Testing the semi-explicit assembly model of aqueous solvation in the SAMPL4 challenge

Libo Li · Ken A. Dill · Christopher J. Fennell

Abstract Here, we test a method, called semi-explicit assembly (SEA), that computes the solvation free energies of molecules in water in the SAMPL4 blind test challenge. SEA was developed with the intention of being as accurate as explicit-solvent models, but much faster to compute. It is accurate because it uses pre-simulations of simple spheres in explicit solvent to obtain structural and thermodynamic quantities, and it is fast because it parses solute free energies into regionally additive quantities. SAMPL4 provided us the opportunity to make new tests of SEA. Our tests here lead us to the following conclusions: (1) The newest version, called Field-SEA, which gives improved predictions for highly charged ions, is shown here to perform as well as the earlier versions (dipolar and quadrupolar SEA) on this broad blind SAMPL4 test set. (2) We find that both the past and present SEA models give solvation free energies that are as accurate as TIP3P. (3) Using a new approach for force field parameter optimization, we developed improved hydroxyl parameters that ensure consistency with neat-solvent dielectric constants, and found that they led to improved solvation free energies for hydroxyl-containing compounds in SAMPL4. We also learned that these hydroxyl parameters are not just fixing solvent exposed oxygens in a general sense, and therefore do not improve predictions for carbonyl or carboxylic-acid groups. Other such functional groups will need their own independent optimizations for potential improvements. Overall, these tests in SAMPL4 indicate that SEA is an accurate, general and fast new approach to computing solvation free energies.

Introduction

In an effort to provide rapid and accurate estimates of hydration free energies for arbitrary solutes, we have recently developed a solvation method called semi-explicit assembly (SEA) [1, 2]. SEA is based on using an assembly approach, whereby we start with different small (atom-sized) spheres, and we simulate their solvation properties in some explicit-solvent model, such as TIP3P water [3]. We harvest these properties as a function of the van der Waals and electrostatic properties of the spheres, then at runtime for a given solute, we assemble appropriate collections of spheres and sum up the hydration free energy for the whole molecule (see Fig. 1).

We have previously found that the SEA model using dipolar semi-explicit water accurately reproduces the results from explicit-solvent free energy calculations [2, 4]. In cases where explicit solvent calculations give good predictions of experimental Henry's law constants, SEA gives similarly good estimates. In cases where explicit solvent calculations provide poor comparisons to experimental quantities, SEA usually does so too. This correspondence between SEA and explicit solvent

L. Li · K. A. Dill
Departments of Chemistry and Physics, Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA

C. J. Fennell (✉)
Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, USA
e-mail: cfennell@gmail.com; christopher.fennell@okstate.edu
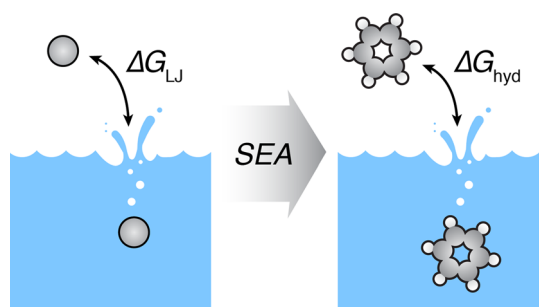
Fig. 1 The SEA model for water. We first calculate hydration free energies (ΔG_{LJ} in the figure) of simple spheres to build a map of the explicit solvent hydration about spherical objects. Using this map, we then assemble a full ΔG of hydration about more complex solutes

comes from the simple sphere calculations that form the foundation for the later assembly procedure. We found in the SAMPL3 event that the accuracy of SEA at reproducing experimental solvation thermodynamics is related to the quality of the force field used to describe the solute molecule [5]. When using the OPLS–AA force field [6] with the SAMPL3 solutes, SEA outperformed both SEA and explicit solvent calculations using GAFF [7] with AM1-BCC partial charges [8, 9]. We observed the relative performance of these force fields to be dependent upon the set of molecules studied, as agreement with experiment seemed to favor one force field combination or the other depending on the SAMPL event [5].

The present work extends our blind tests of the SEA model, within the SAMPL hydration prediction challenge [10–14]. We had three goals here. In addition to the dipolar SEA method we have reported on in the past (not submitted to SAMPL but reported here) we test the quadrupolar version of SEA to check for any measurable benefit in blind predictions. Recently, we have developed a new extension of SEA, called Field-SEA [15], that does away with the discrete surface water representations and replaces them with a adaptive field response term. This model improves predictions for highly charged and/or larger solutes, and we are interested in how it performs in prospective rather than retrospective predictions. Finally, we are interested in testing methods for improving force fields to obtain better agreement with experimental hydration free energies. We have extended some of our previous work in optimizing force field parameters using dielectric constant correction simulations [16] to the general representation of alcohol groups [17]. We evaluate these force field modifications for solutes containing hydroxyl groups (not submitted to SAMPL but reported here) as well as apply these same hydroxyl modifications to carbonyl and carboxylic acid groups to determine if they are general rules or if we need dedicated optimizations for these specific functional groups.

Summary of the Field-SEA method

While the original dipolar and quadrupolar versions of SEA have been described in detail elsewhere [2], the Field-SEA model is a recent development [15]. We provide here a short description of the method as applied in the SAMPL4 hydration prediction event.

The SEA method for predicting hydration free energies, like most implicit and explicit free energy calculations, is composed of two terms (Eq. 1): (1) a nonpolar term ΔG_{np} which captures the free energy cost of opening a cavity in the surrounding solvent and forming dispersion interactions between the solute and surrounding solvent, and (2) a polar term ΔG_{pol} accounting for the effect of electrostatic interactions in the solvation process,

$$\Delta G_{hyd} = \Delta G_{np} + \Delta G_{pol}. \tag{1}$$

As in the dipolar and quadrupolar versions of SEA, Field-SEA uses the standard ΔG_{np} described previously [1].

The primary distinguishing feature of Field-SEA is in the calculation of the ΔG_{pol} term. Here, rather than constructing discrete water representations in the form of free energy dipoles or quadrupoles, we use a procedure very similar to that used for the development of the ΔG_{np} term. We perform TIP3P water [3] explicit solvent free energy of charging calculations on a systematic series of partially and fully charged Lennard–Jones (LJ) spheres. The resulting ΔG_{pol} values are collected in a single contour as a function of both the curvature and the electric field at the solvent-accessible surface.

Applying this precomputed contour requires construction of an initial solvent-accessible surface about the solute molecule of interest. This is initially done using the average distance between solute atoms and the surrounding water (r_w) values extracted from the explicit solvent charging free energy calculations according to the process described for ΔG_{np}[1]. As the collective local electrostatic environment can have a strong effect on what constitutes the solvent-accessible boundary for explicit solvent about a solute molecule, we include a simple boundary refinement procedure to adjust this initial boundary to better correspond to how explicit solvent collectively interacts with the solute molecule. For example, regions with collectively strong electric fields can act to electrostrict surrounding water molecules and contract the surface locally. In this adaptive boundary process, we measure the collective field and curvature at patches on the initial surface and use these quantities to estimate new effective charge values for the immediate atom under each surface patch of interest. We then rebuild this patch of the surface using a lookup of the r_w information using the effective parameters for the local surface atom. To compute the ΔG_{pol}, we compute electric field and curvature values at all M exposed points on this

refined adaptive surface and look up ΔG_{pol} values for these regions. We arrive at an estimate of the total ΔG_{pol} by summing all M of these component terms,

$$\Delta G_{pol} = \sum_{i=1}^{M} a_i \cdot \Delta G_{pol}(\mathbf{E}_i, C_i), \qquad (2)$$

where a_i is the area of the surface patch represented as a unit-less ratio of the exposed area of this patch over the total area for a solute atom with curvature C_i. We approximate C_i using $1/r_{is}$, where r_{is} is the distance between the surface point and its associated solute atom. We calculate the electric field, $\mathbf{E}_i$, by summing the electrostatic contributions from the charges (q_j) on all N solute atoms,

$$\mathbf{E}_i = \frac{1}{4\pi\epsilon_0} \sum_{j=1}^{N} \frac{q_j}{r_{ij}^2} \hat{\mathbf{r}}_{ij}. \qquad (3)$$

Here, r_{ij} is the distance to the surface patch point and the jth solute atom, and $\hat{\mathbf{r}}_{ij}$ the unit vector between the two.

Field-SEA proves quantitatively equivalent to explicit solvent results for monatomic ions, and tests on an extensive variety of nonpolar, polar, and charged small molecule solutes show it often performs better than multipolar SEA in predicting ΔG_{hyd} from TIP3P calculations [15]. We further test its performance here in comparisons with experimental results for the 47 molecules in the hydration free energy challenge of SAMPL4 [18] as well as released explicit solvent results using both GAFF with AM1-BCC partial charges and a version using dielectric constant corrected hydroxyl groups (submissions 004 and 005 respectively) [14].

Hydroxyl force field modifications

Forcefields are parameterized based on matching simulation results to experiments on small-molecule compounds or to quantum-mechanical results. Obtained in the standard way, forcefield parameters are rarely good at giving accurate dielectric constants, when applied to neat liquid states. We recently developed a parameterization method that also ensures parameters are consistent with the known dielectric constants of liquids [16]. In this approach we iteratively optimize the solvent density, dielectric constant $\epsilon(0)$, and heat of vaporization ΔH_{vap}. We used that approach to give dielectric-corrected methanol parameters, starting from standard all-atom force fields [17], and we developed a general procedure for modifying the hydroxyl groups on a solute starting with the same base force field. The results for modification of hydroxyl oxygens in models using GAFF and AM1-BCC partial charges can be summarized by setting the σ_{LJ} and ϵ_{LJ} for the hydroxyl oxygen

to 3.21990 Å and 0.20207 kcal mol^{-1} respectively. These changes are coupled with scaling the hydroxyl hydrogen, oxygen, and bonded carbon (with its bonded hydrogens) partial charges by a factor of 1.20905. Such localized scaling will lead to a change the net charge of this group of atoms. This change is neutralized by scaling of the partial charge on the other atom(s) directly bonded to the hydroxyl carbon atom. In a separate study, we observed a general improvement of neat-liquid $\epsilon(0)$ values for a large set of alcohol molecules when using these dielectric correction (DC) modifications, and we additionally observed closer experimental agreement for hydration free energies calculated from these alcohols [17]. Note that when applying these modifications in hydration free energy calculations, one should apply a W_{pol} polarization work term of 0.409 kcal mol^{-1} per hydroxyl group to account for the energetic cost of polarizing the hydroxyl group beyond the AM1-BCC level [19, 20].

The blind tests afforded by SAMPL4 have been very useful here in testing those parameters. We tested these DC parameters with SEA in the SAMPL4 hydration challenge. It was helpful that the organizers also performed shared explicit solvent free energy calculations using these same DC parameters (submission 005). We show the results of these tests below; however, the hydroxyl DC parameter results were not the submitted results. Instead, we submitted an ambitious application of these DC hydroxyl parameters on carbonyl and carboxylic acid oxygen atoms in addition to the hydroxyl groups (submissions 180 and 181). This was done to test the general applicability of these DC parameters to hydrated oxygen atoms regardless of the functional group. From these results, we hoped to evaluate if independently optimized DC parameters are needed for each functional group of interest.

Simulation methods

All molecules were prepared for SEA hydration calculations starting with GAFF [7] and AM1-BCC [8, 9] parameter assignments using the Amber 11 version of Antechamber [21] starting from the mol2 structures provided by the SAMPL organizers. The resulting Amber coordinates and topologies were converted to GROMACS format with the ACPYPE python script [22] for SEA calculations. GAFF/AM1-BCC DC parameter versions of the 18 solute molecules containing hydroxyl groups were formed using an internally developed perl script. Additionally, the 26 molecules containing one or more hydroxyl, carbonyl, and carboxylic acid moieties were also modified using the same modified hydroxyl atom parameters to perform the general applicability test described above.

All the versions of SEA tested here perform hydration free energy estimations of static structures. While this is not an issue for the mostly rigid structures from SAMPL3 [13], many of the structures in the SAMPL4 challenge contain multiple rotatable bonds. To account for variation in ligand conformations, we report average values from 101 conformations for each ligand sampled from evenly spaced intervals of a 10 ns hydrated solute molecular dynamics trajectory. These calculation were performed following settings reported in previous work [5], with the exception of using GROMACS version 4.5.5 [23, 24].

Results and discussion

Both the earlier versions (dipolar and quadrupolar) and more recent version (Field-SEA) of the SEA model give the same high accuracy in computed hydration free energies compared to experiment that explicit-model MD simulations give. Table 1 and Fig. 2 show that dipolar, quadrupolar, and Field-SEA all predict with high accuracy the explicit-solvent ΔG_{hyd} values. The error bars on these quantities range from ±0.15–0.2 kcal mol^{-1}, and the deviations observed between methods are far below this level. The average unsigned error (AUE) is roughly $1kT$ to explicit solvent and the root-mean-square error (RMSE) is only marginally larger, indicating a lack of significant outliers getting penalized. The mean-signed error (MSE) is only slightly hydrophobic (~ 0.1 kcal mol^{-1}), indicating only a marginal under-solvation of the solutes. Together, these results on the 47 SAMPL4 solutes indicate that SEA predictions, regardless of the specific model, are nearly as accurate as the significantly more costly TIP3P calculations. These results are in agreement with our previous SAMPL results [5] and extensive comparative analyses [2, 15].

Table 1 also shows comparisons to experimental values. Here, all of the prediction errors have doubled relative to those seen in comparisons to TIP3P. Not shown are the RMSE/MSE/AUE of explicit solvent to experiment of 1.40/0.13/1.10. The MSE in these results show a ~ 0.1 kcal mol^{-1} hydrophobic offset relative to the

experimental values. This slightly hydrophobic shift, coupled with the ~ 0.1 kcal mol^{-1} under-solvation of SEA relative to explicit solvent described previously, indicates that further improvement could come from correcting this slightly hydrophobic bias which appears to be partly due to the force field used. Where there are errors in the
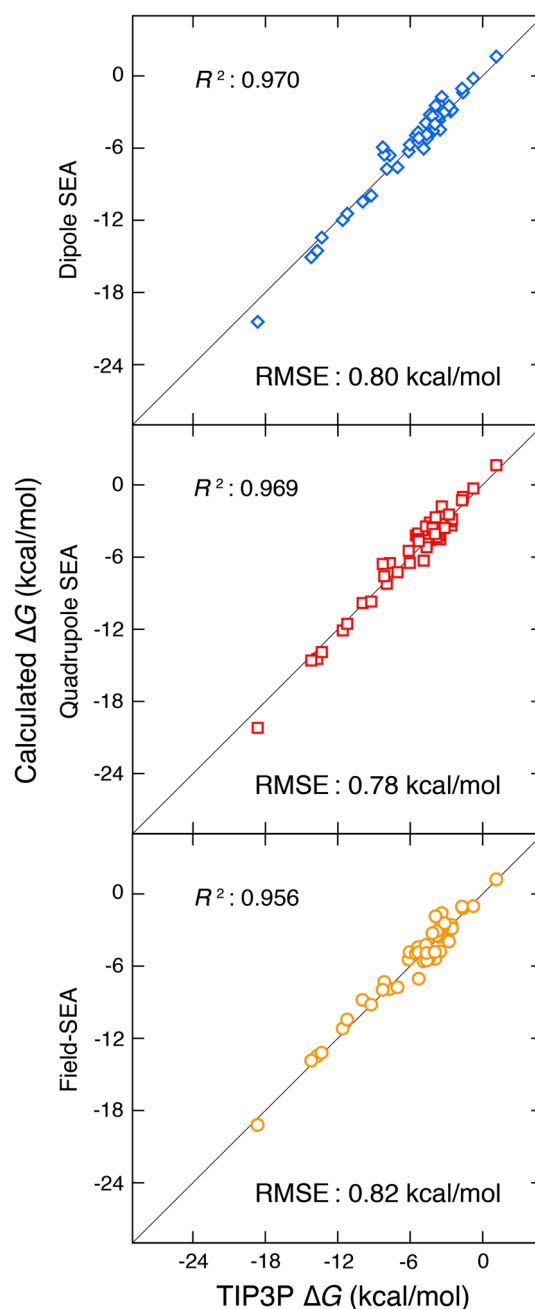
Fig. 2 Correlation plots of calculated ΔG_{hyd} from dipole SEA (*top*), quadrupole SEA (*middle*), and Field-SEA (*bottom*) versus TIP3P explicit solvent calculations using the GAFF/AM1-BCC parameters. These correlation plots show no significant outliers from one variant of SEA to the next, supporting their equivalence in predicting explicit solvent ΔG_{hyd} results

Table 1 Comparisons of ΔG_{hyd} prediction error to explicit solvent and experiment in kcal mol^{-1} for variants of SEA using GAFF/AM1-BCC parameters

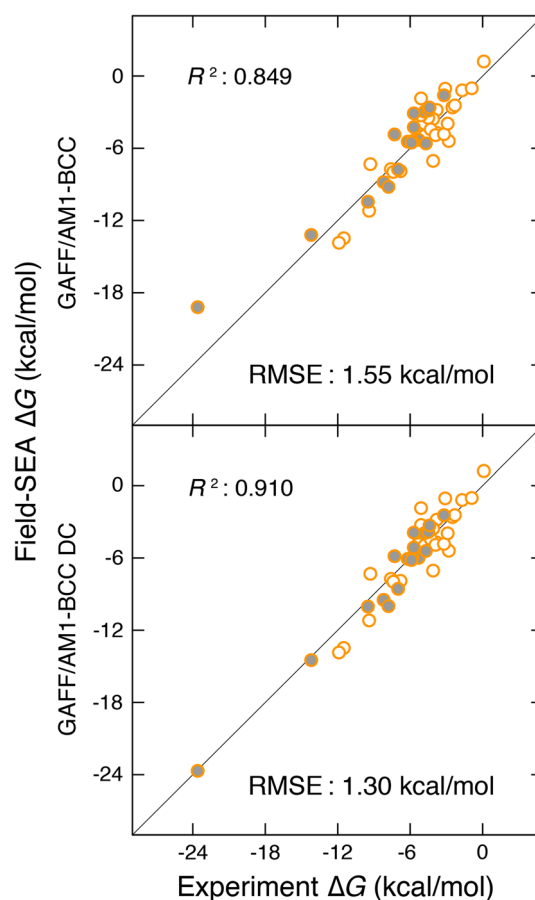|	Explicit solvent			Experiment		
	RMSE	MSE	AUE	RMSE	MSE	AUE
Dipole SEA	0.80	0.15	0.62	1.54	0.29	1.27
Quadrupole SEA	0.78	0.12	0.64	1.52	0.25	1.28
Field-SEA	0.82	0.11	0.67	1.55	0.24	1.22

forcefield/ explicit-solvent model, the same errors should largely appear in the SEA simulations, which are based upon them. In some instances, SEA appears to perform slightly better than the forcefield/ explicit solvent. These are probably mostly fortuitous.

New hydroxyl parameters, derived to be consistent with dielectric constants of neat liquids, improve the agreement with experiments for hydroxyl-containing compounds in the SAMPL4 experiments. Table 2 shows how the various versions of SEA compare with both TIP3P and experiment when using the hydroxyl DC parameters. There are two key points to draw from these results. First, agreement of SEA with TIP3P suffers somewhat when using the hydroxyl DC parameters, though this seems to be more of an issue for dipole and quadrupole SEA than Field-SEA. The reason for this is likely due to the increasing charge magnitude on hydroxyl hydrogen and oxygen atoms—the charge on many of the oxygen atoms is approaching -0.75. One deficiency of the multipolar SEA models is a loss in accuracy with absolute solute charges near one or higher, due to saturation of the discrete semi-explicit response. Because it lacks these discrete multipoles, Field-SEA does not share this deficiency, and we expect Field-SEA to perform increasingly well in more highly polarized environments.

The second key point is that agreement with experiment shows statistically significant improvement, particularly for Field-SEA. Figure 3 shows the performance of Field-SEA compared to experimental results before and after the application of these DC modifications. The 18 shaded points indicating solutes with hydroxyl groups show near uniformly improved correlation with experiment, and the RMSE for the whole set of molecules decreases by 0.25 kcal mol^{-1}. This improvement comes from elimination of a hydrophobic bias relative to experiment in the default hydroxyl parameters. This bias makes the MSE of the whole set in Table 1 slightly positive. Correcting the set of hydroxyl containing molecules with the DC parameters (the MSE for the set of 18 hydroxyl containing solutes decreased from 0.87 to -0.02 kcal mol^{-1}) uncovers a small hydrophilic bias in Table 2 due to the remaining 29 solutes. Though not shown, after application of these

hydroxyl DC parameters, explicit solvent calculations become one of the top predictors in most every statistical category. The results observed here indicate that this tested strategy for improving force field parameters via corrected dielectric behavior holds significant promise for improving general force field quality.

Dielectric-constant-corrected parameters are particular for different functional groups. While the DC modifications described above were specifically derived for alcohol groups, we decided to test if they could be applied indiscriminately to solvent-accessible oxygen atoms, specifically to the oxygen atoms of carbonyl and carboxylic acid groups. We found that they could not. The RMSE degrades to 1.56 kcal mol^{-1} for Field-SEA and 2 kcal mol^{-1} for the multipolar SEA models. Applying these hydroxyl group specific modifications to the carbonyl and carboxylic acid groups eliminates any benefit provided by the initial DC modification described above. The reason for this poor

Table 2 Comparisons of ΔG_{hyd} prediction error to explicit solvent and experiment in kcal mol^{-1} for variants of SEA using GAFF/AM1-BCC DC parameters

	Explicit solvent			Experiment		
	RMSE	MSE	AUE	RMSE	MSE	AUE
Dipole SEA	1.25	0.04	0.73	1.46	−0.19	1.09
Quadrupole SEA	1.28	−0.03	0.80	1.38	−0.26	1.08
Field-SEA	1.05	0.13	0.78	1.30	−0.10	1.02

Fig. 3 Correlation plots of calculated ΔG_{hyd} from Field-SEA versus experiment using the GAFF/AM1-BCC parameters (*top*) and GAFF/AM1-BCC DC parameters (*bottom*). The points with *gray centers* indication solutes with one or more hydroxyl moieties. By improving the dielectric properties of these groups, the hydration behavior falls more in line with experiment, good for a 0.25 kcal mol^{-1} RMSE improvement

performance is likely the need for a specific W_{pol} term for these new functional groups. Without it, the modifications lead to overly hydrophilic hydration of solutes containing these chemical groups. It is also possible that a specific dielectric optimization would not polarize these functional groups to the extent that methanol gets polarized.

Conclusions

We report tests of the family of SEA models in the SAMPL4 hydration prediction challenge. The key findings from this work are as follows: We find that quadrupolar SEA and our new Field-SEA perform as well as or better than the original dipolar SEA model in these blind predictions of solute hydration free energies. We observe that a simple approach for optimizing dielectric behavior of small molecule solute functional groups also simultaneously improves the hydration free energies for molecules containing these functional groups. Finally, while these functional group modifications are transferable to different solutes, we find them to not behave generally enough to apply to like-atoms in different functional groups. Those groups will need their own separately optimized parameters. Our principal conclusion is that the new Field-SEA method gives excellent agreement with experimental solvation free energies, in addition to excellent agreement with explicit-solvent simulations. And the Field-SEA method is faster, more robust, and more accurate than our original SEA models.

Reference

1. Fennell CJ, Kehoe C, Dill KA (2010) Oil/Water transfer is partly driven by molecular shape, not just size. J Am Chem Soc 132:234–240

2. Fennell CJ, Kehoe CW, Dill KA (2011) Modeling aqueous solvation with semi-explicit assembly. Proc Natl Acad Sci USA 108:3234–3239

3. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935

4. Fennell CJ, Dill KA (2011) Physical modeling of aqueous solvation. J Stat Phys 145: 209–226

5. Kehoe CW, Fennell CJ, Dill KA (2012) Testing the semi-explicit assembly solvation model in the SAMPL3 community blind test. J Comput Aided Mol Des 26:563–568

6. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236

7. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25:1157–1174

8. 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 21:132–146

9. 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 23:1623–1641

10. 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 51:769–779

11. Guthrie JP (2009) A blind challenge for computational solvation free energies: introduction and overview. J Phys Chem B 113:4501–4507

12. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ (2010) The SAMPL2 blind prediction challenge: introduction and overview. J Comput Aided Mol Des 24:259–279

13. Skillman AG (2012) SAMPL3: blinded prediction of host–guest binding affinities, hydration free energies, and trypsin inhibitors. J Comput Aided Mol Des 26:473–474

14. Mobley DL, Wymer KL, Lim NM (2014) Blind prediction of solvation free energies from the SAMPL4 challenge. ibid

15. Li L, Fennell CJ, Dill KA (2014) Field-SEA: a model for computing the solvation free energies of nonpolar, polar, and charged solutes in water. J Phys Chem B. doi:10.1021/jp4115139

16. Fennell CJ, Li L, Dill KA (2012) Simple liquid models with corrected dielectric constants. J Phys Chem B 116:6936–6944

17. Fennell CJ, Wymer KL, Mobley DL (2014) Dielectric corrected alcohol in condensed-phase fixed-charge simulations and its role in small molecule hydration. J Phys Chem B (in press)

18. Guthrie JP (2014) SAMPL4, a blind challenge for computational solvation free energies: the compounds considered. ibid

19. Swope WC, Horn HW, Rice JE (2010) Accounting for polarization cost when using fixed charge force fields. II. Method and application for computing effect of polarization cost on free energy of hydration. J Phys Chem B 114:8631–8645

20. Swope WC, Horn HW, Rice JE (2010) Accounting for polarization cost when using fixed charge force fields. I. Method for computing energy. J Phys Chem B 114:8621–8630

21. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25:247–260

22. Sousa da Silva AW, Vranken WF (2012) ACPYPE—AnteChamber PYthon Parser interfacE. BMC Res Notes 5:367

23. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. Comp Phys Comm 91:43–56

24. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447