Rapid prediction of solvation free energy. 3. Application to the SAMPL2 challenge

**Enrico O. Purisima · Christopher R. Corbeil ·
Traian Sulea**

Abstract The SAMPL2 hydration free energy blind prediction challenge consisted of a data set of 41 molecules divided into three subsets: explanatory, obscure and investigatory, where experimental hydration free energies were given for the explanatory, withheld for the obscure, and not known for the investigatory molecules. We employed two solvation models for this challenge, a linear interaction energy (LIE) model based on explicit-water molecular dynamics simulations, and the first-shell hydration (FiSH) continuum model previously calibrated to mimic LIE data. On the 23 compounds from the obscure (blind) dataset, the prospectively submitted LIE and FiSH models provided predictions highly correlated with experimental hydration free energy data, with mean-unsigned-errors of 1.69 and 1.71 kcal/mol, respectively. We investigated several parameters that may affect the performance of these models, namely, the solute flexibility for the LIE explicit-solvent model, the solute partial charging method, and the incorporation of the difference in intramolecular energy between gas and solution phases for both models. We extended this analysis to the various chemical classes that can be formed within the SAMPL2 dataset. Our results strengthen previous findings on the excellent accuracy and transferability of the LIE explicit-solvent approach to predict transfer free energies across a wide spectrum of functional classes. Further, the current results on the SAMPL2 test dataset provide additional support for the FiSH continuum model as a fast yet accurate alternative to the LIE explicit-solvent model. Overall, both the LIE explicit-solvent model and the FiSH continuum solvation model show considerable improvement on the SAMPL2 data set over our previous continuum electrostatics-dispersion solvation model used in the SAMPL1 blind challenge.

Keywords Hydration · Continuum solvation · LIE · Continuum van der Waals · First hydration shell · Prospective study

Introduction

Changes in hydration free energy during complex formation are a crucial element of binding free energies [1–3]. With the use of methods for the prediction of binding free energies becoming commonplace in the field of drug design, there is a need for solvation methods that are both quick and highly accurate [4–6]. With this need, much research has been carried out on the improvement of existing and development of new solvation models at many levels of theory [7–13]. Explicit solvation models such as free energy perturbation (FEP), thermodynamic integration (TI) [1, 14] and the faster linear interaction energy models (LIE) [15, 16], offer detail on the distinct nature of water around the solute and are transferable across a wide range of datasets. The caveat of these solvation models for their application in the field of drug design is their lack of throughput. Implicit solvation models offer a quicker alternative to explicit models by replacing the individual water molecules with a continuous medium [17–20]. For small organic molecules, the loss of molecular detail of the

E. O. Purisima (✉) · C. R. Corbeil · T. Sulea
Biotechnology Research Institute, National Research Council
Canada, 6100 Royalmount Avenue, Montreal, QC H4P 2R2,
Canada
e-mail: enrico.purisima@cnrc-nrc.gc.ca

solvent results in relatively small differences between hydration free energy prediction accuracies calculated with implicit solvent models relative to the explicit treatment [21–25]. When applied to complex biomolecular systems, this loss of detail may become problematic at locations where water does not behave as a continuous medium, for example the individual water molecules ordered in concave pockets at the surface of proteins [4, 26].

An issue arises when addressing the requirement for parameterization. Unlike in the field of prediction of binding affinities where there is a plethora of known experimental data on a diverse set of molecules and targets [27–29], there is a relatively limited collection of measured experimental hydration free energies [22, 30]. Recently, the SAMPL initiative [21, 31] has been established to find new sources for experimental hydration free energies and actively engaged in pushing forward the accuracy of solvation models through blind prediction challenges of hydration free energy. The SAMPL1 blind prediction challenge produced a dataset of 63 functionally diverse neutral drug-like molecules consisting of larger magnitudes of hydration free energies and molecular weights than in present hydration data sets [31]. With the success of the first prediction challenge, the SAMPL initiative decided to broaden the scope of the competition to not only predict hydration free energies, but to offer an open challenge to experimentalists. This was done by not only giving a set of molecules with known but withheld hydration free energies, but by creating a set of molecules with no known experimental values to demonstrate to experimentalists that measurements of hydration free energies for these molecules are attainable.

We participated in the SAMPL1 competition with a continuum electrostatics-dispersion (CED) solvation model with a reasonable success [32]. We have recently revisited the SAMPL1 set with both an explicit-solvent model using the LIE approximation [33] and with FiSH (First Shell Hydration), our refined continuum solvation model [34]. Both these models showed an increase in accuracy over our previous CED model. Both the LIE explicit-solvent model and the FiSH continuum model were prospectively submitted to the SAMPL2 blind challenge. Here, using the newer SAMPL2 dataset, we will present the prediction accuracies of these solvation models, test their sensitivities to several model parameters, and compare their performances relative to previous continuum models.

Materials and methods

SAMPL2 hydration data sets

The SAMPL2 hydration data consisted of 41 polyfunctional neutral compounds subdivided into explanatory,

obscure and investigatory sets. Hydration free energies for the explanatory set (8 compounds) were given before the final submission while the obscure set (23 compounds) values were withheld and therefore resulted in a blind prediction challenge. The investigatory set consists of 10 compounds that currently have no known experimental hydration free energies and therefore represents truly blind but currently unverifiable predictions. Since no experimental hydration free energies are currently available for the compounds in this set, it will not be discussed further with results listed in Table S2 (see Table S1 for model numbers).

Another goal of using the SAMPL2 data set is to test our solvation models for systematic errors. To this end, we have grouped the obscure set into four chemical classes (Table 1): parabens (esters of p-substituted benzoic acid), profens (derivatives of 2-arylpropionate), sugars, and uracils. The remaining molecules are non-classified. In the analyses presented in this paper, the updated values of experimental hydration free energies for cyanuric acid (from the obscure set) and glycerol (from the explanatory set) were used [35].

The SAMPL2 hydration data set was prepared in the same manner as published previously for the preparation of the traditional and SAMPL1 data sets [33]. (In this paper, we will refer to the data set used in the Mobley et al. study [23] as the "traditional" data set in contrast to the more complex and more polyfunctional SAMPL1 data set.) Briefly, the SAMPL2 set was downloaded from the appropriate sources and underwent energy minimization using MMFF94 force-field [36]. The resulting geometries were then used as input for generation of partial charges using the AM1BCC-SP, AM1BCC-OPT [37, 38] (implemented within QUACPAC [39]) and RESP methods as described previously [33].

LIE explicit solvation models

We have recently explored the sensitivity of the LIE explicit-solvent models to various parameters [33]. The method applied in the previous report will be used herein and will only be described briefly. We used the following implementation of the LIE approach:

$$\Delta G_{\text{hyd}}^{\text{LIE}} = \alpha \underbrace{\left(\langle E_{\text{S-W}}^{\text{Coul}} \rangle_{\leq 12\text{Å}} + \langle G_{\text{S}}^{\text{RF}} \rangle_{12\text{Å}-\infty} \right)}_{\text{electrostatic}}$$
$$+ \beta \underbrace{\left(\langle E_{\text{S-W}}^{\text{vdW}} \rangle_{\leq 12\text{Å}} + \langle E_{\text{S}}^{\text{cvdW}} \rangle_{12\text{Å}-\infty} \right)}_{\text{van der Waals}}$$
$$+ \underbrace{\gamma_{\text{cav}} \langle \text{MSA} \rangle + \text{C}}_{\text{cavity}} + \underbrace{\langle E_{\text{aq}}^{\text{intra}} \rangle - \langle E_{\text{gas}}^{\text{intra}} \rangle}_{\text{intramolecular}} \quad (1)$$

Table 1 Composition of chemical classes

Class	Molecule name
Non-classified	Acetylsalicylic acid
Non-classified	Caffeine
Non-classified	Cyanuric acid
Non-classified	Diflunisal
Non-classified	Phthalimide
Non-classified	Sulfolane
Parabens	Butyl-paraben
Parabens	Ethyl-paraben
Parabens	Methyl-paraben
Parabens	Propyl-paraben
Profens	Flurbiprofen
Profens	Ibuprofen
Profens	Ketoprofen
Profens	Naproxen
Sugars	D-glucose
Sugars	D-xylose
Uracils	5-bromouracil
Uracils	5-chlorouracil
Uracils	5-flurouracil
Uracils	5-iodouracil
Uracils	5-trifluoromethyluracil
Uracils	6-chlorouracil
Uracils	Uracil

Table 2 Complete list of all solvation models investigated on the SAMPL2 data sets

Submitted	Method	Charges	Solute	Internal energy
		Explicit solvation models		
	LIE	AM1BCC-SP	Flexible	
	LIE	AM1BCC-SP	Flexible	All
	LIE	AM1BCC-SP	Flexible	Non-bond terms
Y	**LIE**	**AM1BCC-SP**	**Rigid**	
	LIE	AM1BCC-OPT	Flexible	
	LIE	AM1BCC-OPT	Flexible	All
	LIE	AM1BCC-OPT	Flexible	Non-bond terms
	LIE	AM1BCC-OPT	Rigid	
	LIE	RESP	Flexible	
	LIE	RESP	Flexible	All
	LIE	RESP	Flexible	Non-bond terms
	LIE	RESP	Rigid	
		Continuum solvation models		
Y	**FiSH**	**AM1BCC-SP**	**Rigid**	
Y	**FiSH**	**AM1BCC-SP**	**Rigid**	**All**
	FiSH	AM1BCC-SP	Rigid	Non-bond terms
	FiSH	AM1BCC-OPT	Rigid	
	FiSH	AM1BCC-OPT	Rigid	All
	FiSH	AM1BCC-OPT	Rigid	Non-bond terms
	FiSH	RESP	Rigid	
	FiSH	RESP	Rigid	All
	FiSH	RESP	Rigid	Non-bond terms
Y	**RF**	**AM1BCC-SP**	**Rigid**	
	CED	AM1BCC-SP	Rigid	

Highlighted in bold are the models submitted for consideration for the SAMPL2 blind challenge

LIE linear interaction energy explicit-solvent model, *FiSH* first-shell hydration continuum model, *CED* continuum electrostatics-dispersion model, *RF* reaction field continuum electrostatic-only model

where the Coulomb, van der Waals, cavity and internal energy terms all represent averages over snapshots taken from the last nanosecond of a 2 ns MD trajectory. Explicit-solvent MD simulations were carried out with the solute surrounded by a 12 Å shell of TIP3P water, using AMBER 9 [40, 41] with GAFF parameters [42]. Continuum models were applied for corrections outside the explicit-water shell, $\langle G_S^{RF} \rangle_{12\text{Å}-\infty}$ and $\langle E_S^{cvdW} \rangle_{12\text{Å}-\infty}$, for the electrostatic [33] and van der Waals [43, 44] components, respectively. The values of α and β were set to 0.5 and 1.0, respectively. The cavity term was calibrated previously against pseudo-experimental cavity free energies obtained by training to a subset of the traditional dataset [33]. Internal energies in the gas phase, $\langle E_{gas}^{intra} \rangle$, were obtained from separate MD simulations in vacuum. This LIE model was found to give excellent correlative models for both the traditional [33] and drug-like SAMPL1 data sets [31]. On those data sets, partial charges, solute flexibility and internal energy terms influenced the accuracy of the LIE solvation model [33]. All these parameter variations will be revisited here on the SAMPL2 dataset. A complete listing of LIE solvation models used in this study can be found in Table 2 (see also Table S1). The LIE model with AM1BCC-SP charges and rigid solute was submitted for consideration for SAMPL2 blind competition.

FiSH continuum solvation models

The FiSH continuum solvation model has been developed to mimic the accuracy of an explicit-solvent LIE hydration model [34]. This continuum model has been trained on hydration data from explicit-solvent simulations rather than on experimental hydration data in order to enable the calibration of its electrostatic and non-polar components on the corresponding terms available from explicit-solvent simulations but not accessible experimentally. Furthermore, the LIE approach was favoured over FEP-like methods for the calibration of this continuum model due to its term decomposition that is simpler and more compatible with the solvation contributions typically calculated with continuum models, and its similar accuracy in predicting hydration free energies relative to the more expensive FEP approach. A complete report of the FiSH continuum model has been described elsewhere [34] and only a short outline will follow.

The FiSH continuum model includes an electrostatic, a van der Waals and a cavity term similar to the explicit-solvent LIE hydration model it was trained and tested on,

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

The Born radii are derived from the GAFF van der Waals radii with a correction based on the induced surface charge density (ISCD) on the molecular surface [34]. The radius correction is designed to capture the charge asymmetry in the response of water to positively and negatively charged groups [45–50]. The solute reaction field energy, $\Delta G_{\text{hyd}}^{\text{R}}$, is calculated using a boundary element (BEM) solution of the Poisson equation using the BRI BEM program [51, 52]. The correction function parameters were trained on the electrostatic component from the LIE model. The FiSH model also replaces the single surface integral of the Lennard-Jones potential over the molecular surface of the solute by a 2-region representation of the continuum van der Waals term, U^{vdw} [34]. The contribution of the first shell, represented by the solvent-accessible surface (SAS), is calculated separately from the rest of the solvent. The water molecules in the first shell are assumed to be restricted to the SAS and uniformly distributed along this surface. Their contribution to U^{vdw} is calculated as an integral of the Lennard-Jones interaction between the solute atoms and the SAS [34]. The second region, defined to start at the SAS + 2.8 Å (a water diameter), incorporates contributions from the second hydration shell outwards to infinity *via* a traditional van der Waals continuum model. Finally, a term proportional to the molecular surface area of the solute, $\gamma_{\text{cav}} MSA$, and trained on pseudo-experimental free energies is employed to account for the cost of cavity formation. This model reproduced successfully the hydration free energies predicted by the explicit-solvent LIE model, and yielded an improvement over the previously used CED solvation model on the SAMPL1 test set [34]. All parameters applied in this study on the SAMPL2 test set are identical to those developed previously [34].

As in the LIE models, we would like to investigate the inclusion of the difference in the intramolecular energy of the solute between the aqueous and gas phases. Ideally, a complete partition function would be used to determine the hydration free energy model. Due to time constraints imposed by the SAMPL2 competition deadline, a simpler model was proposed. The solute molecule is separately optimized by energy minimization using a distance-dependent ($4R_{\text{ij}}$) dielectric constant or a dielectric constant of 1, to simulate the single dominant conformation attained in aqueous and gas phases, respectively [53]. The structures were conjugate-gradient energy-minimized using the

GAFF force-field [42]. The electrostatic term is then recalculated at a dielectric constant of 1 for both energy-minimized conformations resulting in the following:

$$\Delta G_{\text{hyd}}^{\text{FiSH}}\left(\{r_i^{\text{Born}}\}, \gamma_{cav}\right) = \Delta G_{\text{hyd}}^{\text{R}}\left(\{r_i^{\text{Born}}\}\right) + U^{\text{vdw}}$$
$$+ \gamma_{\text{cav}} MSA + C$$
$$+ \underbrace{\left(E_{\text{intra}}^{\min D=4R} - E_{\text{intra}}^{\min D=1}\right)}_{\text{intramolecular}} \qquad (3)$$

where $E_{\text{intra}}^{\min D=4R}$ and $E_{\text{intra}}^{\min D=1}$ are intramolecular energies. As with the LIE models, we also investigated using all energy terms or retaining just the non-bond terms (1–4 and 1–5 van der Waals and electrostatic). A list of all FiSH solvation models used in this study can be found in Table 2 (see also Table S1). Two of these models were submitted for the SAMPL2 blind competition, both based on AM1BCC-SP charges, with and without all internal energy terms.

Other continuum solvation models

As this is our second time participating in the SAMPL blind prediction challenge, we also wish to demonstrate the improvement of our newer solvation models *versus* our previously used solvation models. We therefore re-investigated the two solvation models used previously in SAMPL1 : an electrostatic-only reaction field model (RF) [51, 52] and a continuum electrostatics-dispersion (CED) solvation model [32]. All parameters are identical to those previously published and therefore will only be described briefly below.

The RF model is a pure electrostatic continuum solvation model consisting of a boundary element solution of the Poisson equation using the program BRI BEM [51, 52]. It has shown surprising correlation when used on the SAMPL1 test set [32]. Even with this success, it has known limitations when dealing with nonpolar molecules [34]. The RF model with a solute dielectric constant, D_{in}, of 1 and a block-scaling factor for the AMBER van der Waals radii, ρ, of 1.1, was employed prospectively in SAMPL2 challenge.

The CED solvation model [32] has the following functional form:

$$\Delta G_{\text{hyd}}^{\text{CED}}(D_{\text{in}}, \rho, \gamma_{cav}, \{B_i\}) = \Delta G_{\text{hyd}}^{\text{R}}(D_{\text{in}}, \rho) + \gamma_{\text{cav}} MSA$$
$$+ \sum_i U_i^{\text{cvdW}}(B_i) + C \qquad (4)$$

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 parameters were trained on a set of 129 neutral polar molecules. The electrostatic contribution,

ΔG_{hyd}^{R}, was calculated using the BRI-BEM program [51, 52]. The cavity contribution is proportional to the total molecular surface area, *MSA*, which was calculated using a variable surface probe. The dispersion-repulsion term, U_i^{cvdW}, was calculated by integrating the 6–12 Lennard–Jones potential over the molecular surface for a set of defined atom types [43, 44, 54–57]. This model yielded very good results on a test set of simple solutes similar to those used for its calibration. Application to the challenging drug-like SAMPL1 dataset revealed the rather limited transferability of the CED solvation model and later prompted the development of the FiSH continuum model. The CED solvation model was not submitted prospectively to the SAMPL2 challenge but will be presented here.

Results and discussion

SAMPL2 submissions

Our submissions for the SAMPL2 blind prediction challenge consisted of 4 submitted models: 1 using the LIE solvation model, 2 using the FiSH solvation model and 1 using the previously used RF model. These models were selected based on their performance on the SAMPL1 and SAMPL2 explanatory set. A fifth model was also submitted using the FiSH model with RESP charges but is not discussed further due to a re-optimization of parameters following the SAMPL2 submission deadline.

The explicit-solvent LIE model based on AM1BCC-SP partial charges, rigid solute geometry, and without inclusion of intramolecular energy change between the gas and aqueous phases was submitted to the SAMPL2 challenge (red bars in Fig. 1). This model achieved mean-unsigned-errors (MUEs) of 1.46 kcal/mol for the explanatory dataset, 1.69 kcal/mol for the obscure dataset and 1.63 kcal/mol for the combined (explanatory and obscure) dataset. For the combined dataset, this model generated absolute hydration free energy predictions highly correlated with the experiment, characterized by a squared correlation coefficient (R^2) of 0.92 and a correlation slope of 0.91.

We submitted to the SAMPL2 challenge two FiSH continuum solvation models featuring AM1BCC-SP solute partial charges. One model, which does not include the change in intramolecular energy of the solute between the gas and water phases (orange bars in Fig. 1), achieved MUEs of 2.29 kcal/mol for the explanatory dataset, 1.71 kcal/mol for the obscure dataset and 1.86 kcal/mol for the combined (explanatory and obscure) dataset. The other model, which includes a simple accounting of the intramolecular energy change as described in the Materials and Methods section (green bars in Fig. 1), achieved MUEs of 2.50 kcal/mol for the explanatory dataset, 2.73 kcal/mol
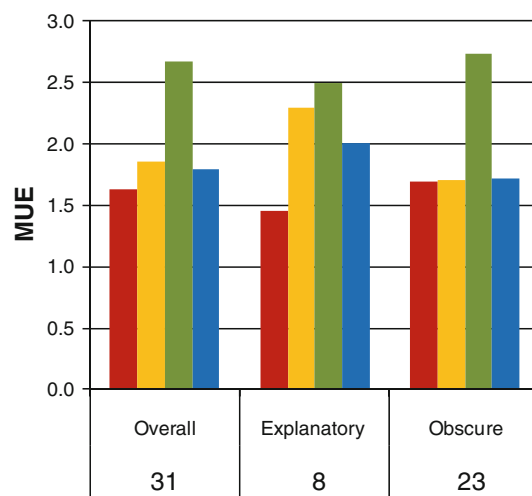

Fig. 1 Mean unsigned error (MUE, kcal/mol) for various solvation models. *Red* = LIE explicit-solvent model with rigid solute and AM1BCC-SP charges. *Orange* = FiSH continuum model with AM1BCC-SP charges. *Green* = FiSH continuum model with AM1BCC-SP charges and including the difference in internal energy. *Blue* = RF solvation model. (The reported MUE values include cyanuric acid. Although poorly predicted, its inclusion in the analysis does not alter the MUE values much.)

for the obscure dataset and 2.67 kcal/mol for the combined dataset. It is apparent that the inclusion of intramolecular terms worsened the FiSH continuum model prediction accuracy, particularly on the obscure set.

Lastly, we submitted the RF model previously used in SAMPL1 (blue bars in Fig. 1). The model achieved MUEs of 2.00 kcal/mol for the explanatory dataset, 1.72 kcal/mol for the obscure dataset and 1.79 kcal/mol for the combined (explanatory and obscure) dataset. The RF model over the complete SAMPL2 dataset yielded a correlation slope close to unity (1.06) and an R^2 of 0.88. Even with these reasonable results the RF model has known deficiencies (prediction of hydration free energy of alkanes) which preclude its general use.

LIE explicit solvation models

In a previous application of the LIE method to the traditional and the drug-like SAMPL1 hydration datasets [33], AM1BCC-SP partial charges yielded slightly better accuracy relative to AM1BCC-OPT and RESP charges, but the accuracy was highly variable on specific functional classes. Therefore, the influence of partial charge set was investigated for the SAMPL2 data set. In this section we present the effect of partial charges on LIE models with flexible solute geometries with and without the inclusion of internal energy terms. Predicted hydration free energies for individual compounds and performance indicators for other LIE models generated are given in Tables S3 and S4. In

this and following sections the numbers in parentheses represent the MUE to experiment.

As in the previous study [33], modest variations in the LIE prediction accuracy occur while varying partial charges on the SAMPL2 hydration dataset (in Fig. 2 red bars for AM1BCC-SP charges, orange bars for AM1BCC-OPT charges and green bars for RESP charges). The RESP partial charges are preferred on the combined SAMPL2 dataset (1.46 kcal/mol), followed by AM1BCC-OPT (1.86 kcal/mol) and AM1BCC-SP (1.94 kcal/mol) partial charges. For those LIE models, RESP charges also led to improved correlation slopes and R^2 values over AM1BCC charges (see Table S4). RESP partial charges are preferred on both explanatory and obscure subsets, with MUEs of 0.88 and 1.67 kcal/mol, respectively.

With flexible solute geometries, RESP charges provide the best predictions in the case of profens (1.20 kcal/mol), sugars (1.77 kcal/mol) and non-classified compounds (1.03 kcal/mol) relative to AM1BCC-OPT and AM1BCC-SP charges. The differences in the profens and non-classified molecules arise from a difference in the polarization of the aromatic rings by AM1BCC-SP and AM1BCC-OPT charges when compared to RESP charges. The underestimation of the sugars' free energies of solvation with AM1BCC charges may result from underpolarization of the anomeric carbon relative to the RESP charging method. In the case of parabens, both AM1BCC-OPT (1.41 kcal/mol) and AM1BCC-SP (1.55 kcal/mol) charges perform similarly with RESP. This trend carries onto uracils, having AM1BCC-OPT (1.25 kcal/mol) charges yielding the most accurate results, followed closely by the AM1BCC-SP (1.66 kcal/mol) and lastly by the RESP (2.31 kcal/mol) charges. The drop in accuracy with the RESP charges may result from a slight overpolarization of the partial charges

over the entire uracil moiety. For a complete list of errors by functional class for all LIE solvation models, see Table S5.

Next, we investigated the effect of solute flexibility during MD on the LIE predictions of hydration free energies. Due to the large number of combinations of charge sets and flexibility conditions, we will focus only on the effect of flexibility of the solute with AM1BCC-SP charges (in Fig. 2, red bars for flexible solute *versus* light blue bars for rigid solute), which, based on previous extensive testing, were found most appropriate for LIE predictions of hydration free energies [33]. The results for all other LIE models based on other solute partial charges, and with or without consideration of intramolecular energy terms, can be found in Table S3 and Table S4.

Overall, for AM1BCC-SP charges LIE predictions based on rigid solute geometries (1.63 kcal/mol) are better than for flexible solute geometries (1.94 kcal/mol). However, the correlation slope (1.01) to experimental data is closer to ideal with flexible solute geometries than with rigid solute geometries (slope of 0.91) (Table S4). When rigidifying the solute the advantage of RESP over AM1BCC partial charges was not observed (see Table S4). For the obscure set, rigidifying the solute improves accuracy (1.69 kcal/mol) over the flexible-solute model (2.10 kcal/mol). The overall increase in accuracy from rigidifying the solute is largely influenced by to the two sugar molecules and partly due to the fortuitous selection of the conformation for the rigid simulation [24]. Looking at the various chemical classes, keeping the solute rigid greatly improves accuracy of the sugars (0.79 kcal/mol) over the flexible solute model (4.62 kcal/mol). However, this result may be a fortuitous since sugars are flexible molecules in water and rigidifying the solute may only provide a compensation that partially corrects the underpolarization of sugars mentioned above. The initial energy-minimized (with a dielectric of 4R)
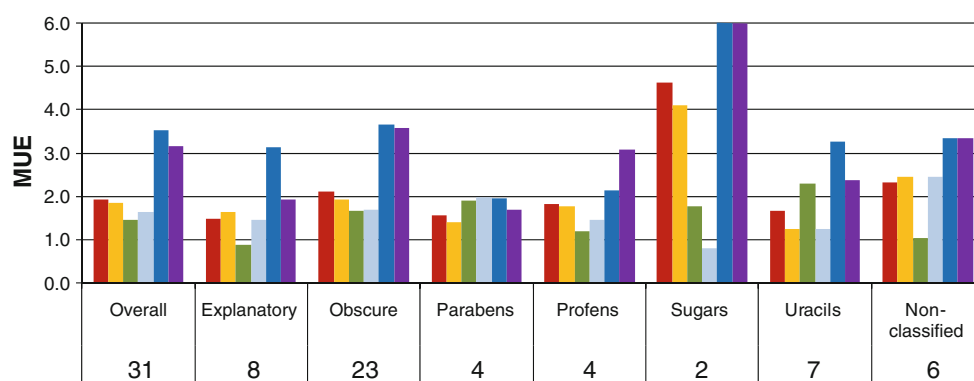

Fig. 2 Mean unsigned error (MUE, kcal/mol) of selected LIE explicit-solvent models. *Red* = AM1BCC-SP charges with the solute flexible. *Orange* = AM1BCC-OPT charges with the solute flexible. *Green* = RESP charges with the solute flexible. *Light blue* = AM1BCC-SP charges with the solute rigid. *Dark blue* = AM1BCC-SP charges with flexible solute and internal energy (all terms). *Purple* = AM1BCC-SP charges with flexible solute and internal energy (excluding covalent terms). *Capped bars* represent errors greater than 6 kcal/mol

structures used for the rigid geometry simulations disrupt or weaken the internal hydrogen bonds of the sugars. This leads to greater interaction with the solvent compensating for the underpolarized charges. In a flexible-geometry LIE MD simulation, these internal hydrogen bonds would be re-formed part of the time, weakening the interaction with solvent. In contrast, if RESP charges are used, the accuracy decreases when rigidifying the solute (4.25 kcal/mol) compared to the flexible-solute model (1.77 kcal/mol). The RESP charges are more polarized and will cause an over-estimation of the solvation free energy of the sugars unless internal hydrogen bonds are formed part of the time, which they are not if a rigid conformation is used.

Adding the change in solute intramolecular energy between the gas and aqueous phases (in Fig. 2, dark blue and purple bars are for inclusion of some or all intramolecular terms *versus* red bars without internal energies) did not improve the LIE predictions for the overall SAMPL2 data set. For AM1BCC-SP charges and flexible solute geometries, the MUE with the inclusion of all intramolecular terms is 3.56 kcal/mol, and is 3.16 kcal/mol if we exclude bonded terms from the internal energy. This is in contrast with previous LIE results on the SAMPL1 hydration dataset [33], where adding non-bonded intramolecular terms improved LIE predictions.

Furthermore, no chemical class from the SAMPL2 hydration dataset could be improved by adding internal energy. This is not surprising, since the obscure set is made up of mostly rigid molecules. Nevertheless, the worsening of prediction accuracy appears larger than expected in some case, for example for the rigid uracils. Upon closer investigation, this inaccuracy is due to large variations in geometries during the simulation, resulting in fluctuations in absolute intramolecular energies of over 1 kcal/mol for more than half of the SAMPL2 data set. Effects of this magnitude can introduce more noise than signal when compared to the absolute values of hydration free energies for the SAMPL2 dataset. The noise introduced by intramolecular energy terms during the MD might also impact on the solute–solvent interaction energy terms of the LIE method, since we have noted earlier that rigidifying the solute improves LIE predictions.

FiSH solvation models

As with the LIE models, we further investigated the effect of the charging method on FiSH continuum solvation model predictions (in Fig. 3 red bars for AM1BCC-SP charges, orange bars for AM1BCC-OPT charges and green bars for RESP charges). The hydration free energies for individual molecules for all solvation models are summarized in Table S6. Overall, the AM1BCC-SP (1.86 kcal/mol)

preformed best followed closely by AM1BCC-OPT (1.96 kcal/mol) and RESP charges (2.07 kcal/mol), with small differences in the slope and R^2 values (Table S7). On the explanatory set, the RESP partial charges performed better (1.69 kcal/mol) than AM1BCC-SP (2.29 kcal/mol) and AM1BCC-OPT (2.27 kcal/mol) charging sets due to a reduction in error for trimethylphosphate and pentachloronitrobenzene (see Table S6). On the obscure set, the AM1BCC-SP (1.71 kcal/mol) and AM1BCC-OPT (1.85 kcal/mol) partial charges performed better than the RESP charges (2.20 kcal/mol) due to the smaller errors in predicting the sugars. Parabens were predicted best with AM1BCC-OPT charges (1.62 kcal/mol), while the AM1BCC-SP (2.24 kcal/mol) and RESP (2.38 kcal/mol) charges performed similarly. The profens class was predicted well by both AM1BCC-OPT (0.49 kcal/mol) and RESP (0.60 kcal/mol) charges, while the for the uracils class AM1BCC-SP (1.12 kcal/mol) charge set outperformed the other charging methods.

Previously published studies have indicated significant changes in the accuracy of implicit solvation models depending on the charging method used [30, 58]. This differs from our results where all charge methods tested with the FiSH continuum model perform similarly across the SAMPL2 set. This insensitivity of the FiSH continuum model to the charging method is likely due to having different values for the parameters used to derive the Born radii that depend on the particular charging method employed. Since the induced surface charge density changes depending on the charging method used, it was necessary to have parameters for each charging method in order to mimic the explicit-solvent LIE model.

As stated previously [34], it is difficult to choose a charge set that is universal, since all charges perform well on some functional groups and poorly on others. For the FiSH continuum model, AM1BCC-SP charges will continue to be recommended since this set of charges performs best across most functional classes and the differences in errors between charge sets are relatively small over the entire set. For a complete list of functional class based MUEs for all FiSH models, see Table S8.

Overall, the addition of the difference in intramolecular energy between the gas and aqueous phases (in Fig. 3 blue bars, to be compared with the red bars) worsens the accuracy of the FiSH model on the SAMPL2 dataset (2.67 kcal/mol vs. 1.86 kcal/mol), There is only a minor change in prediction accuracy for the explanatory set (2.50 kcal/mol vs. 2.29 kcal/mol), but a more important increase in errors for the obscure set (1.71–2.73 kcal/mol), driven by a significant increase in errors for the sugars (0.23–13.99 kcal/mol). Small increases in accuracy in the obscure set are obtained for the uracils and the non-classified group (Fig. 3, Table S8). The FiSH model including

Fig. 3 Mean unsigned error (MUE, kcal/mol) of selected FiSH continuum solvation models. *Red* = AM1BCC-SP charges. *Orange* = AM1BCC-OPT charges. *Green* = RESP charges. *Blue* = AM1BCC-SP charges with internal energy (all terms). *Purple* = AM1BCC-SP with internal energy (excluding covalent terms). *Capped bars* represent errors greater than 6 kcal/mol
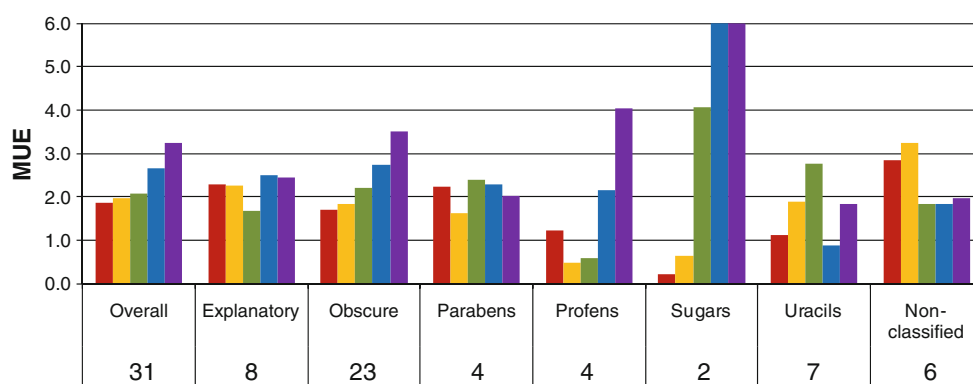
this simple formulation of intramolecular energy change (and a recalibrated cavity contribution) was also applied here to the previously analyzed traditional and SAMPL1 hydration datasets. The inclusion of this intramolecular energy difference resulted in little change in performance for these datasets (Table S9). Overall, the inclusion of only the change in non-bonded intramolecular terms (purple bars in Fig. 3) worsens the prediction accuracy of the FiSH continuum model relative to inclusion of all internal energy terms (blue bars in Fig. 3) for the overall SAMPL2 dataset (3.24 kcal/mol vs. 2.67 kcal/mol), as well as for the SAMPL1 dataset (3.92 kcal/mol vs. 2.19 kcal/mol). As in the case of the LIE solvation model, the increase in error associated with addition of the intramolecular terms is driven by the large errors in the prediction of sugars from the obscure set and glycerol from the explanatory set. When these 3 compounds are removed, the MUE for the remaining SAMPL2 set drops from 2.67 to 1.73 kcal/mol (with the inclusion of all intramolecular terms), which is a small improvement over the FiSH model without inclusion of intramolecular terms (1.93 kcal/mol after removing the sugars and glycerol). The large errors attributed to the sugars and glycerol may be related to the force-field used and suggests that a force-field parameterized for sugars such as GLYCAM [59] may be employed with more success.

Comparison to other continuum solvation models

Lastly, we would like to compare the prediction accuracy of the continuum models we previously used in the SAMPL1 challenge [32] with the LIE and FiSH models used in this SAMPL2 hydration dataset (in Fig. 4, red bars for the LIE solvation model with rigid solute, orange bars for the FiSH model, green bars for the RF model and blue bars for the CED model). The previous continuum models include the CED solvation model and the electrostatics-only RF model (see the Materials and Methods section for model parameters used). Overall, using the AM1BCC-SP partial charge set with all models, the LIE explicit-solvent model with rigid solute performed best (1.63 kcal/mol), followed by the RF model (1.79 kcal/mol) and the FiSH continuum model (1.86 kcal/mol), and lastly by the CED solvation model (3.46 kcal/mol). The LIE explicit-solvent model and the FiSH continuum model yield excellent correlations with reported hydration free energies (Fig. 5ab, see also Tables S4, S7). Even though the simple RF model continues to perform well overall, it is still not appropriate for some functional classes such as alkanes and hydrophobic molecules such as octafluorocyclobutane (4.40 kcal/mol) from the SAMPL2 explanatory subset, (Fig. 5c, see also Table S10) and to a certain degree for uracils (2.91 kcal/mol). The CED continuum model fails

Fig. 4 Mean unsigned error (MUE, kcal/mol) for various solvation models. *Red* = LIE explicit-solvent model with rigid solute and AM1BCC-SP charges. *Orange* = FiSH continuum model with AM1BCC-SP charges. *Green* = RF solvation model. *Blue* = CED solvation model. *Capped bars* represent errors greater than 6 kcal/mol
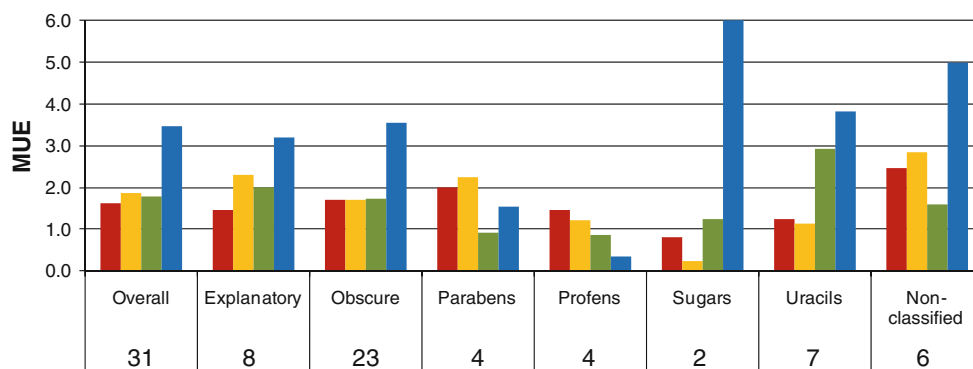
Fig. 5 Scatter plots showing the correlation between hydration free energies (kcal/mol) of the SAMPL2 set determined experimentally and predicted with various solvation models, including: **a** the LIE explicit-solvent model with rigid solute and AM1BCC-SP charges (submitted prospectively); **b** the FiSH continuum model with AM1BCC-SP charges (submitted prospectively); **c** the RF solvation model (submitted prospectively), and **d** the CED solvation model (not submitted prospectively).
Blue = Explanatory set; *Green* = Parabens; *Purple* = Profens; *Grey* = Sugars; *Orange* = Uracils; *Red* = Non-classified compounds
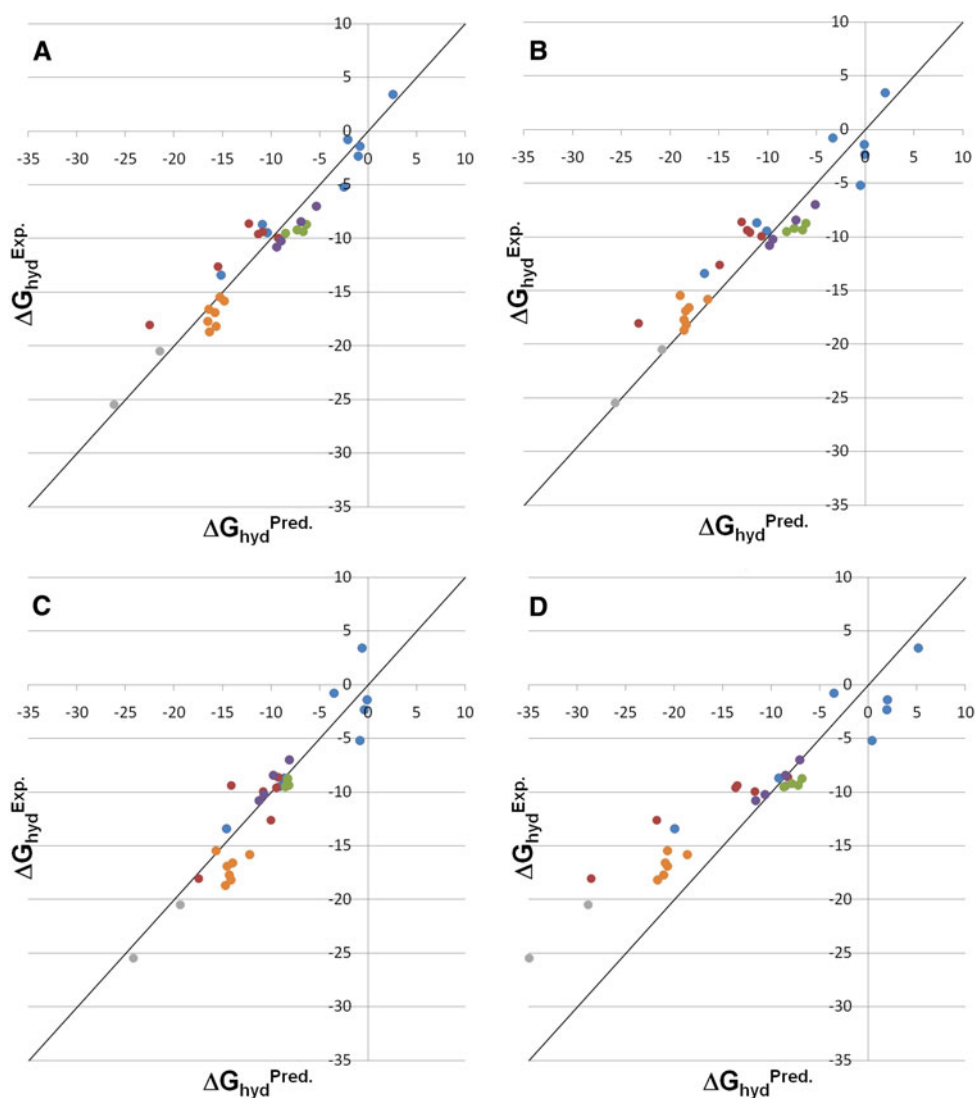
for most subsets and classes (Fig. 5d), yet it performs fairly well for parabens and profens (Fig. 4, see also Tables S10-S12). It is possible that the excessive calibration of continuum van de Waals parameters for the CED model on a given set of functional classes may not be appropriate to predictions on chemical groups that were not present in the training set (e.g., heteroaromatic systems such as uracil). The success of the FiSH model reinforces the core belief that developing a continuum model capable of capturing the physics of an explicit-water solvation model would enable better hydration free energy predictions, mirroring the results seen previously [34]. The FiSH and LIE model performed comparably on all sets and functional classes, showing that our FiSH continuum solvation model is a practical alternative to LIE models based on explicit-solvent MD simulations, thus enabling considerable time savings.

Conclusions

In this paper, using the SAMPL2 hydration dataset, we assessed the sensitivity of the LIE explicit-solvent model and the FiSH continuum model to partial charges and internal energy, and investigated their prediction accuracy comparatively to other continuum solvation models calibrated in our laboratory, the CED continuum model and the RF electrostatics-only solvation model.

In a previous study, we have shown that there is only a modest dependence on the solute partial charge set (between AM1BCC and RESP methods) for both the LIE explicit-solvent model [33] and the FiSH continuum model [34], with AM1BCC-SP charges affording the better performance. This marginal dependence on the charge set was maintained when the models were tested on the overall SAMPL2 hydration dataset, i.e., up to about 0.5 kcal/mol

variation in MUE among various charging methods for both solvation models was noted. For the LIE explicit-solvent models based on flexible solute geometries, the RESP partial charges provided best predictions on the overall SAMPL2 dataset, while the AM1BCC variants performed similarly. This advantage of RESP charges was not reproduced in the case of LIE models based on rigid solute geometries. For the FiSH continuum model, all investigated charge sets performed similarly on the overall SAMPL2 dataset. There was variability in the prediction accuracy for various individual functional classes depending on the charging method used for both the LIE and FiSH models. The AM1BCC-SP charges are still recommended for both models, since these partial charges have demonstrated competitive prediction accuracies on hydration free energies for various datasets covering a wide range of functional classes, and have the added benefit of portability for high-throughput virtual screening applications.

The effect of the solute flexibility was also investigated on the SAMPL2 hydration dataset, by either rigidifying the solute in the case of the LIE explicit-solvent model, or by adding terms to describe the change in solute intramolecular energy for both the LIE explicit-solvent model and the FiSH continuum model. The LIE explicit-solvent model proved slightly more accurate when the solute was held rigid during MD simulations. Since most molecules within the SAMPL2 set are rigid, it is possible that fluctuations around the ground state may have introduced more noise than signal and resulted in the lower accuracy of the flexible-solute-based LIE model. The same reasoning might explain the lack of improvement in the prediction accuracy of the LIE model upon incorporating the change in solute intramolecular energy between the gas and solution phases. We also explored a simple formulation to describe the change in intramolecular energy into the FiSH continuum model, by taking the energy difference between solute conformations energy-minimized with a distance-dependent dielectric constant and in vacuum. This model was met with limited success, where some of the more rigid molecules received slightly improved predictions and in the case of the sugars the prediction was dramatically degraded.

Overall, both the LIE explicit-solvent model and the FiSH continuum solvation models provided more accurate and transferable predictions of hydration free energies on the SAMPL2 set than the CED solvation model previously employed in the SAMPL1 blind challenge. Partial charges modestly affect the overall accuracy of these models, but accuracy does vary across functional classes. Further work is required in order to account for the solute conformational change between gas and water phases.

References

1. Gilson MK, Zhou HX (2007) Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct 36:21
2. Honig B, Sharp K, Yang AS (1993) Macroscopic models of aqueous solutions: biological and chemical applications. J Phys Chem 97(6):1101
3. Rashin AA (1993) Aspects of protein energetics and dynamics. Prog Biophys Mol Biol 60(2):73
4. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol 153(SUPPL. 1):S7
5. Mancera RL (2007) Molecular modeling of hydration in drug design. Curr Opin Drug Discov Dev 10(3):275
6. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3(11):935
7. Gallicchio E, Zhang LY, Levy RM (2002) The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. J Comput Chem 23(5):517
8. Chambers CC, Hawkins GD, Cramer CJ, Truhlar DG (1996) Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. J Phys Chem 100(40):16385
9. Kang YK, Némethy G, Scheraga HA (1987) Free energies of hydration of solute molecules. 2. Application of the hydration shell model to nonionic organic molecules. J Phys Chem 91(15):4109
10. Marten B, Kim K, Cortis C, Friesner RA, Murphy RB, Ringnalda MN, Sitkoff D, Honig B (1996) New model for calculation of solvation free energies: correction of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding effects. J Phys Chem 100(28):11775
11. Sitkoff D, Sharp KA, Honig B (1994) Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 98(7):1978
12. Tan C, Yang L, Luo R (2006) How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. J Phys Chem B 110(37):18680
13. Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319(6050):199
14. Reddi MR, Erion MD (2001) Free energy calculations in rational drug design. Kluwer/Plenum Publishers, New York
15. Aqvist J, Medina C, Samuelsson JE (1994) A new method for predicting binding affinity in computer-aided drug design. Protein Eng 7(3):385
16. Carlson HA, Jorgensen WL (1995) An extended linear response method for determining free energies of hydration. J Phys Chem 99(26):10667
17. Chen J, Brooks Iii CL, Khandogin J (2008) Recent advances in implicit solvent-based methods for biomolecular simulations. Curr Opin Struct Biol 18(2):140
18. Roux B, Simonson T (1999) Implicit solvent models. Biophys Chem 78(1–2):1
19. Simonson T (2001) Macromolecular electrostatics: continuum models and their growing pains. Curr Opin Struct Biol 11(2):243
20. Baker NA (2005) Improving implicit solvent simulations: a poisson-centric view. Curr Opin Struct Biol 15(2):137
21. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS (2008) Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. J Med Chem 51(4):769
22. Mobley DL, Bayly CI, Cooper MD, Dill KA (2009) Predictions of hydration free energies from all-atom molecular dynamics simulations. J Phys Chem B 113(14):4533

23. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA (2009) Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. J Chem Theory Comput 5(2):350

24. Mobley DL, Dill KA, Chodera JD (2008) Treating entropy and conformational changes in implicit solvent simulations of small molecules. J Phys Chem B 112(3):938

25. Nicholls A, Wlodek S, Grant JA (2009) The SAMP1 solvation challenge: further lessons regarding the pitfalls of parametrization. J Phys Chem B 113(14):4521

26. Li Z, Lazaridis T (2007) Water at biomolecular binding interfaces. Phys Chem Chem Phys 9(5):573

27. Scior T, Bernard P, Medina-Franco JL, Maggiora GM (2007) Large compound databases for structure-activity relationships studies in drug discovery. Mini-Rev Med Chem 7(8):851

28. Wang R, Fang X, Lu Y, Yang CY, Wang S (2005) The PDBbind database: methodologies and updates. J Med Chem 48(12):4111

29. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem 47(12):2977

30. Rizzo RC, Aynechi T, Case DA, Kuntz ID (2006) Estimation of absolute free energies of hydration using continuum methods: accuracy of partial charge models and optimization of nonpolar contributions. J Chem Theory Comput 2(1):128

31. Guthrie JP (2009) A blind challenge for computational solvation free energies: introduction and overview. J Phys Chem B 113(14):4501

32. Sulea T, Wanapun D, Dennis S, Purisima EO (2009) Prediction of SAMPL-1 hydration free energies using a continuum electrostatics-dispersion model. J Phys Chem B 113(14):4511

33. Sulea T, Corbeil CR, Purisima EO (2010) Rapid prediction of solvation free energy. 1. An extensive test of linear interaction Energy (LIE) for continuum model development. J Chem Theory Comput (in press)

34. Corbeil CR, Sulea T, Purisima EO (2010) Rapid Prediction of Solvation Free Energy. 2. The First-Shell Hydration (FiSH) Continuum Model. J Chem Theory Comput (in press)

35. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ (2010) The SAMPL2 blind prediction challenge: introduction and overview. J Comput Aided Mol Des. doi:10.1007/s10822-010-9350-8

36. Halgren TA (1999) MMFF VII. Characterization of MMFF94, MMFF94 s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. J Comput Chem 20(7):730

37. Jakalian A, Jack DB, Bayly CI (2002) Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem 23(16):1623

38. Jakalian A, Bush BL, Jack DB, Bayly CI (2000) Fast, efficient generation of high-quality atomic charges. AM1-BCC Model: I. Method. J Comput Chem 21(2):132

39. (2005) QuACPAC. 1.1.1. Sante Fe, NM: OpenEye Scientific Software, Inc

40. Case DA, Cheatham Iii TE, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. J Comput Chem 26(16):1668

41. Case DA, Darden TA, Cheatham TE III, CLS WangJ, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani

F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, Kollman PA (2006) AMBER. 9. University of California, San Francisco, San Francisco

42. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general Amber force field. J Comput Chem 25(9):1157

43. Floris FM, Tomasi J, Ahuir JLP (1991) Dispersion and repulsion contributions to the solvation energy: refinements to a simple computational model in the continuum approximation. J Comput Chem 12(7):784

44. Floris F, Tomasi J (1989) Evaluation of the dispersion contribution to the solvation energy. A simple computational model in the continuum approximation. J Comput Chem 10(5):616

45. Mobley DL, Barber Ii AE, Fennell CJ, Dill KA (2008) Charge asymmetries in hydration of polar solutes. J Phys Chem B 112(8):2405

46. Latimer WM, Pitzer KS, Slansky CM (1939) The free energy of hydration of gaseous ions, and the absolute potential of the normal calomel electrode. J Chem Phys 7(2):108

47. Rashin AA, Honig B (1985) Reevaluation of the Born model of ion hydration. J Phys Chem 89(26):5588

48. Roux B, Yu HA, Karplus M (1990) Molecular basis for the born model of ion solvation. J Phys Chem 94(11):4683

49. Babu CS, Lim C (1999) Theory of ionic hydration: insights from molecular dynamics simulations and experiment. J Phys Chem B 103(37):7958

50. Purisima EO, Sulea T (2009) Restoring charge asymmetry in continuum electrostatics calculations of hydration free energies. J Phys Chem B 113(24):8206

51. Purisima EO (1998) Fast summation boundary element method for calculating solvation free energies of macromolecules. J Comput Chem 19(13):1494

52. Purisima EO, Nilar SH (1995) A simple yet accurate boundary element method for continuum dielectric calculations. J Comput Chem 16(6):681

53. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S Jr, Weiner P (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 106(3):765

54. Huron MJ, Claverie P (1972) Calculation of the interaction energy of one molecule with its whole surrounding. I. Method and application to pure nonpolar compounds. J Phys Chem 76(15):2123

55. Zacharias M (2003) Continuum solvent modeling of nonpolar solvation: improvement by separating surface area dependent cavity and dispersion contributions. J Phys Chem A 107(16):3000

56. Levy RM, Zhang LY, Gallicchio E, Felts AK (2003) On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute-solvent interaction energy. J Am Chem Soc 125(31):9523

57. Tan C, Tan YH, Luo R (2007) Implicit nonpolar solvent models. J Phys Chem B 111(42):12263

58. Shivakumar D, Deng Y, Roux B (2009) Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. J Chem Theory Comput 5(4):919

59. Woods RJ, Dwek RA, Edge CJ, Fraser-Reid B (1995) Molecular mechanical and molecular dynamical simulations of glycoproteins and oligosaccharides. 1. GLYCAM_93 parameter development. J Phys Chem 99(11):3832