] The solute and solvent dielectric constants were taken to be 1.0 and 78.5, respectively. The dielectric boundary was taken to be the solvent-excluded surface (also known as the molecular surface) as generated and triangulated using a marching tetrahedra algorithm and a solvent probe radius of 1.4 Å.[30,31] The induced surface charge density (ISCD) distribution at the dielectric boundary was automatically obtained as part of the solution of the Poisson equation. The atom-based ISCD was determined by assigning surface patches to the nearest atom and averaging the ISCDs of the patches associated with a particular atom. All calculations of the ISCD-based Born radii (eqs 3−5) used GAFF[39] van der Waals radii as the initial value, r^0.

Parameter Fitting. Optimization of parameters in the linear and nonlinear correction functions (eqs 3−5) was carried out in order to minimize the mean unsigned error (MUE) of the electrostatic component of solvation calculated with a continuum model from that calculated with an explicit-solvent LIE model, for a given set of compounds.

In the case of spherical model solutes, parameter optimization for the nonlinear models (eqs 4 and 5) was carried out with the Solver plug-in in Microsoft Excel. These values were then used as starting points for parameter optimization against the training data set of real molecules, for which the Nelder−Meade (aka downhill simplex) algorithm using the TCL8.4 math::optimize library was employed. Optimized parameters in eqs 3−5 are given as Supporting Information (Table S2). Bootstrapped statistical analyses were carried out for 5000 samples using the boot library within the R software.[51] In the case of the linear function in eq 3, initial values for the c_+ and c_- parameters were taken from our previous fitting to hexagonal neutral bracelets as model compounds.[23]

Other Continuum Solvation Models. The transferability of the FiSH model will be assessed by comparison to previously developed continuum solvation models, a continuum electrostatics-dispersion (CED) model and a continuum model consisting of only reaction field electrostatics (RF), both of which have been developed and used previously.[21] The CED model consists of continuum reaction field electrostatics, continuum solute−solvent van der Waals interactions, and surface-area-based cavity cost. Unlike in FiSH, where the parameters were trained on explicit water simulation, the parameters were calibrated against the experimental hydration free energy data.[21] The CED model uses a solute dielectric of 1, Born radii that were 0.9 of the AMBER van der Waals radii, and a continuum van der Waals model with 25 atom types, all of which were taken from the previous study. Since the CED parametrization lacked continuum van der Waals parameters for the iodine atom, CED predictions were not obtained for all molecules from the current data sets containing iodine. In the RF model, the scanning of the scalar for the AMBER van der Waals radii, used as the Born radii, and solute dielectric parameters in a previous study suggest optimal values of 1.1 and 1, respectively, for acceptable and transferable prediction of small-molecule hydration.[21] These values are then used for reaction field calculations by solving the Poisson equation using a boundary element method. Both models were used as implemented within the BRI-BEM program.

Results and Discussion

We will begin by presenting the data obtained for new formulations of the electrostatic and van der Waals components of the FiSH model, which were calibrated and tested against explicit-water LIE data. We will then analyze the cavity cost and the total nonpolar contribution to solvation vis-à-vis the solute surface area. The performance of the generated FiSH models versus LIE explicit models, experimental data, and earlier continuum models will be presented in detail. Functional group analysis of errors will be used to detect trends in the FiSH model predictions.

Electrostatic Component of the FiSH Model. As seen in Table 2, even the parametrization of charge-asymmetry-corrected continuum electrostatics on the spherical model alone improved significantly the agreement with the explicit-solvent electrostatic solvation data across the three molecule sets—training, testing, and SAMPL1 relative to using the GAFF radii. For example, in the linear model, the MUEs go from 1.676, 1.557, and 3.506 kcal/mol to 0.513, 0.479, and 0.731 kcal/mol, respectively, for the three sets. Similar improvements can also be seen for the slope and squared correlation coefficient (R^2).

Parametrization on real molecules further improves significantly the agreement between the explicit and continuum models of electrostatic hydration with nonlinear correction functions and marginally with the linear correction function. Results on the testing and SAMPL1 data sets indicate similar performances for the three correction functions: MUE values below 0.5 kcal/mol on the testing data set and slightly above 0.7 kcal/mol for the SAMPL1 data set, in all cases highly

Table 2. Comparing the Electrostatic Component of FiSH Models with the Electrostatic Component of the LIE Explicit-Solvent Model[a]

set	r^{Born} correction	trained on spherical solutes MUE	slope	R^2	trained on molecules MUE	slope	R^2
training	original[b]	1.676 ± 0.090	1.348 ± 0.038	0.889 ± 0.016			
	linear[c]	0.513 ± 0.033	0.980 ± 0.015	0.957 ± 0.007	0.506 ± 0.033	0.982 ± 0.014	0.958 ± 0.007
	Atan+Gauss[d]	1.196 ± 0.078	1.149 ± 0.028	0.900 ± 0.013	0.480 ± 0.031	0.978 ± 0.015	0.962 ± 0.006
	rational [e]	0.859 ± 0.065	1.153 ± 0.020	0.935 ± 0.010	0.531 ± 0.035	0.977 ± 0.008	0.953 ± 0.008
testing	original	1.557 ± 0.071	1.434 ± 0.029	0.918 ± 0.011			
	linear	0.479 ± 0.023	1.022 ± 0.012	0.964 ± 0.004	0.468 ± 0.022	1.018 ± 0.012	0.965 ± 0.004
	Atan+Gauss	1.043 ± 0.060	1.211 ± 0.021	0.920 ± 0.009	0.414 ± 0.021	1.001 ± 0.014	0.970 ± 0.003
	rational	0.758 ± 0.046	1.175 ± 0.015	0.953 ± 0.006	0.444 ± 0.024	0.988 ± 0.014	0.963 ± 0.005
SAMPL1	original	3.506 ± 0.319	1.490 ± 0.025	0.967 ± 0.011			
	linear	0.731 ± 0.077	1.029 ± 0.020	0.982 ± 0.006	0.704 ± 0.072	1.027 ± 0.019	0.983 ± 0.005
	Atan+Gauss	2.732 ± 0.222	1.315 ± 0.027	0.962 ± 0.011	0.732 ± 0.074	1.055 ± 0.014	0.985 ± 0.004
	rational	2.219 ± 0.237	1.281 ± 0.024	0.974 ± 0.008	0.734 ± 0.077	1.005 ± 0.019	0.981 ± 0.006

[a] Statistics are given as averages ± standard deviation for 5000 bootstrapped samples. Errors are in kcal mol^{-1} units. [b] GAFF vdW radii. [c] Equation 3. [d] Equation 4. [e] Equation 5.

correlative and with slopes very close to unity. By comparison, the original Born-radius uncorrected continuum electrostatic model differed from the LIE explicit-solvent electrostatic model by MUEs larger than 1.5 kcal/mol for the training and testing data set and 3.5 kcal/mol for the SAMPL1 data set. Throughout the rest of the paper, all results discussed or presented will be with the parameters that have been trained on molecules.

These results highlight the benefits of ISCD-dependent Born radii to account for charge asymmetry effects, as well as the improvements afforded by training on real molecules for the nonlinear model. The linear correction function appears to provide robust and competitive results when compared with the nonlinear correction functions. However, as presented in the Theory and Implementation section, our computational experiments on spheres clearly show that correction should be nonlinear with respect to the induced surface charge density. The linear function most likely gives good results since the ISCD range explored by the neutral molecules in our data sets is rather narrow (between −0.01 and +0.01 e/Å^2, see Figure S2, Supporting Information), for which the linear approximation is still applicable (see Figure 2). With charged molecules, the range of ISCD will be expanded and the linear correction will most likely fail. For example, the nitrogen of a terminal alkyl ammonium would have an ISCD of around −0.025 e/Å^2, which falls outside of the linear region seen in Figure 2 and justifies the use of a nonlinear function. A more complete study (outside the scope of this work) is needed for charged molecules, but for now, the nonlinear model seems most appropriate because of its greater generality. Given the comparable performances obtained with the two nonlinear correction functions, the rational function (eq 5) is preferred over the arctan + Gaussian function (eq 4) due to a lower number of fitted parameters and will be featured for the rest of the paper. The correlation between the FiSH continuum electrostatic component and the explicit-solvent LIE electrostatic term for the training, testing, and SAMPL1 data sets is shown in Figure 4.

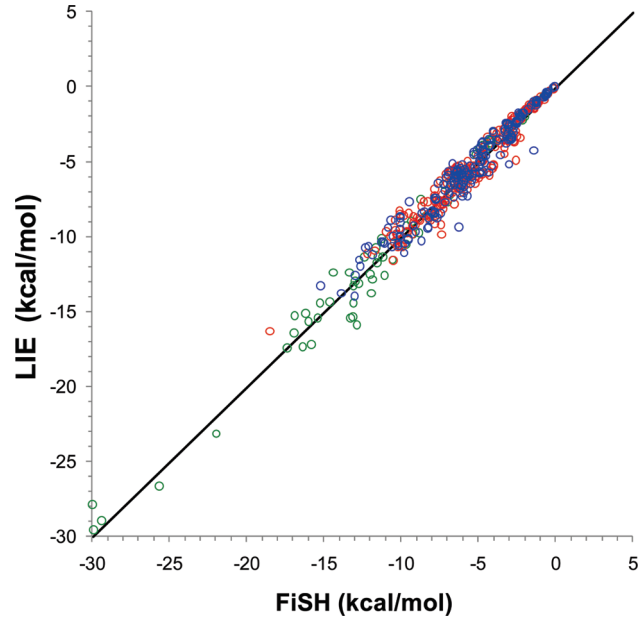

Figure 4. Comparison between the electrostatic component of the FiSH model and the electrostatic component of the LIE explicit-solvent model on the training (blue symbols), testing (red symbols), and SAMPL1 data sets (green symbols). The FiSH continuum electrostatic component is calculated with the optimized rational function (eq 5) for the Born radii. LIE data on single-conformation solutes are taken from a separate study.[24] Both models are derived using AM1BCC-SP partial charges. The diagonal line indicates ideal correlation.

van der Waals Component of the FiSH Model. Comparison with LIE data from explicit-solvent MD simulations with AM1BCC-SP solute charges and rigid solute geometries indicates that our two-zone continuum van der Waals model reproduces the explicit-solvent van der Waals contribution to solvation with MUEs below 0.6 kcal/mol for the testing set and about 1.4 kcal/mol for the SAMPL1 data set (Table 3). The model slightly overestimates the explicit-solvent van der Waals interactions, especially for the SAMPL1 molecules (Figure 5). This may reflect some additivity problems in the continuum model for highly polyfunctional molecules. The FiSH model addresses some of the nonhomogeneity of the solvent distribution function in directions radially away from the solute surface. However, it still assumes a uniform distribution tangential to the solute surface in the first shell.

Table 3. Comparing the van der Waals Component of the FiSH Model against the van der Waals Component of the LIE Explicit-Solvent Model

set	MUE[a]	slope	R^2
training	0.519 ± 0.034	0.900 ± 0.012	0.946 ± 0.004
testing	0.584 ± 0.025	0.882 ± 0.010	0.966 ± 0.004
SAMPL1	1.403 ± 0.110	0.884 ± 0.020	0.925 ± 0.110

[a] Errors are in kcal mol^{-1}.

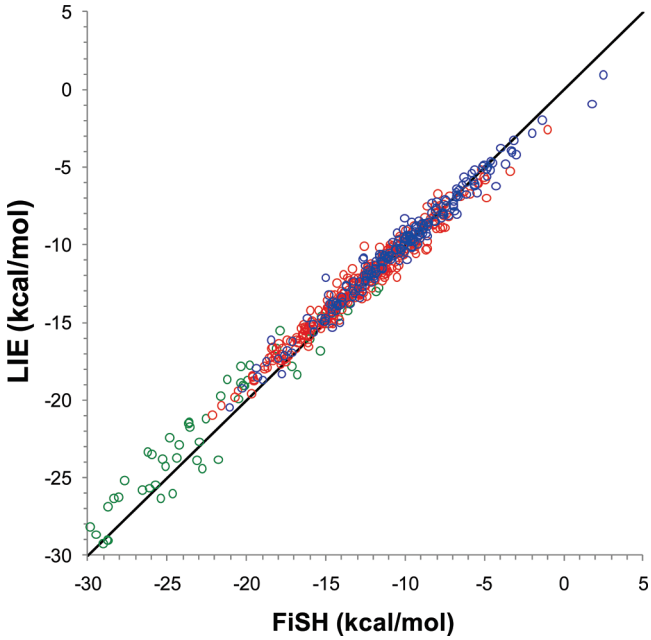

Figure 5. Comparison between the van der Waals component of the FiSH model and the van der Waals component of the LIE explicit-solvent model on the training (blue symbols), testing (red symbols), and SAMPL1 data sets (green symbols). LIE data on single-conformation solutes and AM1BCC-SP partial charges are taken from a separate study.[24] The diagonal line indicates ideal correlation.

This can be a gross approximation with ordered water molecules that may be present in highly polyfunctional molecules.

Cavity and Total Nonpolar Contributions of the FiSH Model. The cost of cavity formation in water was treated as a linear dependence on the molecular surface area, MSA, of the solute (eq 2) and fitted to a pseudoexperimental cavity cost for the training data set. This cost was obtained by subtracting the FiSH continuum electrostatic and van der Waals contributions, described earlier, from the experimental hydration free energy. A robust linear relationship was obtained (Figure 6), characterized by a bootstrapped correlation coefficient of 0.906 ± 0.016 and a slope and intercept (γ and C, respectively, in eq 2) of 0.115 ± 0.003 kcal mol^{-1} Å^{-2} and −4.276 ± 0.386 kcal/mol, respectively (Table 4). We note that the microscopic surface tension of water, γ, is close to the macroscopic one, reiterating the findings obtained with the LIE explicit-solvent model.[24] This linear relationship extends very well to the testing set and SAMPL1 data set (Figure 6), which is further supported by similar cavity parameters γ and C derived by fitting directly to these data sets. The slightly larger slope (γ) obtained in the case of the SAMPL1 data set is partially due to a few sulfoneurea

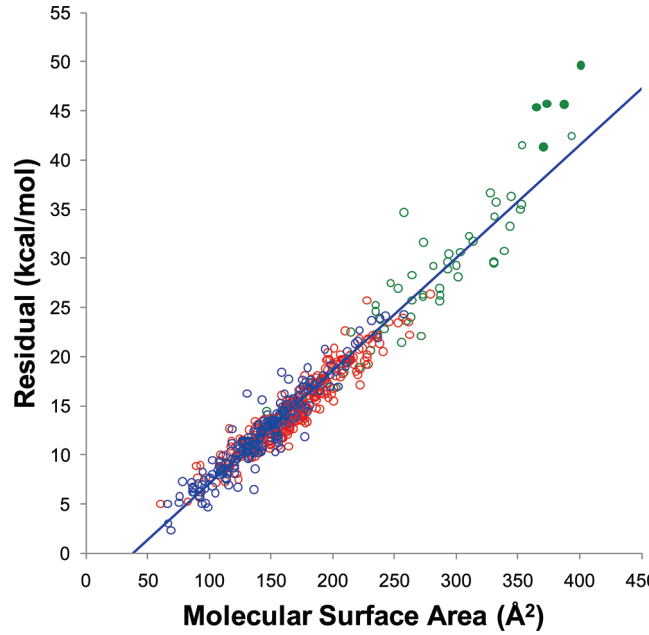

Figure 6. Deriving the cavity contribution for the FiSH model. Linear relationship between pseudoexperimental (residual) cavity contribution and the MSA for the training (blue symbols) and testing (red symbols) data sets and for the SAMPL1 (green symbols) data set. Only the regression line for the training data set is shown, since this is used to predict the cavity contribution for the testing and SAMPL1 data sets. Filled circle points correspond to 5 sulfoneurea analogs from the SAMPL1 data set.

analogs with larger MSA values. Overall, the data obtained for the cavity component of the FiSH continuum model mirror closely those obtained in a study of LIE models of hydration.[24] A direct comparison between the cavity parameters with the FiSH continuum model and the corresponding LIE explicit model is also given in Table 4, where the presented LIE data were derived using AM1BCC-SP partial charges and single-conformation geometries for the solutes. We see that the macroscopic surface tension is consistently slightly larger for the FiSH continuum model relative to the LIE explicit model. This compensates for the modest underestimation of the explicit solute−solvent van der Waals interactions by the FiSH continuum model (Figure 5, Table 3). Intercepts are also consistently more negative in the case of the FiSH continuum model relative to the LIE explicit model.

The total nonpolar solvation component, i.e., the van der Waals contribution plus cavity cost, does not correlate with the solute MSA, due to strong anticorrelation between these contributions leading to cancellation of large opposing numbers (Figure S3, Supporting Information). A weak correlation is seen only in the case of the SAMPL1 data set (Figure S3B). These results mirror LIE data from a previous study demonstrating the FiSH continuum solvation model's ability to mimic an explicit-solvent model.[24] Together with earlier reports from FEP calculations,[25,52] these results stress the requirement for separate accounting of van der Waals and cavity terms.

The FiSH model draws its roots from a continuum electrostatics-dispersion (CED) solvation model,[21] which we

Table 4. Parameters for the Cavity Cost That Can Be Derived from Linear Relationships between the Pseudo-Experimental (Residual) Cavity *versus* the Solute Molecular Surface Area, for the Indicated Hydration Data Sets[a]

	FiSH			LIE		
set	slope (γ)	intercept (C)	R^2	slope (γ)	intercept (C)	R^2
training	0.115 ± 0.003	−4.276 ± 0.386	0.906 ± 0.016	0.108 ± 0.002	−3.488 ± 0.282	0.923 ± 0.015
testing	0.103 ± 0.002	−2.739 ± 0.346	0.903 ± 0.011	0.095 ± 0.002	−1.674 ± 0.291	0.913 ± 0.009
SAMPL1	0.127 ± 0.006	−7.204 ± 1.378	0.902 ± 0.025	0.118 ± 0.005	−5.625 ± 1.231	0.904 ± 0.026

[a] γ is in kcal mol^{-1} Å^{-2} and C is in kcal mol^{-1} units.

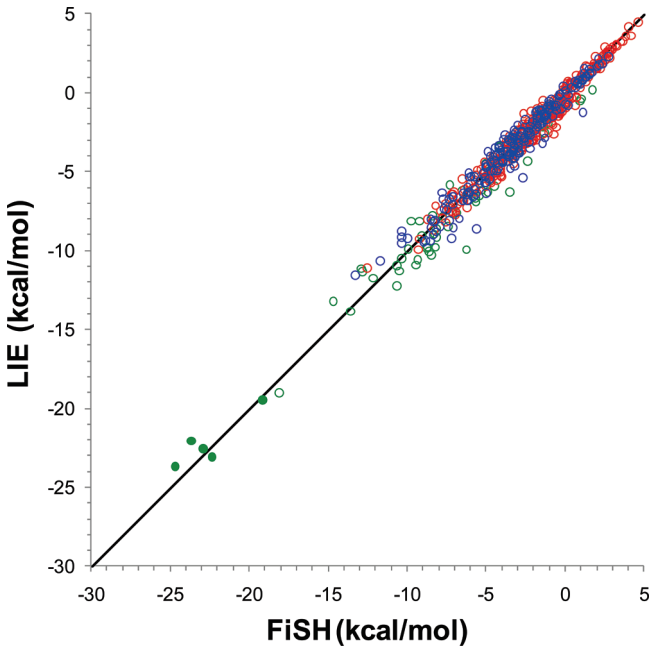

Figure 7. Comparison between hydration free energy predictions with the FiSH model (this study) and with the LIE explicit-solvent model[24] for the training (blue symbols) and testing (red symbols) data sets and for the SAMPL1 (green symbols) data set. Filled circles correspond to 5 sulfoneurea analogs from the SAMPL1 data set. The plotted data correspond to the FiSH model with AM1BCC-SP partial charges, and cavity parameters derived from the training data set. The LIE data are for AM1BCC-SP partial charges and single-conformation solutes and are taken from a separate study.[24] The diagonal line indicates ideal correlation.

have previously employed in the SAMPL1 prospective challenge.[27] An important aspect that differentiates the FiSH model from that earlier model is the reduction of parameters fitted to experimental hydration data in an attempt to improve model transferability. In the FiSH model, only the two cavity parameters require fitting to the experiment, the microscopic surface tension of water, γ_{cav}, and a constant, C. The van der Waals and electrostatic components were calibrated against the corresponding components derived from explicit-solvent simulations using the linear interaction energy (LIE) approach.[11−14,24] The philosophy adopted here is that the transferability of continuum solvation models can be increased by emulating the physics captured by explicit solvation models.

Performance of FiSH Model versus LIE Explicit Model. The primary objective of this study is to develop a continuum model that mimics as closely as possible an explicit solvation model. Performance testing was carried out on the 301 compounds of the testing set and 63 compounds from the SAMPL1 data set. In Figure 7, we plot the hydration free energies predicted with the FiSH continuum model versus those calculated with the LIE explicit-solvent model (based on AM1BCC-SP partial charges and single-conformation representations of the solutes). It is apparent that the continuum model developed here reproduces closely the explicit model al the level of hydration free energies. In quantitative terms, the FiSH continuum model predicts the explicit model data with MUE values of ~0.5 kcal/mol and slightly below 1 kcal/mol for all three data sets, respectively, with correlation slopes and coefficients close to unity (Table 5). There are no major outliers even for the SAMPL1 data set that include more complex, drug-like compounds (Figure 7). We have seen in the previous sections that the excellent agreement carries on to the hydration component terms as well.

Performance of FiSH Continuum Model versus Experimental Data. The absolute performance of the developed FiSH continuum solvation model is tested against the experimental hydration free energies for the testing set and the SAMPL1 drug-like data set. As seen in Figure 8, the FiSH continuum model predictions achieve a fairly good correlation with the experiment. In the case of the testing set, MUE is close to 1 kcal/mol, with a slope close to unity and R^2 above 0.8 (Table 5). We note that the MUE obtained with the FISH model is only 0.1 kcal/mol higher than that obtained with the corresponding LIE model (0.906 kcal/mol).[24] For testing on SAMPL1, MUE is slightly larger than 2 kcal/mol, with a slope and R^2 around 0.6 and of 0.8, respectively. These results are slightly better than those

Table 5. Comparing the Hydration Free Energy Predictions of the FiSH Model with Predictions from the Explicit-Solvent LIE Model and with Experimental Hydration Free Energies[a]

	FiSH versus LIE			FiSH versus experiment		
set	MUE	slope	R^2	MUE	slope	R^2
training	0.524 ± 0.033	0.953 ± 0.017	0.946 ± 0.009	0.985 ± 0.066	0.914 ± 0.042	0.806 ± 0.028
testing	0.469 ± 0.020	0.968 ± 0.010	0.968 ± 0.004	1.075 ± 0.052	0.938 ± 0.030	0.826 ± 0.018
SAMPL1	0.958 ± 0.084	0.930 ± 0.018	0.968 ± 0.011	2.173 ± 0.250	0.599 ± 0.043	0.805 ± 0.056

[a] LIE data are for AM1BCC-SP partial charges and rigid solutes.[24] Errors are in kcal mol^{-1} units.

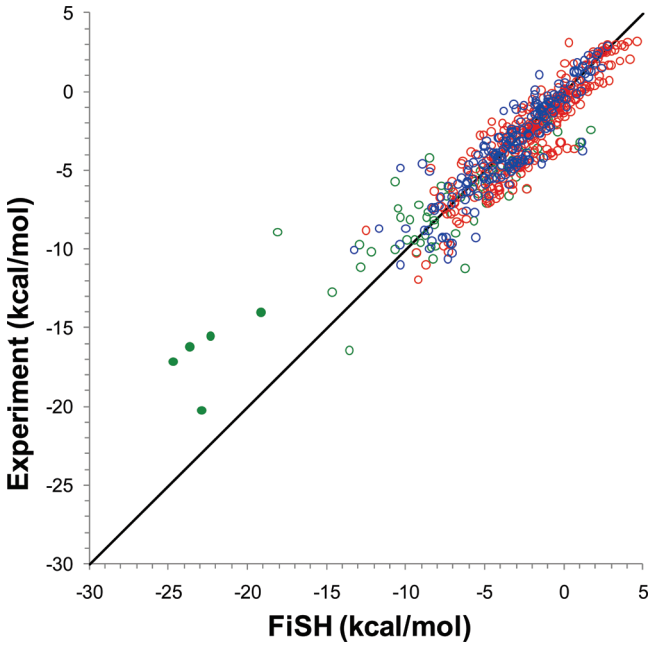

Figure 8. Correlation between hydration free energy predictions with the FiSH model and experimental hydration free energies for the training (blue symbols) and testing (red symbols) data sets, and for the SAMPL1 (green symbols) data set. Filled circles correspond to 5 sulfoneurea analogs from the SAMPL1 data set. The plotted data correspond to the FiSH model with AM1BCC-SP partial charges, and cavity parameters derived from the training data set. The diagonal line indicates ideal correlation.

obtained from the explicit-solvent LIE model with full solute flexibility (MUE of 2.25 kcal/mol), and a little worse than from the LIE model with rigid solute (MUE of 1.92 kcal/mol), for the same partial charge set.[24]

While these data correspond to AM1BCC-SP partial charges, we also calibrated the FiSH model with different partial charge sets, AM1BCC-OPT and RESP. We retrained the cavity component on the training set each time we changed the charge set. The parameters for the cavity term do not vary too much depending on charge (Table S4, Supporting Information). The overall performance of the FiSH models does not depend much on these different charge sets, similar to what was observed with the explicit-solvent LIE models (see Table S5, Supporting Information).[24] Comparable performances in terms of MUEs (training, testing, SAMPL1) were obtained by employing the AM1BCC-SP (0.995, 1.075, 2.173 kcal/mol) or RESP (1.173, 1.068, 2.156 kcal/mol) partial charges in the FiSH models in terms of MUEs. In terms of correlation coefficients and slopes, RESP charges yielded improved slopes (0.793−0.997) compared to those of the AM1BCC-SP charges (0.599−0.938), yet with smaller correlation coefficients (0.660−0.801 for RESP vs 0.805−0.826 for AM1BCC-SP). This decrease in correlation coefficient going from RESP to AM1BCC-SP charges may be partly because the RDF peaks used to define the SAS for the continuum van der Waals term were originally obtained from MD simulations using AM1BCC charges.

Table 6. Listing of Function Groups Used for Error Analysis

functional group	# of members
other	81
alkane	20
alkene	13
alkyne	3
aromatic	18
halogen	57
F	3
Cl	31
Br	12
I	4
OH	27
1° OH	10
2° OH	4
3° OH	2
phenyl OH	11
amine	10
alkyl amine	7
aryl amine	3
carboxylic acid	2
ester	30
amide	2
ether	8
ketone	12
aldehyde	8
nitro	1
cyano	3
hypervalent S	1
thiol	5

Functional Group Analysis of FiSH Continuum Model Predictions. We have separately examined the performance of the derived continuum model on specific chemical classes, on the basis of monofunctional compounds that could be found in the testing set. A majority of functional groups that are commonly encountered in typical drug-like compounds were assessed (see Table 6) in this analysis of FiSH model prediction errors. A similar analysis was previously carried out on the explicit-solvent LIE model of hydration.[24] As seen in Figure 9 (data tabulated in Table S6, Supporting Information), the functional group based error profile of the FiSH continuum model mirrors closely that of the LIE explicit model. The changes in prediction errors between these two models are within 1 kcal/mol for all functional groups investigated. This further emphasizes that the FiSH continuum model succeeded in its primary objective, that is, to mimic a physics-based explicit-solvent hydration model.

In terms of mean-unsigned errors to experimental data, the FiSH continuum model performs well on alkanes (0.394 kcal/mol), alkenes (0.846 kcal/mol), aromatic hydrocarbons (0.691 kcal/mol), chlorinated (0.541 kcal/mol) and iodinated compounds (0.644 kcal/mol), aryl amines (0.410 kcal/mol), esters (0.747 kcal/mol), ethers (0.825 kcal/mol), and ketones (0.500 kcal/mol), with MUE values well below the average for the entire testing set of 1.08 kcal/mol (for AM1BCC-SP partial charges). Interestingly, even though aromatic compounds are predicted well by both the FiSH continuum model and the LIE explicit model when compared to the experiment, the continuum model predictions underestimate LIE predictions by a considerable margin (0.76 kcal/mol). For brominated compounds, neutral carboxylic acids, aldehydes,

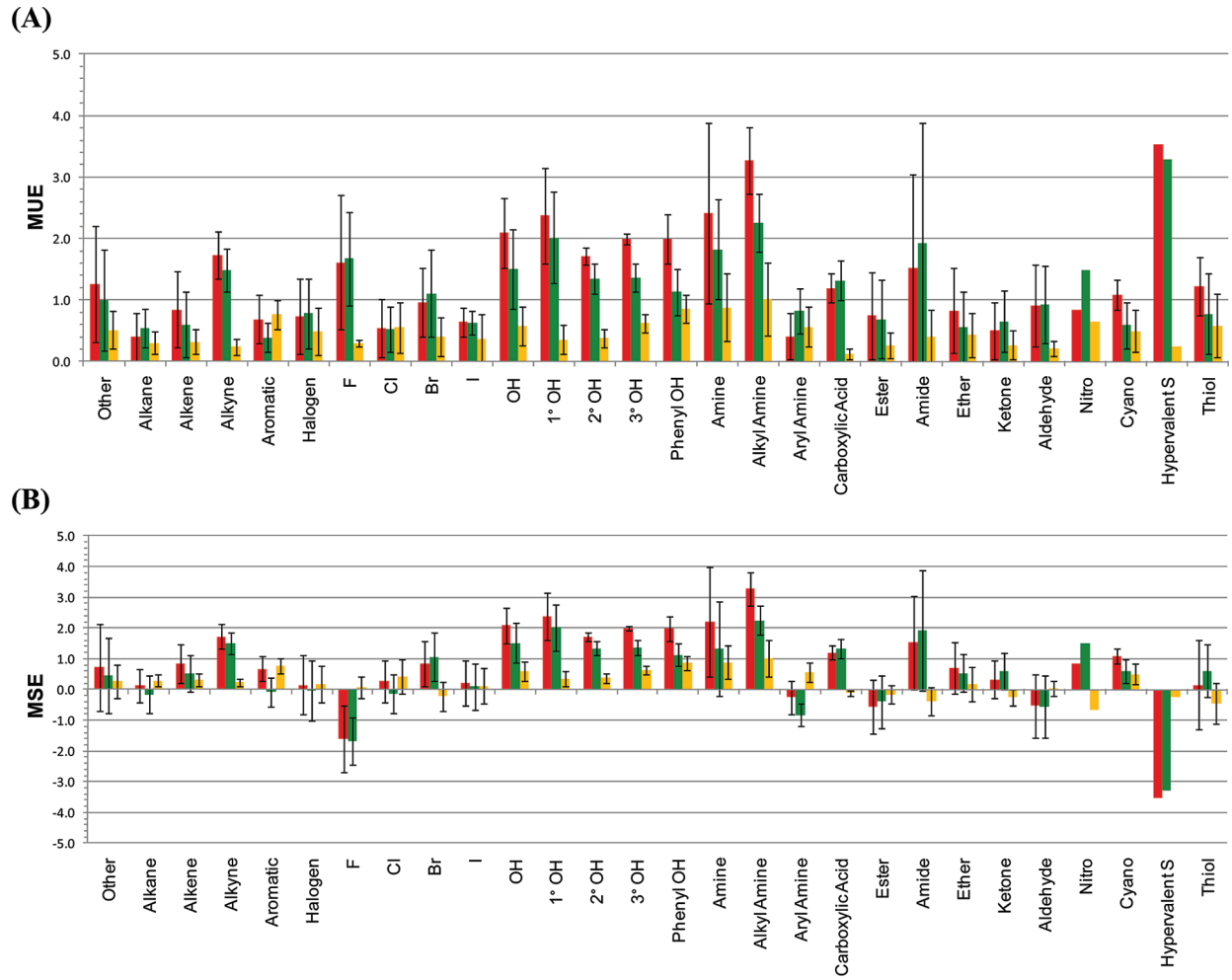

Figure 9. Comparative functional group analysis of prediction errors for the FiSH model and the LIE explicit-solvent model. FiSH model versus experiment (red bars); LIE explicit model versus experiment (green bars); FiSH model versus LIE explicit-solvent model (orange bars). The LIE data are for AM1BCC-SP partial charges and single-conformation solutes and are taken from a separate study.[24] (A) MUE ± SD values. (B) MSE ± SD values.

cyano derivatives, and thiols, the predictions are close to the MUE value of the entire data set, with either the continuum or explicit model having a marginal advantage.

Problematic functional classes for the FiSH continuum model, having MUE values larger than 1.5 kcal/mol, include alkynes (1.730 kcal/mol), fluorinated compounds (1.611 kcal/mol), alcohols (2.090 kcal/mol) and phenols (1.986 kcal/mol), neutral aliphatic amines (3.270 kcal/mol) and amides (1.525 kcal/mol). As seen in Figure 9, these are the same chemical classes that are problematic with the corresponding explicit-solvent LIE model. Similar mean signed errors (FiSH, LIE) were obtained with the two models in the case of alkynes (1.730, 1.491 kcal/mol), fluorinated compounds (−1.611, −1.675 kcal/mol), and amides (1.525, 1.920 kcal/mol). Also, similarly to what was observed with the LIE method, the hydration free energy predictions for alkynes can be significantly improved by employing a FiSH continuum model based on RESP charges (MUE reduced from 1.73 to 0.61 kcal/mol, see Table S6, Supporting Information). Fluorinated compounds were among the few functional classes that were overestimated (Figure 9B). In our FiSH continuum model, the Born radius for the F atom is about 1.72 Å, which is typically less than often used, but which we find is appropriate to mimic well the explicit solvent based data. Hence, nonempirical improvements in the hydration free energy prediction of fluorinated compounds (e.g., not simply based on ad-hoc adjustment of F radius) have to be sought in force field modifications like Lennard-Jones potential parameters or atomic partial charges. Indeed, the FiSH continuum model based on RESP charges does provide some relief in the case of fluorinated compounds, with MUE being reduced from 1.61 to 1.04 kcal/mol. For some chemical classes, particularly the alcohols, phenols, and aliphatic amines, the FiSH continuum model performs poorly (MUEs of 1.72 to 3.27 kcal/mol), and the underestimation of experimental data is accentuated (by 0.3 to 0.9 kcal/mol) with the FiSH continuum model compared to that with the LIE explicit solvent model. In these cases, again, significant improvements can be obtained by employing RESP charges (Table S6). Unfortunately, this does not extend generally to other functional classes, as RESP partial charges degrade the overall predictions obtained with AM1BCC partial charges for both the FiSH and LIE models (Table S6).

Comparison of FiSH Model with Other Continuum Models. The FiSH model is compared with the CED and RF models in Table 7 on the testing set and the SAMPL1 data set. On the training and testing sets, there was little

Table 7. Comparing the Hydration Free Energy Predictions of the FiSH Model with Those from Previously Developed Continuum Solvation Models, Continuum Electrostatics-Dispersion (CED) Solvation Model and Reaction Field (RF) Electrostatics-Only Model[a]

	FiSH		
set	MUE	slope	R^2
training	0.985 ± 0.066	0.914 ± 0.042	0.806 ± 0.028
testing	1.075 ± 0.052	0.938 ± 0.030	0.826 ± 0.018
SAMPL1	2.173 ± 0.250	0.599 ± 0.043	0.805 ± 0.056
	CED[a]		
set	MUE	slope	R^2
training	0.762 ± 0.063	0.881 ± 0.040	0.877 ± 0.034
testing	0.874 ± 0.046	0.872 ± 0.044	0.879 ± 0.023
SAMPL1	2.729 ± 0.331	0.542 ± 0.041	0.818 ± 0.051
	RF[b]		
set	MUE	slope	R^2
training	1.140 ± 0.064	1.150 ± 0.054	0.789 ± 0.035
testing	1.186 ± 0.048	1.309 ± 0.041	0.848 ± 0.017
SAMPL1	1.631 ± 0.188	0.751 ± 0.059	0.780 ± 0.064

[a] CED model at $D_{in} = 1$, $\rho = 0.9$, and 25-atom-type c-vdW parameters, as described previously.[21] [b] RF model at $D_{in} = 1$, $\rho = 1.1$, as described previously.[21] [a] All iodinated molecules have been removed. AM1BCC-SP charges used throughout. Errors are in kcal mol^{-1} units.

variation in the performance of these models in terms of MUE (training, testing), with FiSH model predictions (0.995, 1.075 kcal/mol) marginally outperformed by those of the CED model (0.762, 0.874 kcal/mol). We note that the FiSH model had the correlation slope closest to 1 among the three models. Functional group analysis of mean-unsigned-errors (FiSH vs CED) shows that the difficult functional groups for the FiSH continuum model (alkynes, 1.730 vs 1.127 kcal/mol; fluorinated compounds, 1.611 vs 0.617 kcal/mol; alcohols, 2.090 vs 0.680 kcal/mol; phenols, 1.986 vs 1.083 kcal/mol; and neutral aliphatic amines, 3.270 vs 1.397 kcal/mol) receive better predictions with the CED model (Figure 10, Table S6). The CED model also improves the predictions of brominated compounds (0.985 vs 0.436 kcal/mol), neutral carboxylic acids (1.197 vs 0.140 kcal/mol), and thiols (1.229 vs 0.482 kcal/mol), whereas FiSH continuum model predictions were better on chlorinated compounds (0.541 vs 0.911 kcal/mol), aryl-amines (0.410 vs 1.183 kcal/mol), and cyano derivatives (1.087 vs 1.603 kcal/mol). The slightly better performance of the CED solvation model is understandable since it has many more parameters and was fitted on experimental data for monofunctional compounds from the traning and testing data sets, whereas the FiSH continuum model was trained on LIE data from explicit solvent simulations and therefore mimics LIE's shortcomings.

The CED solvation model, however, has serious transferability problems for the SAMPL1 data set, noted previously,[21] and highlighted in Table 7, that are outside its applicability domain. The predictions on the SAMPL1 data set are improved with the FiSH model (2.173 kcal/mol) relative to the CED solvation model (2.729 kcal/mol), with a MUE decrease of 0.55 kcal/mol and a slightly larger correlation slope (0.599 vs 0.542) to experimental data. In terms of transferability, the increase in MUE from testing set to SAMPL1 data set is 1.1 kcal/mol in the case of the FiSH model, and 1.9 kcal/mol in the case of the CED solvation model. Hence, the FiSH continuum model is more transferable. This supports the hypothesis that the transferability of continuum solvation models can be increased by fitting to physics-based explicit solvation models rather than directly to experimental data. Although performing worst on traditional compounds, the simple RF continuum solvation model outperforms the complex FiSH continuum model on the SAMPL1 data set by a further decrease of 0.55 kcal/mol in MUE and an increase in correlation slope, with a good transferability reflected by only a 0.5 kcal/mol decrease in MUE from the testing set to the SAMPL1 data set. However, the RF continuum model is not practical because it fails on all types of hydrocarbons and primarily on alkanes, noting that hydrocarbon moieties are ubiquitous in organic molecules. Alkanes, alkenes, aromatics, and alkynes all are overestimated (MSE of 1.0 to 2.7 kcal/mol) with the RF model, while large errors are also obtained for fluorinated and iodinated compounds, and for amides (Figure 10, Table S6).

Conclusions

In this paper, we propose a novel continuum solvation model, the First-Shell Hydration (FiSH) model, as an attempt to capture the physics of an explicit solvation model by focusing on the first shell of water around the solute while maintaining the speed provided by the continuum approach. The FiSH continuum model consists of an electrostatic, van der Waals, and cavity contribution to solvation, with only the latter fitted to experimental data. Changes have been introduced to the definition of the continuum electrostatic and van der Waals components, which have been calibrated against explicit-solvent MD simulations *via* the linear interaction energy (LIE) method. The central premise of this study is that the transferability of the continuum model can be increased by reducing the number of parameters fitted directly to the experiment, and by emulating the physics captured by an explicit solvation model. A continuum model designed to mimic an explicit solvent force field model will inherit the transferability and generality of the force field model, for better or for worse.

To capture first hydration shell effects with the FiSH model, we first incorporated charge asymmetry[20,34−37] into the continuum electrostatics model. This was achieved through a modification of our earlier approach of defining the Born radii of atoms as a function of the ISCD.[23] Multiple functional forms were explored and trained on explicit solvent simulations. A nonlinear function with four parameters yielded optimal correlations to explicit water simulations and gave drastic improvements over the initial continuum electrostatic model. A hybrid continuum van der Waals model introduced in this paper creates a first shell of solvent restricted to and distributed uniformly over the SAS. A second region, starting one solvent diameter away from the SAS and extending to infinity, is treated

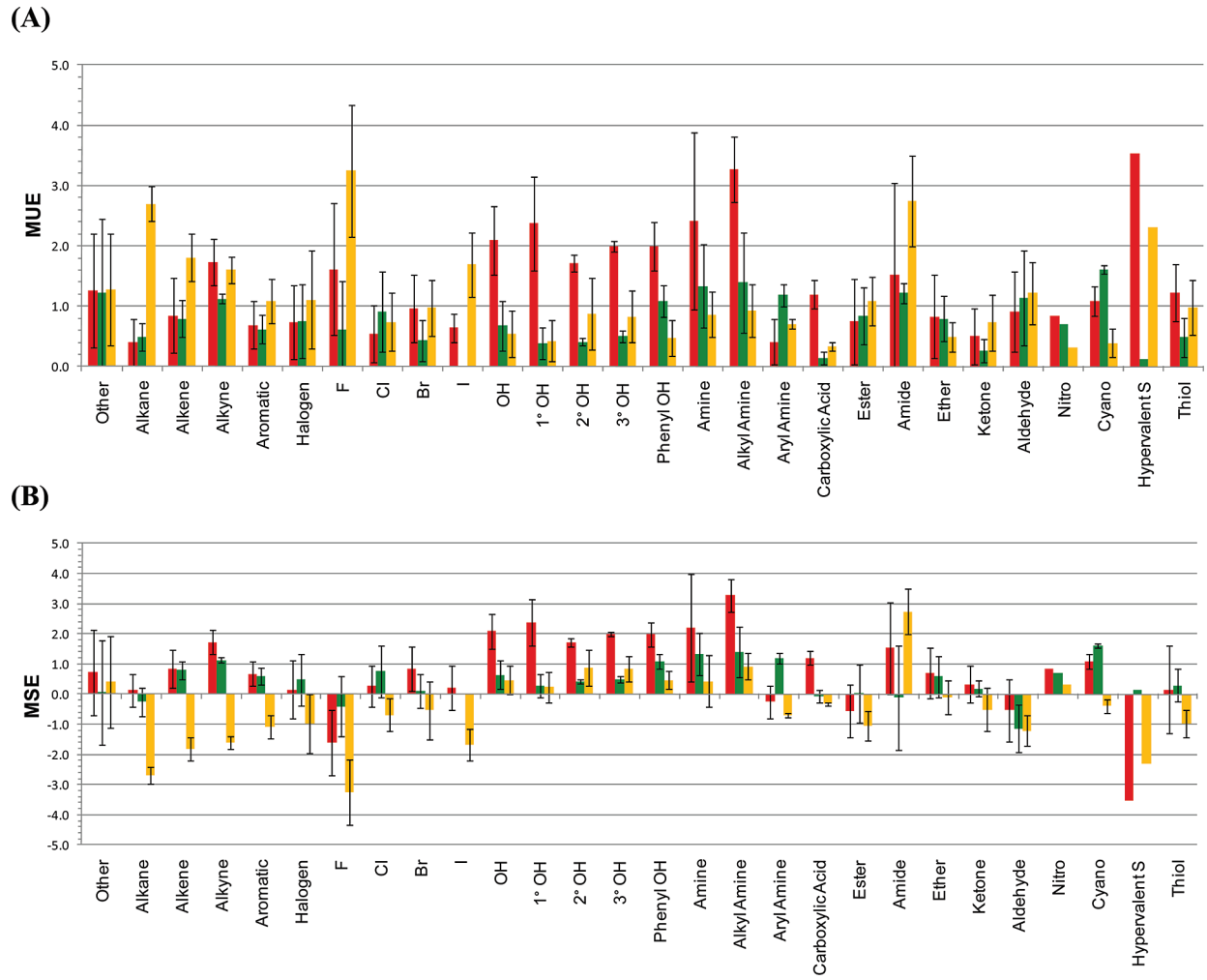

Figure 10. Comparative functional group analysis of prediction errors for the FiSH model, the continuum electrostatics-dispersion (CED) solvation model, and the reaction field (RF) electrostatics-only model. FiSH model versus experiment (red bars); CED model versus experiment (green bars); RF model versus experiment (orange bars). The CED model was employed here with $D_{in} = 1$, $\rho = 0.9$, and 25-atom-type c-vdW parameters, as described previosuly.[21] The RF model was employed here with $D_{in} = 1$, $\rho = 1.1$, as described previously.[21] (A) MUE $\pm$ SD values. (B) MSE $\pm$ SD values.

as a uniform continuum. This model does not require the large number of fitted parameters used in the previous CED model,[21] instead relying entirely on force field parameters for the Lennard-Jones potentials. Testing of the FiSH van der Waals continuum model against the explicit-solvent van der Waals data showed an excellent performance on simple compounds and moderate performance on more complex, drug-like molecules.

The primary objective of the FiSH continuum model, to mimic the hydration free energies from an explicit-solvent model, has been achieved. It predicts the explicit-solvent LIE data with MUEs of about 0.5 kcal/mol for the training and testing data sets and slightly below 1 kcal/mol for the drug-like SAMPL1 data set, with correlation slopes and coefficients close to unity for all three data sets. The excellent agreement carries on to the hydration component terms, as well as to various chemical functional groups commonly present in small organic molecules. The absolute performance against experimental data obtained with the FiSH continuum model is as good as that afforded by the explicit-solvent LIE model, i.e., MUEs of about 1 kcal/mol for the training and testing sets and slightly above 2 kcal/mol for the SAMPL1 data set. Another similarity to the explicit-solvent model is the weak dependence of the overall performance of the FiSH continuum model on the tested partial charge sets. There is, however, an uneven impact of the charging method across functional classes, with RESP charges providing better prediction than AM1BCC charges on certain chemical classes that are poorly predicted (e.g., alkynes, fluorinated compounds, alcohols, phenols, and aliphatic amines), but worse predictions on others.

Another objective that has been achieved with the FiSH continuum model is the improvement of transferability relative to previously developed CED solvation model that has been (over)fitted against experimental data. Comparatively, the transferability of the FiSH continuum model is improved by about 0.8 kcal/mol between simple compounds from the training and testing data sets over the more complex molecules found in the SAMPL1 data set when compared to the CED solvation model. On the basis of a very

acceptable performance, the sound physical foundation of the FiSH continuum model is an important attribute that should not be overlooked when compared to other models in terms of global fitness measures.

Acknowledgment. This is National Research Council of Canada publication number 50689.

Supporting Information Available: Composition of the hydration data sets with experimental transfer free energies (Table S1). Optimized parameters for Born radii correction functions (Table S2). LIE data for spherical solutes (Table S3). Cavity parameters calibrated on the training subset for various FiSH models (Table S4). Errors for various charging methods with FiSH continuum model (Table S5). Raw data for Figures 9 and 10 (Table S6). Plots of Born radii correction functions trained on spheres or bracelets (Figure S1). Distribution of atomic ISCD within the training, testing, and SAMPL1 data sets (Figure S2). Correlation plots for the van der Waals and total nonpolar components of the FiSH continuum model against the solute molecular surface area (Figure S3). This material is available free of charge via Internet at http://pubs.acs.org.

References

(1) Gilson, M. K.; Zhou, H. X. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(2) McInnes, C. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494–502.

(3) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(4) Kang, Y. K.; Némethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4109–4117.

(5) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(6) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385–16398.

(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) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.

(9) Tan, C.; Yang, L.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 18680–18687.

(10) Reddy, M. R.; Erion, M. D. *Free Energy Calculations in Rational Drug Design*; Springer-Verlag: New York, 2001.

(11) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. *Protein Eng.* **1992**, *5*, 215–228.

(12) Aqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng.* **1994**, *7*, 385–391.

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

(14) Su, Y.; Gallicchio, E.; Das, K.; Arnold, E.; Levy, R. M. *J. Chem. Theory Comput.* **2006**, *3*, 256–277.

(15) Chen, J.; Brooks, C. L., III; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

(16) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(17) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243–252.

(18) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(19) Raschke, T. M.; Levitt, M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6777–6782.

(20) Mobley, D. L.; Barber, A. E.; Fennell, C. J.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 2405–2414.

(21) Sulea, T.; Wanapun, D.; Dennis, S.; Purisima, E. O. *J. Phys. Chem. B* **2009**, *113*, 4511–4520.

(22) Nicholls, A.; Wlodek, S.; Grant, J. A. *J. Phys. Chem. B* **2009**, *113*, 4521–4532.

(23) Purisima, E. O.; Sulea, T. *J. Phys. Chem. B* **2009**, *113*, 8206–8209.

(24) Sulea, T.; Corbeil, C. R.; Purisima, E. O. *J. Chem. Theory Comput.* **2009**, DOI: 10.1021/ct9006025.

(25) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(26) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(27) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

(28) Purisima, E. O.; Nilar, S. H. *J. Comput. Chem.* **1995**, *16*, 681–689.

(29) Purisima, E. O. *J. Comput. Chem.* **1998**, *19*, 1494–1504.

(30) Chan, S. L.; Purisima, E. O. *J. Comput. Chem.* **1998**, 1268–1277.

(31) Bhat, S.; Purisima, E. O. *Proteins* **2006**, *62*, 244–261.

(32) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616–627.

(33) Floris, F. M.; Tomasi, J.; Pascual-Ahuir, J. L. *J. Comput. Chem.* **1991**, *12*, 784–791.

(34) Latimer, W. M.; Pitzer, K. S.; Slansky, C. M. *J. Chem. Phys.* **1939**, *7*, 108–111.

(35) Rashin, A. A.; Honig, B. *J. Phys. Chem.* **1985**, *89*, 5588–5593.

(36) Roux, B.; Yu, H. A.; Karplus, M. *J. Phys. Chem.* **1990**, *94*, 4683–4688.

(37) Babu, C. S.; Lim, C. *J. Phys. Chem. B* **1999**, *103*, 7958–7968.

(38) Chorny, I.; Dill, K. A.; Jacobson, M. P. *J. Phys. Chem. B* **2005**, *109*, 24056–24060.

(39) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(40) Born, M. *Z. Phys.* **1920**, *1*, 45–48.

(41) Cerutti, D. S.; Baker, N. A.; McCammon, J. A. *J. Chem. Phys.* **2007**, *127*, 155101–155112.

(42) Tan, C.; Tan, Y. H.; Luo, R. *J. Phys. Chem. B* **2007**, *111*, 12263–12274.

(43) Orozco, M.; Luque, F. J. *Chem. Phys. Lett.* **1997**, *265*, 473–480.

(44) Westergren, J.; Lindfors, L.; Höglund, T.; Lüder, K.; Nordholm, S.; Kjellander, R. *J. Phys. Chem. B* **2007**, *111*, 1872–1882.

(45) Almlof, M.; Carlsson, J.; Aqvist, J. *J. Chem. Theory Comput.* **2007**, *3*, 2162–2175.

(46) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

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

(48) 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.

(49) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(50) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(51) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005.

(52) Shivakumar, D.; Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.

CT9006037