Title

Adapting the semi-explicit assembly solvation model for estimating water-cyclohexane partitioning with the SAMPL5 molecules

Journal

Journal of Computer-Aided Molecular Design, 30(11)

ISSN

0928-2866

Authors

Brini, Emiliano
Paranahewage, S Shanaka
Fennell, Christopher J
et al.

Publication Date

2016-11-01

DOI

10.1007/s10822-016-9961-9

Peer reviewed

HHS Public Access

Author manuscript

J Comput Aided Mol Des. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

J Comput Aided Mol Des. 2016 November ; 30(11): 1067–1077. doi:10.1007/s10822-016-9961-9.

Adapting the semi-explicit assembly solvation model for estimating water-cyclohexane partitioning with the SAMPL5 molecules

Emiliano Brini

Laufer Center for Physical and Quantitative Biology, Stony Brook University Stony Brook, NY 11794, USA Tel.: 631 632 5420 emiliano.brini@stonybrook.edu

S. Shanaka Paranahewage and **Christopher J. Fennell**

Department of Chemistry, Oklahoma State University Stillwater, Oklahoma 74078, USA

Ken A. Dill

Laufer Center for Physical and Quantitative Biology, Stony Brook University Department of Chemistry, Stony Brook University Department of Physics and Astronomy, Stony Brook University Stony Brook, NY 11794, USA

Abstract

We describe here some tests we made in the SAMPL5 communal event of `Semi-Explicit Assembly' (SEA), a recent method for computing solvation free energies. We combined the prospective tests of SAMPL5 with followup retrospective calculations, to improve two technical aspects of the field variant of SEA. First, SEA uses an approximate analytical surface around the solute on which a water potential is computed. We have improved and simplified the mathematical model of that surface. Second, some of the solutes in SAMPL5 were large enough to need a way to treat solvating waters interacting with `buried atoms', i.e. interior atoms of the solute. We improved SEA with a buried-atom correction. We also compare SEA to Thermodynamic Integration molecular dynamics simulations, so that we can sort out force field errors.

Keywords

SAMPL; SEA; Solvation Free Energy; Partitioning; Distribution Coefficient

1 Introduction

We have recently developed a computational solvation model called Semi-Explicit Assembly (SEA) [1–3]. It parses a solvation free energy calculation into: (i) a presimulation step in which the solvation of a collection of model spheres is simulated by MD in an explicit model of water, then (ii) a runtime step that assembles the appropriate model spheres using a `regional additivity' relationship. Thus, SEA free energies are both fast to compute (at runtime), and yet give physical accuracies that about the same as much slower force field (FF) MD simulations for the same water model. SEA has been shown to accurately predict the solvation free energy of small molecules in water and electrolyte solutions [3–5]. In previous SAMPL competitions SEA showed an accuracy in blind predictions of the

solvation free energy of small solutes in water comparable to MD simulations (RMS error of 1.6 and 1.5 kcal/mol respectively) while being about 5 orders of magnitude faster [6–10].

SAMPL5 provided new challenges. First, rather than air-water partitioning of relatively small molecules, SAMPL5 sought the distribution coefficient values (log*D*) between water and cyclohexane (CYH) of 53 different drug-like compounds. In order to maintain this level of performance and accurately estimate log*D* values for these molecules, the various techniques need to accurately predict the solvation free energy of the compounds in both solvents. Second, the compounds of SAMPL5 are larger on average than those in previous SAMPL experiments,

Here, we made two types of tests. First, we tested the existing SEA in SAMPL5 in its canonical prospective (i.e. blind-test) mode. Again, we find SEA to give comparable results with molecular dynamics modeling, but much faster. Second, after SAMPL5 ended, we studied the successes and failures, and refined the SEA-tables, treated multiple solute and solvent representations, and developed a more systematic treatment of the large-molecule corrections. In short, SAMPL5 has been of great value to us for improving this solvation model.

2 Theory

Here, we give a general description of SEA. Then, we describe our approach to handling non-surface atoms in the solute molecules.

SEA estimates the solvation free energy of a molecule using a series of additive terms pulled from pre-computed free energy contours. Each additive term describes how the solvent interacts with a small region of the solute. As common in many approaches, the process of solvation of a solute from vacuum to bulk solvent is described as a two step process. First a dummy molecule, with appropriate Lennard-Jones (LJ) σ and ϵ parameters but no charges, is grown in the bulk solution from vacuum. Second, the partial charges of the atoms are turned on while the molecule is surrounded by bulk solvent. The non-polar (*NP*) solvation free energy of a molecule is the reversible work associated with the first step, while the polar (*P*) solvation free energy of a molecule is the reversible work associated with the second step. The sum of the two is the (total) solvation free energy of a molecule [11–18]. In SEA the solvation free energy of a molecule (Δ*G*) is calculated as a sum of additive terms,

$$\Delta G = \sum_{i}^{\text{atoms}} \Delta G_i^{NP}(\sigma,\epsilon)\frac{A_i}{A_i^0} + \sum_{j}^{\text{surf. dots}} \Delta G_j^{P}(C,E)\frac{A_j}{A_j^0}. \quad (1)$$

On the right side of Eq.1 the first term represents the *NP* contribution to the solvation free energy of the solute, and the second term is the *P* contribution.

The *NP* solvation free energy is calculated as a sum of atom-wise contributions. The contribution of each atom is proportional to its solvation free energy when packed in its molecular environment, treated as that of an isolated atom with local environment perturbed

LJ parameters ($\Delta G_i^{NP}(\sigma,\epsilon)$). The LJ parameters used come from a LJ function fit to the collective LJ interaction field about the atom of interest. This contribution is weighted according to the ratio of the exposed solvent accessible surface (SAS) when the atom is part of the molecule (A_i) and when it is isolated (A_i^0). The term $\frac{\Delta G_i^{NP}}{A_i^0}$ is interpolated from the *NP* free energy contour and can be thought of as a surface tension for solvent at that particular region on the surface of the molecule [1,2].

The second term of eq.1 represents the *P* solvation free energy of a solute. The interaction between charges is long range in nature, making it difficult to define a purely local-environment *per atom contribution* to the *P* solvation free energy. Still the *P* solvation free energy can be computed as a sum of *semi-local* contributions. Each region is defined as a segment of the SAS with a given curvature (*C*, which is equal to 1/*R* where *R* is the SAS radius) that experiences a given electric field (*E*) generated by all the atoms of the solute. The contribution of each region is proportional to the solvation free energy of a charged sphere that creates a SAS with the same curvature and electric field as the molecular region ($\Delta G_j^P(C,E)$). As before we need to weight this contribution with an area term that describes the size of the SAS of the region (A_j) compared to the one of the isolated test sphere (A_j^0) [3].

It is important to note that, since charges attract water molecules, the SAS of the *NP* dummy molecule and of the *P* molecule are different. They need to be computed separately. Also the *P*-SAS curvature and electric field are not strictly independent of one another, so the *P*-SAS needs to be converged upon following an iterative process.

What makes SEA a particularly fast method to compute the solvation free energy is the possibility to pre-compute the *NP* and *P* solvation free energies for a set of ideal spheres with systematically varied LJ parameters and charges(see Sec. 3.1 for grid details). This expensive computation needs to be done only once for a given solvent at a given state point. At run-time it is then possible to combine this information, stored in look-up tables or as nonlinear-fit functions, to estimate the solvation free energy of the solute in a fraction of the time [4,5].

Because of the size of the solute molecules of the SAMPL5 competition, we developed a new way to deal with the contribution of buried-atoms to the *NP* solvation free energy of a molecule. Why we need to do this is illustrated in the 2D cartoon of Fig. 2. Both molecules in the figure have the same surface area and displace the same volume of water. If we consider only the contributions of the surface atoms, the two molecules will have the same *NP* solvation free energy, since in both cases only the gray atoms are on the surface of the molecules. To distinguish between molecules A and B it is necessary to consider a term that accounts for the (attractive) van der Waals interaction between the red atom of the solute B and the solvent. In other words we need to account for the conditional solvation free energy of the red atom [16,19,20]. It should be noted that support for such occluded volume elements was considered in the original *NP* SEA term, though as the water boundary is close to the solute and the previously investigated solutes tended to be mostly solvent exposed, so

such buried contributions were typically negligible [1]. In SAMPL5 both the cyclohexane boundary is further out and the target solute molecules are larger on average than those previously considered, meaning that buried contributions are increasingly important for such solvation calculations.

The contribution of a buried-atom to the *NP* solvation free energy of a small molecule primarily depends on two factors: 1) the atom LJ parameters and 2) the chemical nature of the solvent. In molecules with buried-atoms, like the ones commonly seen in the SAMPL5 set, an atom is most likely buried by atoms it has a chemical link with (i.e. the central carbon of a neopentane molecule). We set such contribution ($\Delta G^{NP,ba}$) to be equal to

$$\Delta G^{NP,ba} = \sum_{i}^{\text{buried atoms}} \frac{V_i}{V_{CYH}} \Delta G^{NP}_{CYH}, \quad (2)$$

where ΔG^{NP}_{CYH} is an estimate of the *NP* solvation free energy of cyclohexane in the solvent, V_i is the volume of the void buried-atom cavity, and V_{CYH} is the molecular volume of CYH. The assumptions on which this simple correction relies holds for small/medium size molecules with few buried-atoms, like the ones of SAMPL5. These assumptions probably do not hold when dealing with larger molecules with more packed 3D structures, like folded proteins. For such cases, this simple correction will likely need to be substituted with a more general treatment. Note that in the *P* solvation free energy, buried atoms are already fully accounted for *via* the contribution of their partial charges to the electric field at the SAS.

3 Computational Methods

The SEA contours we used in this work are based on four different solvent model. For water we alternately use TIP3P and H2O-DC [21,22]. For CYH we use a dielectrically corrected united atom version (CYH-DC) and an all atom version based on the GAFF using AM1-BCC partial charges (CYH-AA) [23, 24]. For a discussion about characteristics and differences of these solvent FF we refer to a related study in this issue by Paranahewage *et al.* [25]. We compare results from these options in the retrospective analyses. The prospective SAMPL5 submissions used the default TIP3P water solvent contours and a CYH *P* contour built by scaling the TIP3P nonlinear fit function to a roughly 15% of its full magnitude. This value was chosen in an attempt to have the existing function pass through a limited set of sphere free energy calculations in CYH-DC solvent. For the retrospective analyses, new nonlinear fit functions for the *P* contour were extracted from an extensive set of sphere calculations in each of the 4 solvents as described below.

3.1 SEA Contour Determination

In order to build the SEA look-up tables for these solvent models, we must pre-compute the solvation free energies for a set of uncharged and charged spheres. For the uncharged spheres, free energy calculations were performed on spheres with σ values of 0.6 Å to 7.0 Å with a 0.8 Å step interval and ϵ values of 0.015625, 0.0625, 0.25, and 1.0 kcal/mol. This is a set of 36 calculations in each solvent and is thus a subset of the original *NP* free energy

contour calculated for TIP3P [1]. These subset contours were smooth and quite similar in curvature to the more detailed TIP3P *NP* contour, so more detailed *NP* contours were constructed for each solvent by linearly scaling the TIP3P *NP* contour points such that the total contour smoothly passed through the 36 points. Similarly, the *NP* solvent distance (*R*) files for constructing the SAS boundaries were built from these 36 spheres by using the first peak of the g(r) between the solute and the center of geometry of the solvent molecules in solvated sphere calculations.

For the charged spheres, free energy calculations were performed on spheres with sizes spanning σ 2.0 Å to 7.0 Å with a 0.2 Å step interval and an ϵ value of 0.125 kcal/mol. The sphere charges values spanned −2.0 to 2.0 in steps of 0.1. Thus, 1040 sphere charging free energy calculations were performed in each solvent. Nonlinear fits to the field SEA *P* contour function were performed in Mathematica 10 [26] to develop new *P* free energy functions [3]. The data set used here for each solvent is significantly more extensive than in the original field SEA work, providing an opportunity to potentially simplify the *P* contour function. To this end, for positively and negatively charged spheres we performed both single function *uniform* nonlinear fits and nine function *piece-wise* nonlinear fits spanning select charge value ranges as performed in the original field SEA study [3]. We found that for both the uniform and piecewise fits, we were able to simplify the *P* contour function to

$$\Delta G_j^P(C,E) = B - E^2 / \left(A_0 + A_1 C + A_2 C^2\right), \quad (3)$$

for the positively charged sphere solutes and

$$\Delta G_j^P(C,E) = B - E^2 / \left(A_0 + A_1 C + A_2 C^2 + A_3 C^3\right), \quad (4)$$

for the negatively charged sphere solutes, often with the *B* value being 0. As stated previously, *E* is the electric field and *C* is the curvature at a given SAS point. The *B* and A_n values are simply nonlinear fit parameters. Note that these functions are not truly necessary as one could simply interpolate over the *P* free energy points. The benefit of using equations 3 and 4 is computational performance.

3.2 Thermodynamic Integration Calculations

The free energies associated with the solvation of the test spheres were computed using thermodynamic integration (TI). For each *NP* TI calculation, λ steps of (0.0 0.05 0.1 0.2 0.3 0.4 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1.0) were used, and soft core potentials where employed[27]. Each *P* TI calculation used 6 λ steps evenly distributed from 0.0 to 1.0. The simulations were performed using version 5.0.4 of the GROMACS package [28–32]. The temperature was held constant at 298.15 K with Langevin dynamics [33] with an inverse friction coefficient of 2 ps, and the pressure was set to 1 atm using the Parrinello-Rahman barostat [34]. Following 300 ps of equilibration, each TI window was sampled for 5 ns using a 2 fs timestep for integrating the equations of motion with the leap-frog algorithm. Lennard-Jones interaction where computed using a shifted cutoff at 1.2 nm, and energy and

pressure long-range dispersion corrections were applied. Interaction between charges where computed using PME [35] with 0.12 grid spacing and a real space cut-off of 1.2 nm.

3.3 Buried-Atom Correction

Considering that the correction of the *NP* solvation free energy associated with the presence of buried-atoms highlighted in of eq. 2 is going to be a fraction of the *NP* solvation free energy of a molecule; we decided to apply this correction only to CYH solvation. During SAMPL we used an estimate of the experimental value of the solvation free energy of CYH in CYH (−4*kcal/mol*) and an experimental value of its volume (0.1724 nm^3)[36]. Retrospectively we used the *NP* solvation free energy of CYH in CYH as calculated using TI and the appropriate FF. In these cases the solvation free energy are equal to −5.30 kcal/mol for CYH-DC and to −4.57 kcal/mol for CYH-AA. The molecular volumes are 0.185 nm^3 and 0.1875 nm^3 respectively. Those value were obtained using TI calculation similar to the one of the test spheres. In each case, the buried-atom volume was calculated by subtracting the volume of the solvent accessible atoms from the total volume of the solute, with both of these volumes calculated numerically using the double cubic lattice method [37].

3.4 SEA Solvation Free Energy Calculations

The FF used to represent the 53 SAMPL solute are GAFF and a scaled version of GAFF (GAFF-scaled) where atomic charges were scaled 20% and sigma values were linearly scaled with changes in atom charge magnitudes in order to compensate for the change in solute volume. SEA is agnostic with respect to the FF used to describe the solute. Therefore we often comment the results of the two different solute FF together. We would therefore refer to the “106 SAMPL molecules” meaning the 53 SAMPL solute described according to the two different FF. SEA predictions were calculated on solute centroid structures obtained by clustering the fully coupled state trajectories of Paranahewage *et al.* [25].

The *NP* solvation free energies predicted by SEA were the average of 10 independent calculations on each solute molecule, where the *P* solvation free energies were averaged over 50 independent calculations. Each independent calculation used a different randomized dot surface. These molecular dot surfaces are composed of non-overlapping atomic dot surfaces, each with approximately 180 dots at a distance r_w from the atom center, where r_w is the radius of first hydration shell interpolated from the model sphere pre-computations. Inaccessible points due to neighboring atoms were culled out.[1] Additionally the position of the P SAS was iteratively refined 3 times each calculation.[3] log *P* was calculated using the following equation

$$log\ P = \frac{\Delta G_{\mathrm{water}} - \Delta G_{\mathrm{CYH}}}{ln\,(10)\ k_{\mathrm{B}}T}, \quad (5)$$

where Δ*G* is the solvation free energy of the solute in the labeled solvent, ln(10) converts from base *e* to base 10 logarithm, k_B is the Boltzmann constant, and *T* is the temperature (i.e. 300 K). No specific considerations of the tautomeric, protonation, or dimerization states

of the solute molecules beyond those provided by the organizers were considered, so the estimated log P values were reported as log$D$ values in experimental comparisons, though we refer to them as log P when comparing between calculations.

4 Results and Discussion

The results below include both prospective and retrospective analyses. The prospective SEA results were those submitted to the SAMPL5 experiment using a preliminary set of free energy contours and an initial estimate of the buried-atom correction. The retrospective SEA results involve both analysis on how the newly introduced corrections affect the quality of our prediction and explorations for further improvement. Since the quality of SEA predictions of experimental equilibria is strongly correlated with the ability of the FF to properly describe solvation process, most of the retrospective comparisons are between SEA and explicit solvent TI calculations in an effort to identify potential improvements in the employed techniques.

4.1 SAMPL5 submission shows that the chosen FF matters greatly for accurate predictions

The SEA SAMPL5 submissions mark it as a middle performer of the participants in comparisons with experiment. With our expected better submission showing a mean signed error (MSE), root mean square error (RMSE) and average unsigned error (AUE) of 3.04, 4.9, and 3.8 log$D$ units for the complete set of 53 molecules, it sits almost exactly in the middle of complete submissions for the RMSE and AUE metrics, being approximately 1.5 and 1 log$D$ units off of the best performing technique. As the computational cost for making specific predictions is quite minimal with SEA, this is an encouraging outcome.

Considering the individual molecule predictions, it becomes clear why the SEA GAFF-scaled submission was a significant improvement over the GAFF submission. The scatter-plot on the left side of Fig. 3 shows the correlation between the results obtained with SEA and TI using the two different FF options. While the orange points are balanced such that they are clustered near log $P = 0$, the blue points are skewed to favor positive log P values in the upper right quadrant, indicating that both SEA and TI favor solvation in the cyclohexane phase. This is further supported by the comparisons of SEA predictions for all 53 SAMPL5 molecules with experiment shown on the right side of Fig. 3. The predictions with GAFF (blue bars) are systematically shifted to positive values while the GAFF-scaled predictions are more evenly balanced, in better agreement with the experimental line. This indicates how critical the solute and solvent FF choices are if one expects to obtain accurate predictions with a generalized physical model like SEA. The GAFF-scaled were introduced in an effort to provide a better condensed-phase depiction of a solute than standard GAFF with AM1-BCC partial charges, and the GAFF-scaled FF results significantly outperform GAFF in these SEA calculations. However, this outcome is highly influenced by the approximate treatment of the cyclohexane phase in these predictions, and it appears that the approximate cyclohexane phase used here is slightly too polar and overly stabilizes all solutes.

Fig. 4 shows how the P and NP components of the submitted SEA results compare against the more computationally expensive TI calculations based on the same FF combinations. The first row of graphs shows the comparison between NP solvation free energy of the 53

SAMPL5 molecules. The second row shows the same comparison for the *P* component. The first column of graphs is relative to calculation in water, the second is relative to calculations in cyclohexane (CYH-DC for TI and the approximate CYH-DC for SEA). Orange points are relative to calculations carried using GAFF and AM1-BCC to represent the solute molecule, while the blue points represent calculation using the scaled version of GAFF discussed in section 3. In all the graphs the vertical axis is relative to the solvation free energy value predicted by SEA and the horizontal one is relative to the TI calculation. As the blue and orange points are similarly well correlated, the observed agreement between SEA and TI appears independent of the FF-solvent combination used. This supports the assertion that SEA is agnostic with respect to the FF used to describe the solute. The agreement between SEA and TI prediction is reasonable in the four graphs. The *NP* SEA calculation in water show a systematically lower slope. This could potentially be corrected by including the buried atoms effect in the hydration calculation. The differences between *P* calculations are mostly due to the quality of the table used to compute the *P* contribution to the solvation free energy with SEA. As the SEA *P* calculations report lower free energies than TI, it appears that the approximate CYH-DC contour is too polar to properly describe charging in the explicit CYH-DC environment. Additionally, a diagonal separation between the GAFF and GAFF-scaled results in the water *P* calculations highlights the fact that the default GAFF FF is less polarized than the GAFF-scaled FF, and this combination of too polar CYH and low polarity solute lead GAFF to perform poorly in this set of calculations. We discuss more about these points in the sub-section 4.2.

4.2 More accurate contours and explicit buried-atom corrections improve the predictive accuracy of SEA

To test the possible points of improvement for SEA predictions discussed above, we decided to craft an explicit TI based treatment for the buried-atom volume correction as well as highly detailed *P* contours for all solvents of interest. These are retrospective analyses and are used to indicate the potential of the SEA solvation technique given more fully developed pre-computations, as well as identify possible strategies for future improvements.

Cyclohexane is a larger and more nonpolar solvent than water. In order to describe such a solvent with the SEA approach, we needed to reconsider how the solvent molecules arrange and pack around the solute in the first solvation shell. There are two primary options: we can consider the first solvation shell as described 1) by the position of the center of mass of neighboring solvent molecules or 2) by the position of the closest heavy atom of the neighboring solvent (see the picture of Fig. 5). The “closest atom” choice would place an emphasis on the solute since it defines the cavity where the solute resides and would be somewhat like a solvent excluded surface (SES) in contrast to the SAS normally utilized in SEA. The “COM choice” puts the more consideration on possible orientational states of the surrounding solvent, and it is consistent with the process already used in forming the water SAS. Additionally, use of the COM is consistent with coarse-graining processes where relevant states are united into an averaged representation of a solvent molecule, a process related to that done here. In Fig. 5 we can see how the two choices affect the prediction of the *NP* solvation free energy for 106 SAMPL5 solute points of comparison. Using the COM solvent boundary is noticeably more well correlated with TI calculations.

A consideration that became important in this SAMPL competition was the treatment of buried-atoms. Of the 53 solute molecules, only 15 do not have buried-atoms. In section 2 we highlighted a simple correction that estimates the effect that buried-atoms have on the solvation free energy of a molecule. The effect of such correction is presented in the two plots of Fig. 6. Both plots consider the *NP* component of solvation in cyclohexane, just using an CYH-AA or CYH-DC as the explicit model. Including this correction does improve the SEA correlation with TI in the case of the CYH-DC solvent. The correction shifts the points in the right direction for CYH-AA solvent, but it seems to overestimate the contribution of buried atoms to the solvation free energy. The correction applied is a volume contribution to the solvation free energy equal to a volume contribution of solvation CYH in CYH. As mentioned in section 3, during SAMPL we used an experimental estimate of this free energy contribution and the experimental density of CYH. An alternative is to use the *NP* solvation free energy of CYH in CYH obtained from simulation using the same solvent FF parameters. Fig. 6 shows how these two specific options alter the correlations, with the light blue points using the experimentally based correction and the dark blue dots using the term derived from explicit simulations. This more consistent description of the CYH *NP* solvation free energy improves the prediction accuracy, in particular for the CYH-DC model. CYH-AA seems to be somewhat over-corrected. In both cases, this correction is somewhat empirical, and it would be better to have a more general accounting of complex buried-atom effects on the *NP* solvation term. We are currently working on a more rigorous treatment of the contribution of buried-atoms that give a properly weighted accounting based on how deep an atom is buried in the molecule.

For the SAMPL competition we did not introduce any algorithmic changes in the way the *P* solvation free energy is calculated. However, for each new solvent, a new *P* free energy contour and water boundary table needs to be pre-computed. Our prospective effort used an estimated table for CYH. Retrospectively we decided to investigate the possible benefit from using a more systematically and rigorously detailed set of fits for the solvents.

SEA ability to effectively predict the *P* solvation free energy of a solute is linked to the quality of the fit function that describes how the solvation free energy of a solute-region changes with the electric field and the curvature of the SAS. During SAMPL5 we followed the approach described in *Li et al.* [3] to build the P functions for H2O-DC, and the two CHY models. Retrospectively we decided to investigate if we could improve our approach function. To do so we developed a new set of calculations of solvation free energy that covers more uniformly the curvature - electric field space of the probes. Thanks to that we were able to proceed to a more systematic fitting using Eq. 3 and 4. In particular we were able to obtain single fitting functions that were covering the whole curvature-charge space. Also the new fit predict no contribution to the *P* solvation free energy coming from molecular-region where the electric field is equal to 0. This allowed us to remove smoothing functions employed to enforce that. In Fig. 7 is shown a comparison between the approach used in SAMPL5 (blue) and this new “cleaner” one (orange). The result are bench-marked versus TI calculation results. Following this new approach to deriving *P* solvation tables we can decrease the computational load, obtain more rigorous fit, and at the same time maintain a good accuracy of the prediction.

Fig. 8 highlights the effect that the introduced improvement have on the SEA predictions. The orange dots represent SEA calculation performed as we did in SAMPL5, the red ones represent SEA calculation computed retrospectively both with the field surface and the TI volume correction. TI predictions are also reported in blue. The gray line represents perfect agreement between the predictions and experimental results. The dashed lines divide the graphs in the four quadrant. Points laying in the first and third quadrant agree at least qualitatively on the propensity of a solute for one solvent or the other. Points in the other quadrants are "false positive/negative" predictions. It is important to notice that an error of $k_B T$ in the estimates of the difference in solvation free energy of a solute in the two solvents propagates in Eq. 5 into an error of 0.61 log unit in the prediction of the partition coefficient. We can see how the red points are less scattered and more in agreement with experimental results. The SEA results are slightly more scattered than TI as they are performed on single clustered conformations of the solutes, but they fall within the same prediction envelope as explicit solvent calculations for the selected solute and solvent FF. Total retrospective prediction RMSE, MSE, and AUE improved of about 1 log unit for every force field/solvent combination. These results are a significant improvement over the original SAMPL5 submissions, bringing them in line with explicit solvent log$D$ predictions.

5 Conclusion

We describe here our tests and improvements of the field variant of a computational model of solvation free energies called SEA. SEA is a method that mimics explicit solvent by first pre-simulating a set of component spheres in a given explicit solvent model. In our case, that explicit model is TIP3P. Then at runtime, Field SEA uses very fast additivity relations to compute the free energy of solvation of an arbitrary solute. We use the GAFF force field and a variant of it we call GAFF-scaled. We tested SEA prospectively (in the blind SAMPL5 event) and then made improvements retrospectively (after the event). We reach three conclusions here. First, we describe a simpler and better polynomial function that represents the field surface used by SEA. Second, the solute molecules in SAMPL5 are bigger and more complex than the solutes in SAMPL3 and SAMPL4. We found that a simple way to handle buried atoms improves predictions relative to the earlier version of Field SEA. And, third, as we have found before in both retrospective tests and in the prospective tests of SAMPL3 and SAMPL4 [3–5] SEA gives predictions of comparable accuracy to the underlying force field and explicit-solvent model from which it is derived. The advantage is that SEA is much faster. The additivity relations that give SEA its much greater speed seem to hold also with big molecules like the ones of SAMPL5. The primary current sources of error are likely, as discussed above, the quality of the free energy contours used and the approximate correction of the effect that the buried atoms have on the nonpolar solvation free energy. Also it is important to keep in mind that the quality of the force fields and of the explicit solvent models impacts strongly the ability of SEA to accurately reproduce experimental data, and improved molecular mechanics representations of explicit solutes and environments should result in improved predictive accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors appreciate the support from National Institutes of Health Grant GM063592 and GM107104. Portions of the computing for this project were performed at the Laufer Center, XSEDE allocation CHE150012 to CJF, and the OSU High Performance Computing Center at Oklahoma State University supported in part through the National Science Foundation grant OCI1126330.

References

1. Fennell CJ, Kehoe C, Dill KA. Oil/Water transfer is partly driven by molecular shape, not just size. J. Am. Chem. Soc. 2010; 132:234–240. [PubMed: 19961159]
2. Fennell CJ, Kehoe CW, Dill KA. Modeling aqueous solvation with semi-explicit assembly. Proc. Natl. Acad. Sci. USA. 2011; 108:3234–3239. [PubMed: 21300905]
3. Li L, Fennell CJ, Dill KA. Field-SEA: a model for computing the solvation free energies of nonpolar, polar, and charged solutes in water. J. Phys. Chem. B. 2014; 118:6431–6437. [PubMed: 24299013]
4. Kehoe CW, Fennell CJ, Dill KA. Testing the semi-explicit assembly solvation model in the SAMPL3 community blind test. J. Comput. Aided Mol. Des. 2012; 26:563–568. [PubMed: 22205387]
5. Li L, Dill KA, Fennell CJ. Testing the semi-explicit assembly model of aqueous solvation in the SAMPL4 challenge. J. Comput. Aided Mol. Des. 2014; 28:259. [PubMed: 24474161]
6. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. J. Med. Chem. 2008; 51:769–779. [PubMed: 18215013]
7. Guthrie JP. A blind challenge for computational solvation free energies: Introduction and overview. J. Phys. Chem. B. 2009; 113:4501–4507. [PubMed: 19338360]
8. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ. The SAMPL2 blind prediction challenge: introduction and overview. J. Comput. Aided Mol. Des. 2010; 24:259–279. [PubMed: 20455007]
9. Skillman AG. SAMPL3: blinded prediction of host–guest binding affinities, hydration free energies, and trypsin inhibitors. J. Comput. Aided Mol. Des. 2012; 26:473–474. [PubMed: 22622621]
10. Mobley DL, Wymer KL, Lim NM, Guthrie PJ. Blind prediction of solvation free energies from the SAMPL4 challenge. J. Comput. Aided Mol. Des. 2014; 3:135–150.
11. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. Journal of the American Chemical Society. 1990; 112:6127–6129.
12. Sitkoff D, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. The Journal of Physical Chemistry. 1994; 98:1978–1988.
13. Hawkins GD, Cramer CJ, Truhlar DG. Pairwise solute descreening of solute charges from a dielectric medium. Chemical Physics Letters. 1995; 246:122–129.
14. Gilson MK, McCammon JA, Madura JD. Molecular dynamics simulation with a continuum electrostatic model of the solvent. Journal of Computational Chemistry. 1995; 16:1081–1095.
15. Tsui V, Case DA. Theory and applications of the generalized born solvation model in macromolecular simulations. Biopolymers. 2000; 56:275–291. [PubMed: 11754341]
16. Wagoner JA, Baker NA. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. Proc. Nat. Acad. Sci. USA. 2006; 103:8331–8336. [PubMed: 16709675]
17. Mongan J, Simmerling C, McCammon JA, Case DA, Onufriev A. Generalized Born model with a simple, robust molecular volume correction. Journal of Chemical Theory and Computation. 2007; 3:156–169. [PubMed: 21072141]
18. Tan C, Tan Y-H, Luo R. Implicit nonpolar solvent models. The Journal of Physical Chemistry B. 2007; 111:12263–12274. [PubMed: 17918880]
19. Hummer G. Hydrophobic force field as a molecular alternative to surface-area models. J. Am. Chem. Soc. 1999; 121:6299–6305.

20. Pitera JW, van Gunsteren WF. The importance of solute-solvent van der waals interactions with interior atoms of biopolymers. J. Am. Chem. Soc. 2001; 123:3163–3164. [PubMed: 11457039]
21. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 1983; 79:926–935.
22. Fennell CJ, Li L, Dill KA. Simple liquid models with corrected dielectric constants. J. Phys. Chem. B. 2012; 116:6936–6944. [PubMed: 22397577]
23. Wang J, et al. Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J. Comput. Chem. 2004; 25:1157–1174. [PubMed: 15116359]
24. Jakalian A, Bush BL, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. J. Comput. Chem. 2000; 21:132–146.
25. Paranahewage, SS., Gierhart, CS., Fennell, CJ. Predicting water-to-cyclohexane partitioning of the sampl5 molecules using dielectric balancing of force fields. current issue
26. Wolfram Research Inc.. Mathematica 10.4. 2016.
27. Steinbrecher T, Mobley DL, Case DA. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. J. Chem. Phys. 2007; 127:214108. [PubMed: 18067350]
28. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Comp. Phys. Comm. 1995; 91:43–56.
29. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. Gromacs: Fast, flexible, and free. J. Comput. Chem. 2005; 26:1701–1718. [PubMed: 16211538]
30. Hess B, Kutzner C, van der Spoel D, Lindahl E. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. 2008; 4:435–447. [PubMed: 26620784]
31. Pronk S, et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013; 29:845–854. [PubMed: 23407358]
32. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015; 1:19–25.
33. Goga N, Rzepiela A, de Vries A, Marrink S, Berendsen H. Efficient algorithms for langevin and dpd dynamics. Journal of chemical theory and computation. 2012; 8:3637–3649. [PubMed: 26593009]
34. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. J. Appl. Phys. 1981; 52:7182–7190.
35. Essman U, Perela L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh ewald method. J. Chem. Phys. 1995; 103:8577–8592.
36. BenNaim A, Marcus Y. Solvation thermodynamics of nonionic solutes. J. Chem. Phys. 1984; 81:2016–2027.
37. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comput. Chem. 1995; 16:273–284.

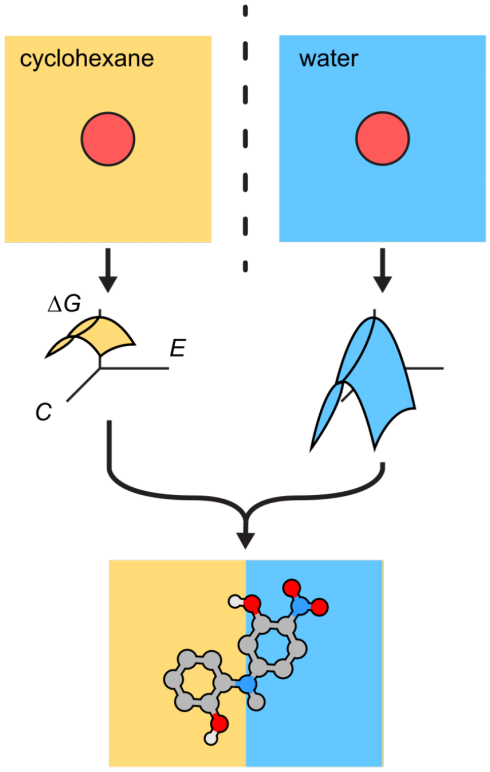

Fig. 1. SEA can predict the partition coefficient of a molecule in two solvents

SEA accurately estimates the solvation free energy of a molecule combining pre-computed free energy terms. Each one describes how the solvent interacts with a small region of the solute. The pre-computation is computationally expensive, but it needs to be done only once for a given solute at a given state point. At run-time SEA is fast since it only needs to assemble pre-computed data. Different solvent react differently to the presence of the solute: different look-up tables need to be built for different solvent. With SEA it is in principle possible to compute the solvation free energy of any molecule in any solvent. From these it is possible to evaluate the partition coefficient of a molecule in any pair of solvents.

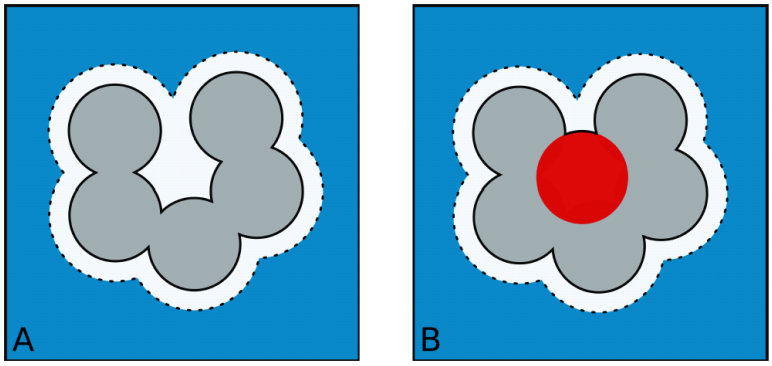

Fig. 2. The `buried-atom problem'

Some solute molecules are big enough to require more than just treating in the nonpolar term: (A) how solvent waters interact with surface atoms (gray). (B) The red sphere shows a buried atom, with which solvating waters will also interact. Here, we give a buried-atom correction.

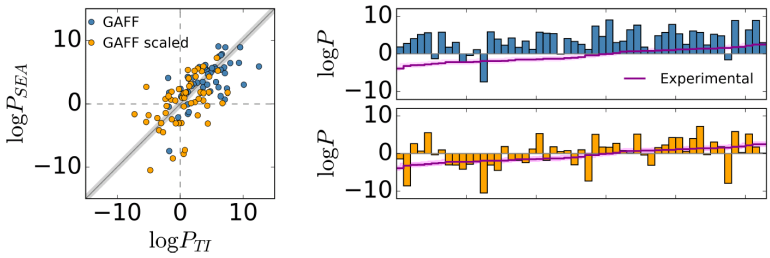

Fig. 3. Comparing Field SEA predictions to its underlying model (TI MD simulations) and experimental results

for the 53 SAMPL5 $\log D$ values. The deviation of points from the 45° line indicate where the SEA method differs from its underlying force field model. Shaded areas in the graphs represent an uncertainty of 0.61 $\log P$ units. This value comes from propagating a solvation free energy error of $k_B T$ in Eq.5. We compared two force fields for the solute: GAFF (blue), and GAFF-scaled (orange) with experimental data (purple line). Force field-wise GAFF-scaled is better than GAFF, which favors the cyclohexane phase too strongly.

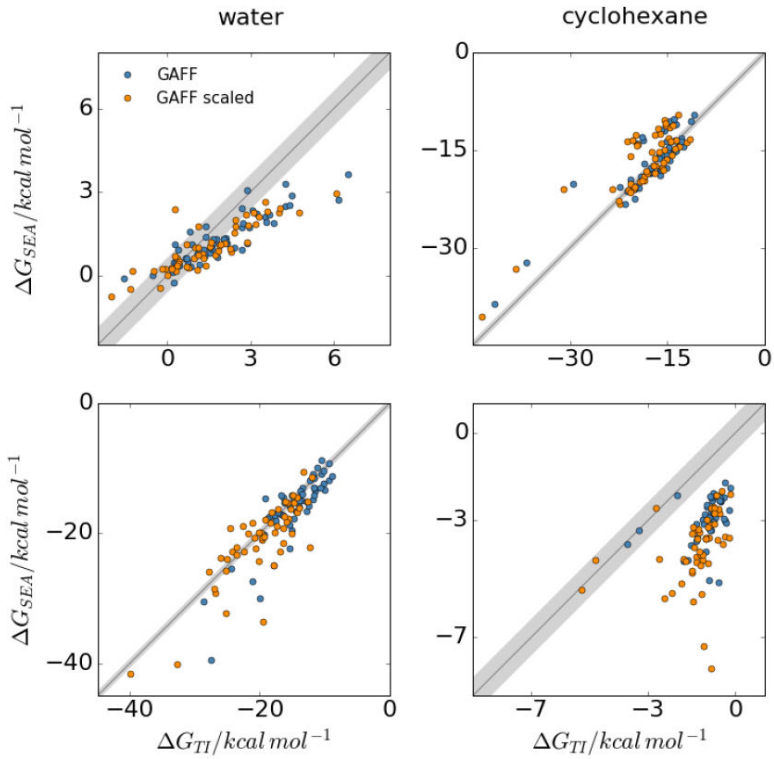

Fig. 4. Comparing polar and nonpolar terms in water and cyclohexane, for SEA vs. the TI simulations

The calculations are reported for two different solute force fields: GAFF (blue dots) and GAFF-scaled (orange dots). Shaded areas represent an uncertainty of $k_B T = 0.6\ kcal\ mol^{-1}$ in the determination of the solvation free energy. We note that our average statistical error is about 0.3 kcal mol^{-1}, but physically an error of $k_B T$ is more sensible. The errors (deviations from the diagonal line) are independent of the force field, and SEA shows reasonable agreement with TI. Note the differences in scales for the NP and P terms in the different solvents.

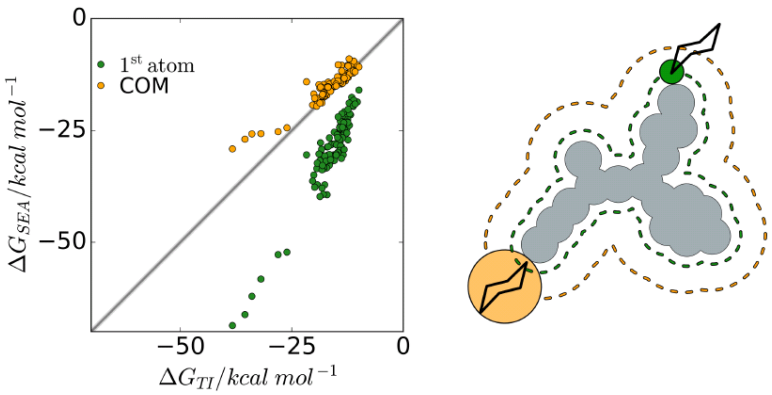

Fig. 5. What's the best cyclohexane solvation shell?

(Right) The gray object is an arbitrary solute. For *NP* solvation, the position of a solvating cyclohexane molecule can be determined either as cyclohexane's closest small atom (green circle at the top, predicting a tight solvation shell), or as cyclohexane's center of mass (COM) (orange circle at the bottom, predicting a loose solvation shell). (Left) Comparing SEA with the TI simulations shows that the COM is a better model of the solvation shell.

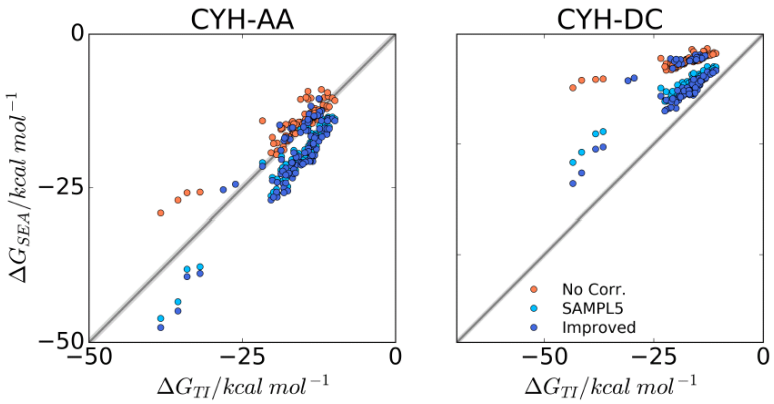

Fig. 6. Improvement from the buried-atom correction for *NP* solvation

When the effect of buried atoms is ignored (oragane points) the *NP* solvation free energy calculated using SEA tends to be underestimated (i.e. the solute is less soluble) compared to TI calculations. This effect is more evident in CYH-DC solvent. When the effect of buried atom is considered the points shifts in the right direction (blue and light blue points; see text for details). The applied correction (highlighted in eq.2) is quite approximate and this causes to be overestimated for CYH-AA.

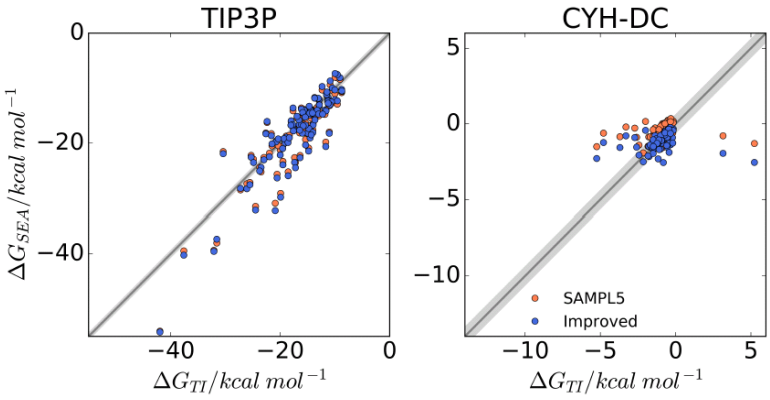

Fig. 7.

Polar solvation free energy calculated using SEA and TI for two different solvents for 106 comparison point for the SAMPL5 solutes. SEA value are reported on the vertical axes and TI values are reported on the horizontal ones. Each SEA calculation was performed using a *uniform* single function fit (orange) or a *piece-wise* multiple function fit (blue) polar table. It is clear that the more complicated *piece-wise* fit is not necessary to accurately describe the *P* solvation free energy of a solute in water or CYH.

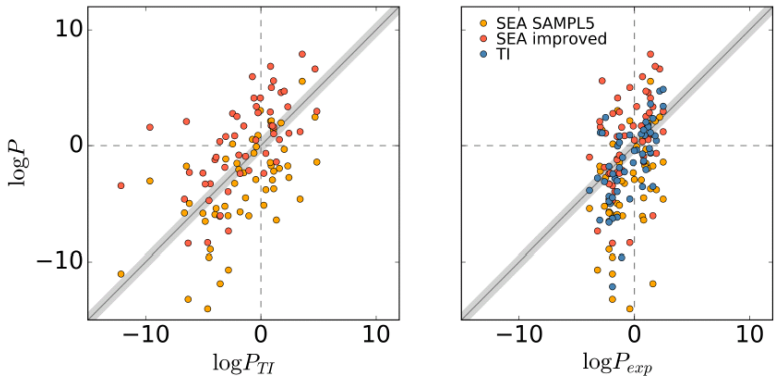

Fig. 8. Comparing SEA, TI, and experiments

On the left we show a comparison of log *P* SEA prediction against TI calculations. Yellow dots represent SAMPL submission and red ones represent the improved version of SEA. We can note how the predictions have been shifted upward. On the right we show the comparison of the same data with respect to experimental values. We also show in blue the results of the TI calculations. With the latest version of SEA tables and volume correction we see improvements of 1.4 log *P* units in RMSE, 2.1 log *P* unit in AUE and 1.1 log *P* unit in MSE.