Title

The Semi-Explicit Assembly Solvation Approach and Selected Applications

Permalink

https://escholarship.org/uc/item/90m652dg

Author

Kehoe, Charles Ward

Publication Date

2012

Peer reviewed|Thesis/dissertation

**The Semi-Explicit Assembly Solvation Approach
and Selected Applications**

by

Charles W. Kehoe

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

I am heavily indebted to several people for their guidance, support, and patience during the course of my graduate research career at UCSF. Foremost among them are Ken Dill, my advisor, and Chris Fennell, the postdoc who worked closely with me throughout my thesis project. They always made themselves available, and provided more advice and encouragement than I could have hoped for. This was largely true of everyone else in the Dill lab, too, who were great colleagues and friends.

My orals and thesis committee members likewise provided valuable advice and insights. Members included Matt Jacobson, Ajay Jain, Vojko Vlachy, and Brian Shoichet. I'm especially grateful to Ajay, who made a point of providing career guidance whenever grad school had me lost and confused. The UCSF facilities, as well as NSF and NIH funding, also played essential roles in enabling us to get things done.

Anyone who has been heavily involved in research knows that it can be a long and difficult road at times. Fortunately, my family and friends were there to help me take breaks and have fun along the way. I'm particularly grateful to my wife, Jennifer Bavani Krishnan Kehoe, whose patience and support was endless and invaluable.

Previous Publications

Chapters 2, 3, and 4 have already been published as articles in peer-reviewed journals. The publication details appear at the beginning of each chapter.

Statement from Ken Dill, thesis advisor, on chapter co-authors:

Charlie played an important role in the papers comprising chapters 2 and 3. He designed, coded, and maintained most of the computer program used by SEA to analyze molecules and produce solvation results. He also developed postprocessing scripts to compare SEA's results on large data sets to those of other solvation models, classified by functional groups, which aided greatly in understanding and improving the results. Additionally, he developed the technique of correlating the water's dipole response to the electric field. This pivotal final step brought SEA's results nearly in line with those of explicit solvent.

For chapters 4 and 5, Charlie is listed as first author because he was responsible for nearly all of the content. He designed and performed nearly all of the research, created most of the figures, and wrote most of the text. As mentioned previously, the actual SEA method used was carried out by code mostly written by him. The other authors played a significant advisory role, but contributed a small amount to the actual research, figures, and text for each chapter.

Charlie played a critical early role in the development of this research, and he has extended it in individually driven and unique ways. The dissertation documents his initial efforts and his development into an original researcher and a foremost specialist in the science of solvation at the molecular level.

The Semi-Explicit Assembly Water Model and Selected Applications

Charles W. Kehoe

We present a new solvation approach, capable of accelerating the computations required for explicit solvent modeling by several orders of magnitude. This technology allows researchers to apply the most accurate solvation models available without sacrificing speed. Semi-Explicit Assembly (SEA) runs at nearly the same speed as the fastest solvation models currently in use. We make this possible by distilling out and reproducing the most important physics modeled by explicit solvent. This approach can be applied to any model of any dipolar solvent, even at the quantum level (though we do not demonstrate that here). Since water is the solvent found in almost all biological interactions, we have so far limited our work to aqueous solvation.

SEA is based on a set of precomputations which sample solvation behavior around single atoms with varying properties, and record the response of explicit solvent to all common sets of atomic parameters. The most important component turns out to be water's reproducible, but asymmetric, response to the electric field on the molecular surface. Treating waters as particles, rather than as a dielectric field, also allows us to reproduce important physical details.

The first half of this work presents all the details of the SEA approach, along with several sets of results demonstrating its potential. We show that, given a set of conformations to work with, we can match the accuracy of explicit solvent at a similar speed as the simple Generalized Born solvation model. We present several sets of molecules which we can model at explicit solvent accuracy, but in which standard approaches like γA and Poisson-Boltzmann (PB) miss important details.

The second half of this work presents additional applications. The first is a blind solvation test using multiple forcefields and water models. We achieved the second-highest accuracy to experimental values in the test, but our approach was also one of the fastest used. The second application is a visualization tool, named SurfMap, which allows users to explore several types of solvation results using interactive three-dimensional models. SurfMap demonstrates additional solvation details which SEA catches, but PB does not.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Author: Charles Kehoe

There is considerable interest today in molecular modeling and simulation. Academic and commercial researchers are increasingly finding that simulation offers a level of detail, control, and convenience that experimental methods often cannot match. Protein structures can be obtained via X-ray crystallography, estimated from known structures of their relatives,[87] or sometimes determined ab initio based on many layers of statistical information.[75] Simulations can then probe these structures for atomically detailed information about folding,[101,135] misfolding,[133] and stability,[14] as well as interactions with ligands[89] and other biomolecules.[76] There is considerable interest (and funding) from the pharmaceutical research community in technologies such as ligand docking[70] and rational drug design.[3,88] And interest will only grow further as the software improves, and hardware continues to get cheaper.

Water is the solvent in which nearly all biochemistry takes place. It has several unusual properties which make possible biology as we know it, and consequently requires careful treatment in a simulation setting. There are several common ways of simulating water, broadly classifiable as either *implicit* (treating water as a smooth dielectric continuum) or *explicit* (treating water as a collection of many individual molecules). Predictably, implicit models are fast to calculate, but not that accurate, while explicit models are very computationally intensive, but produce more correct results when used correctly.[94]

Primarily due to their high performance, implicit solvation models are still widely used in biomolecular simulation. The Generalized Born[125] family of solvation models, for example, simply modify the solute's coulomb interactions to account for the surrounding dielectric solvent. Models based on the Poisson-Boltzmann equation[23] require numerical integration, but are still reasonably fast. However, these methods require carefully recalculated atomic radii to produce accurate results.[119] When used with the standard radii provided by an all-atom forcefield, their numerical accuracy suffers as compared to explicit solvent results, as we shall see in chapters 3 and 4. Because they treat water as a continuum of infinitely small particles, they are also unable to correctly reproduce certain biochemical phenomena, such as the correct PMFs of salt bridges.[129]

Explicit solvent and its variants are sometimes also used to explore the behavior of small and large molecular systems. However, their high computational expense – sometimes several orders of magnitude slower than implicit methods – limits their applicability. The ideal situation, of course, would be to achieve the accuracy of explicit solvent methods at the speed of implicit solvent. This may seem like an impossible dream, but we have been able to achieve it in several molecular simulation areas.

How is it possible to skip all the work that explicit solvent does to get to its carefully refined answers? The answer lies in critically evaluating the physics being simulated, and the particular answers one is looking for. This allows several simplifications of the problem that lead to dramatically improved computational speed.

We have focused on modeling the correct physical and energetic interactions between molecules and the surrounding solvent. Especially for uncharged molecules, a significant majority of water's response to a solute occurs in the first layer of water molecules around the solute (or the first solvation shell). Also, the advantages of explicit solvent based on "discrete size effects", such as correct salt bridge behavior, usually result from interactions between solute atoms and first-shell water molecules. So the first simplification we used was to consider the first shell of water molecules explicitly, and use an implicit model for the rest of the solvent. This approach has been tried before,[103] but it still requires explicit simulation of many water molecules, and it usually introduces the new problem of how to create and destroy waters as they exchange in and out of the first shell. We wanted to go farther with our model.

The next step in focusing our computational effort came in noticing that the individual motions of waters were not that important to the energetic behavior of the whole system. Simulating these motions over time amounts to a huge computational expense. One of water's unique properties is its many degrees of rotational freedom, which all must be integrated separately. Worse, some explicit water models have up to five charge centers, all of which interact with all of the charge centers of every other water. (In computer science, this type of $O(n^2)$ computational complexity is often frowned upon.) But since each water that exchanges out of the first solvation shell is quickly replaced by (or was originally pushed out by) another, what actually matters is the favorability of certain *locations* and *orientations* for surface water molecules. In order to eliminate explicit water simulation entirely, we needed to boil down explicit water's behavior (the pun is accidental) to just these essentials. How we implemented these optimizations is the subject of chapter 3.

The technical discussion up to this point has mostly focused on methods of treating water's *polar* solvation effects: its dielectric-like response with associated finite-size effects. A somewhat simpler question, and one that we address first, is that of the solvent's *nonpolar* effects: the energetic consequences or advantages of dispersion interactions. For calculating free energies of solvation, these effects require completely removing the solute from simulation in a slow, controlled process. The energetic consequences are usually somewhat small, but they can be fantastically expensive to compute because of the convergence problems involved. This is another area where our Semi-Explicit Assembly (SEA) approach shines: again, it focuses on the exact

answers we are looking for, and reframes the entire calculation in terms of them. The SEA nonpolar approach calculates the effective nonpolar character of every point on the solute surface, and then estimates (quite accurately) explicit solvent's response to this landscape. Further details can be found in chapter 2.

In later chapters, we illustrate SEA's potential by showing its fast, accurate performance on a variety of data sets. We compare our results to those of similar methods, showing that we provide real advantages in both performance and accuracy. Work is ongoing in making these approaches available to a wider audience and in more diverse applications, including molecular dynamics. We hope you enjoy reading about Semi-Explicit Assembly.

Chapter 2

Oil/Water Transfer Is Partly Driven

by Molecular Shape,

Not Just Size

Abstract

We present a new approach to computer modeling of solvation free energies of oil in water. In *Semi-Explicit Assembly*, we first precompute structural and thermal properties of TIP3P waters around different Lennard-Jones spheres. This tabulated information is then used to compute the nonpolar solvation properties of arbitrary solutes. By accumulating interactions from whole regions of the solute molecule, Semi-Explicit Assembly more properly accounts for effects of solute shape and solves problems that appear as nonadditivities in traditional γA approaches.

Semi-Explicit Assembly involves little parameter fitting because the solute and water properties are taken from existing force fields. We tested the predictions on alkanes, alkynes, linear and planar polyaromatic hydrocarbons, and on a diverse set of 504 molecules previously explored by explicit solvent simulations. We found that not all hydrocarbons are the same. Hydrocarbons have 'hot spots', places where first-shell waters interact more strongly with the molecule than at other locations. For example, waters are more attracted to hover over hydrocarbon rings than at the edges. By accounting for these collective regional effects, Semi-Explicit Assembly approaches the physical accuracies of explicit solvent models in computing nonpolar solvation free energies, but because of the pre-computations and the regional additivities, it is nearly as fast to compute as γA methods.

2.1 Introduction

Various processes in nature – the folding of proteins, the self-assembly of lipid bilayer membranes and soap micelles, the chromatographic separations of materials, the binding of drugs to proteins, and the partitioning of environmental toxins into fish oils – are driven, at least in part, by the solvation or desolvation of oil-like molecules in water. Two approximations have been commonly used in modeling the molecular solvation of oil in water:

(1) The solute-solvent interface is assumed to be a miniature version of a macroscopic liquid interface. Key knowledge of *hydrophobic interactions* derives from bulk-phase experiments such as measurements of the interfacial tension γ between oil and water, where $\Delta G = \gamma A$ is this nonpolar free energy of transfer; it increases in proportion to the interfacial area A. Microscopic solvation processes such as protein folding are often treated as sums of transfers of subcomponents, such as an oil moiety from water to oil. For example when a protein folds, its oil-like amino acids are transferred from a state of exposure to water to a state of burial in a nonpolar core. The free energies for such processes are often estimated by a quantity of the same form, $\gamma_i A$, where A is a microscopic property – the surface area of the oil molecule (which can be estimated in different ways), and γ_i is a parameter chosen for a particular type of chemical moiety i. This is the approach generally taken in "implicit" computer models of water.

(2) Solvation energies are approximated using group additivities. A main approach to computing solvation free energies for complex processes is to assume additivity and sum the free energies of component parts. Central to this enterprise are *hydrophobicity scales*, which are lists of free energies of transfer – typically between an oil or vapor phase and water – of model compounds that represent the component parts. There are more than 30 hydrophobicity scales for the amino acids alone[13,27,72,102,116,143,144] and many more for simple hydrocarbons and chemical groups.[8,18,116,121] This model-compound / hydrophobicity-scale approach rests on the underlying assumptions of additivity and transferability. Model-compound / hydrophobicity-scale studies would have little value if the component quantities measured in a simple oil/water experiment were not applicable to more complex media such as the interiors of lipid bilayers, protein cores, or nonpolar chromatographic stationary phases, extending to situations beyond just the direct measurements themselves. Such additivity approaches require the assumption that one methylene group or one amino acid somewhere in the molecule is equivalent to another methylene group or amino acid somewhere else. In this way, solvation free energies are assumed to only depend upon the numbers and types of substituents, and not their geometric arrangements.

Moreover, hydrophobicity scales depend on the premise of equivalence, namely that one type of oil is essentially the same as another type of oil. As Tanford and Nozaki noted in one of their first publication on such scales,[102] in order to have a 'scale' that spans from some extreme of maximum nonpolarity to the other extreme of maximum

polarity requires a 'gold standard' of nonpolarity. Which type of oil best represents the essence of 'nonpolarity'? If oils were all different, then it would be impossible to capture the spirit that somehow all protein cores or all lipid bilayers have the same property of being 'hydrophobic'.

Some limitations of the γA approach

Some of the problems with these simple approaches to molecular solvation are known.

(1) Solute shape matters too, not just surface area. The γA model treats only the dependence of solvation free energy on solute surface area and not on solute shape. Yet, water adopts very different structures and thermal properties around highly curved or nonlinear solutes than around planar solutes or large (protein-sized) objects having the same surface area.[22, 25, 57, 115, 116, 123, 126, 137] One result is that $\Delta G/A$ measured from interfacial tensions at planar surfaces is 75 cal mol^{-1} Å^{-2} from interfacial tension measurements, but only $\Delta G/A \approx 30$ cal mol^{-1} Å^{-2} for small-molecule hydrocarbon/water transfer[115] and 5 cal mol^{-1} Å^{-2} for air/water transfer, the value typically used in implicit models.[24, 113]

(2) Dispersion interactions do not have the same form as cavity formation costs. Dissolving a solute in water entails: (1) opening a cavity in water, which involves unfavorable water ordering or unfavorable hydrogen-bond breaking in

water, then (2) inserting the solute, which involves favorable dispersion interactions of the solute with the water. Both terms are treated in the scaled particle theory approach,[52,108] for example. Often, both terms are assumed to have the same mathematical form and are captured in a single γA quantity; in this approach both cavity formation and the attractive dispersion interactions are assumed linearly dependent on the solute surface area. However, Pitera and van Gunsteren showed that this simplification leads to underestimating the true attractive aspects of nonpolar solvation, a nearly 50 kcal/mol oversight for small proteins.[109] A better accounting of dispersion interactions has been a driving force for new methods for treating nonpolar solvation.[43,79,128,136]

(3) Different oil phases are different. Additivity sometimes doesn't work. While Tanford and Nozaki did show that partitioning into water is not strongly dependent on the types of oil in some cases,[102,130] more recent studies have shown that partitioning can be substantially dependent on what oil is used for the oil phase,[111] indicating the limitations of this assumption. The atom arrangements, densities, and chemical character differ between different molecules. Solvation free energies can sometimes be non-additive because of these microscopic details.[32,85] Treating chemically distinct solute surfaces additively with a uniform γ parameter misses these effects.

The two standard routes to improved solvation modeling are: (1) to include additional parameters,[28,33,46,106,128,136] or (2) to perform 'explicit water' computer simulations, but at considerably greater computational expense and loss of simplicity.[93,117]

Here, we present a third approach, which we call *Semi-Explicit Assembly.* We use parameters and water models that are taken directly from explicit-water forcefields, so our approach does not involve 'learning' or parameterization from databases. Semi-Explicit Assembly retains the simplicity of a type of additivity, but it is *regional*, collectively capturing results from multiple solute groups at the same time. In this way, it correctly captures effects that would be described as non-additivities in the simpler group-additivity approaches. Also, as a consequence of this additivity and of a pre-calculation step, this approach is computationally nearly as fast and simple as γA methods, and is much faster than explicit solvent simulations. Nevertheless, we find that the quality of the modeling is close to that of explicit solvent simulation modeling.

2.2 The Semi-Explicit Assembly approach to nonpolar solvation

Our aim in Semi-Explicit Assembly is to capture the parameters and much of the physics from explicit solvent modeling within a rapidly computable implicit framework. To do this, we use fully explicit solvent simulations to pre-compute the behaviors of waters around a series of nonpolar solute spheres having different radii and attractive dispersion interactions. After this one time precomputation, we probe the local interaction identity of an arbitrary solute molecule and assemble its nonpolar solvation free energy.

Pre-computations of Lennard-Jones spheres in explicit water

In computer modeling, molecules are usually represented as collections of bonded spheres. Steric repulsion and attractive dispersion interactions are most often handled using a standard Lennard-Jones (LJ) pair potential,

$$
V_{\mathrm{LJ}}(r_{ij}) = \begin{cases} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] & r_{ij} \leq r_c \\ 0 & r_{ij} > r_c, \end{cases} \tag{2.1}
$$

where the size (σ) and well-depth (ϵ) parameters account for the steric and dispersive elements respectively[a], r_{ij} is the distance between particles i and j, and r_c is an interaction cutoff distance. To gather the physics of solvation using LJ spheres, we start by performing explicit solvent free energy calculations to compute their nonpolar solvation free energy (ΔG) spanning a wide range of σ and ϵ values. These calculations are based on a constructing a thermodynamic cycle connecting simulations of the LJ sphere in two different media. We transfer the solute between vacuum and water, and obtain ΔG for this transfer process. This is similar to a combined scaled-particle theory approach,[52,108,126] where cavity formation and interaction activation steps are carried out simultaneously. It should be noted that total solvation free energies include both a polar and nonpolar part. As indicated earlier, we are interested exclusively in the nonpolar part throughout this study, so atomic partial charges are always set to zero.

---

[a]We use Lorentz-Berthelot combination rules, where σ_{ij} is an arithmetic mean of the individual particle diameters ($\sigma_{ij} = [\sigma_i + \sigma_j]/2$) and ϵ_{ij} is a geometric mean of the individual particle
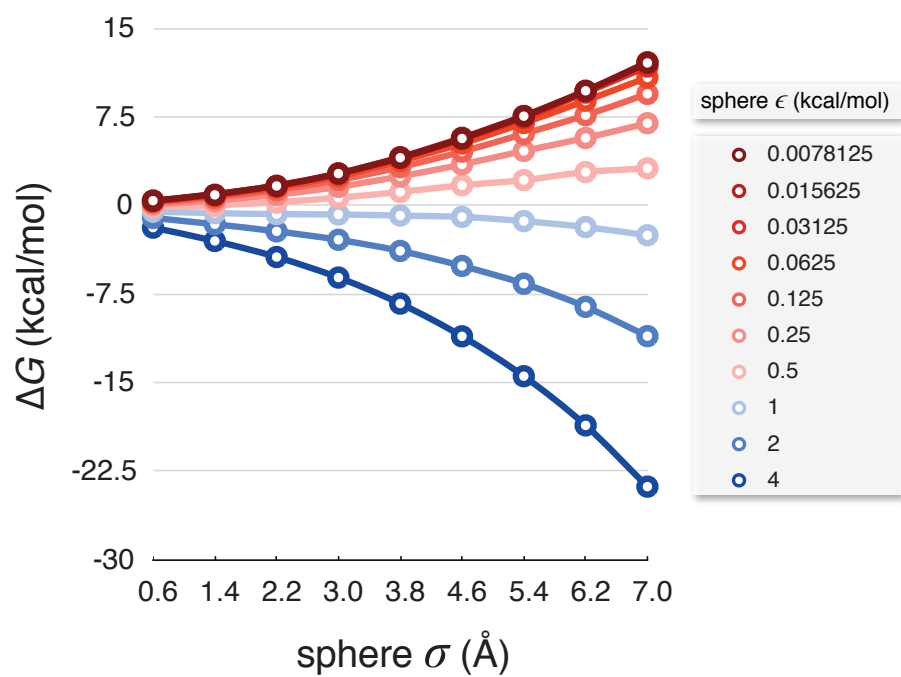
Figure 2.1: Nonpolar solvation free energy (ΔG) of single LJ spheres in TIP3P water at 300 K as a function of their σ and ϵ parameters. Unfavorable ΔG values are red. Favorable ΔG values are blue.

Figure 2.1 shows the pre-computed values of ΔG across a range of LJ spheres solvated in the TIP3P water model.[64] Increasing the LJ well-depth gives more favorable solvation free energies. As the well-depth decreases, the ΔG values converge toward the previously observed ΔG limit for purely hydrophobic hard spheres.[22,57,58] A crossover from unfavorable to favorable solvation occurs around a well-depth value of $\epsilon = 0.75$ kcal/mol. Individual atoms in molecular simulations rarely have dispersion attractions this strong, but we include these simulations because we find that collections of atoms can have attractive potentials of this magnitude.

The pre-computation step that generates Figure 2.1 is computationally expensive, but it is only performed once for any given temperature, pressure, or solvent model. After the values in this plot are determined, they can be applied in much faster computations for any given solute.

At the same time we compute ΔG values, we also construct a table of average separation distances between the solute and first-shell water. We collect these distances (r_{w}) from radial distribution functions of water oxygen atoms with respect to the centers of each type of LJ sphere; see Figure 2.2a. These distances are collected in a table as a function of σ and ϵ of the LJ spheres.

---

interaction well-depths $\left(\epsilon_{ij} = \sqrt{\epsilon_i \cdot \epsilon_j}\right)$.
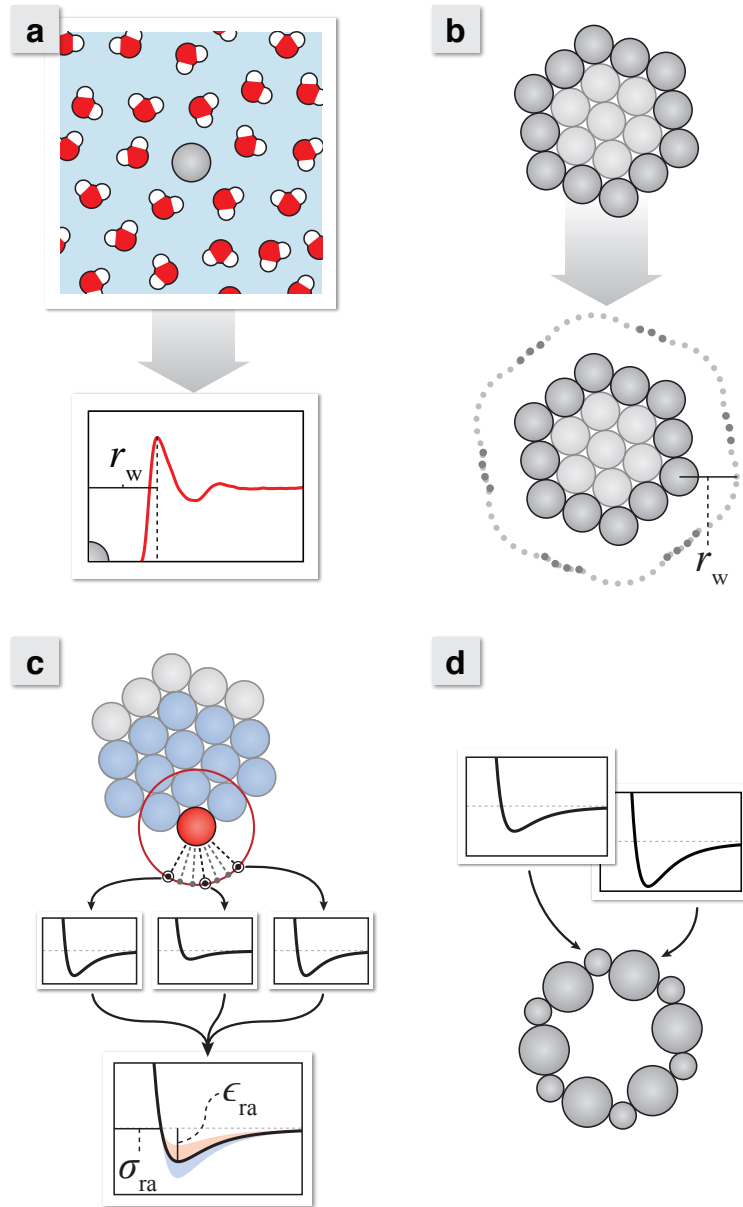
Figure 2.2: The process for incorporating non-additive environmental effects on the solute surface atoms. (a) Sample LJ spheres in explicit water and build a map of water distances (r_{w}) as a function of σ and ϵ. (b) Construct the solvent accessible surface (SAS) using the distances from the explicit solvent map. (c) Probe the LJ potential of the solute along the line connecting each SAS dot to its surface atom. Average these potentials for each surface atom, and extract new "effective" LJ parameters (σ_{ra} and ϵ_{ra}) from this curve. (d) Use these effective potential parameters when calculating the solvation free energy. Note that edge atoms will have more attractive ϵ_{ra} values than corner atoms because of the greater number of atoms near to the probe particle.

Assembly of molecular solvation free energies

The explicit solvent pre-computations provide a detailed picture of how the chosen water model will solvate simple nonpolar spheres. To estimate the nonpolar solvation free energy of arbitrary solute molecules, the results from these representative atomic systems need to be brought to the unique solute surfaces. This assembly process is shown in Figures 2.2b, 2.2c, and 2.2d.

(1) Compute solvent-accessible surface (SAS) of the solute. For every atom of the solute molecule, with its given radii and LJ parameters, we look up (or interpolate) from the pre-computed table of r_w values, the average contact distance of the surrounding solvent. We form spherical accessibility boundary points around each solute atom from these interpolated r_w values and cull out points that are inaccessible due to other neighboring atoms. This generates an initial molecular SAS; (Figure 2.2b). This SAS differs from that of Lee and Richards[74] in two ways: (1) we do not use a hard sphere probe, so, in principle, our solvation boundary expands or contracts with pressure and temperature, and (2) the interactions governing solvent accessibility will not be with only a single nearest-neighbor solute atom (see below), hence we capture contributions from other nearby atoms.

(2) Compute a *region-averaged* dispersion-potential field. Now we construct local dispersion potential fields at different points in the solvation shell around the solute. First, we define a vector from a water dot point of the initial SAS to the

center of the associated solute surface atom. In Figure 2.2c, the current surface atom is colored red. The dashed lines show vectors connecting this target solute atom to its SAS dot sites. Starting at the SAS dot sites, we probe the regional LJ potential field along these vectors. This regional field we use encompasses all solute atoms within the r_c of the target surface atom, the blue particles in Figure 2.2c. The gray particles are outside this cutoff and therefore ignored. By including more surrounding interactions, longer r_c values will result in a more accurate depiction of the total solute dispersion potential. Rather than use an infinitely long cutoff, we found that

$$r_c = 2r_{\mathrm{max}} + r_{\mathrm{ww}}, \tag{2.2}$$

gives results that were converged within the calculated error for the overall nonpolar free energy. Here, r_{max} is the maximum r_{w} found for the atoms making up the solute, and r_{ww} is the water–water packing distance extracted from a water–water radial distribution function ($\sim$2.7 Å).

In this probing process, the LJ interactions are accumulated between the solute atoms within the region described above and a probe particle along the dot site vector. The pairwise LJ potential (Equation 2.1) is dependent on both σ_{p} (the probe σ parameter) and atom σ parameters. So now σ_{p} becomes a parameter that is determined as discussed below. This probe particle is progressively stepped closer to the surface in order to construct a potential as a function of probe particle position, and this potential is stored for each surface dot. As shown in Figure 2.2c, the wells of potentials

calculated for surface sites in closer proximity to more solute atoms (those nearer to solute edges rather than corners) will tend to be deeper. After constructing potentials for each surface dot about a solute surface atom, we average them together to generate a region-averaged dispersion potential (V_{ra}) which incorporates shape and the attractive interactions of nearby collections of solute atoms. We then extract region-averaged parameters for each surface atom (σ_{ra} and ϵ_{ra}) by fitting this curve to an LJ potential,

$$V_{\mathrm{ra}}\left(r\right) \approx 4\sqrt{\epsilon_{\mathrm{ra}} \cdot \epsilon_{\mathrm{p}}} \left[\left(\frac{(\sigma_{\mathrm{ra}} + \sigma_{\mathrm{p}})}{2r} \right)^{12} - \left(\frac{(\sigma_{\mathrm{ra}} + \sigma_{\mathrm{p}})}{2r} \right)^{6} \right], \qquad (2.3)$$

where r is the distance between the probe and the target surface atom. As V_{ra} is an average of collective atom dispersion potentials, fitting it to a single LJ potential is an approximation. It should be noted that the averaging procedure is technically unnecessary. We could retain a more detailed map of the dispersion potential based on these more numerous surface points rather than on an averaged, per atom basis. We have tested both routes, and they are equivalent for the nonpolar solvation of small molecules shown below, so we use the per atom averaging step for convenience.

(3) Reduce the region-averaged field to a single effective LJ interaction. Assign these newly derived σ_{ra} and ϵ_{ra} parameters to the associated surface atoms (Figure 2.2d). This procedure encodes information about the full solute structure and interactions into the solvent exposed regions of the molecule.

From the steps above, we obtain free energy component quantities of the solute that can be added to get the total nonpolar solvation free energy,

$$\Delta G = pV_{\mathrm{v}} + \sum_{i=1}^{N} f_i \Delta G_i. \tag{2.4}$$

Here, f_i is the fraction of the surfaced exposed for atom i, and ΔG_i is that atom's free energy term extracted (via a linear interpolation) from the map pictured in Figure 2.1 using the region-averaged parameters as our σ and ϵ values. The pV_{v} "void" term is the cavity formation cost due to the buried particles within the molecule (the lightened atoms in Figure 2.2b)[b]. For small molecules, this void term is often zero because all the solute atoms also happen to be surface atoms. We found setting the pV_{v} term to zero for all the molecules studied within a good approximation. If one is interested in the absolute nonpolar solvation of macromolecular structures, optimization of this void term will become increasingly important.

As the SAS is a set of discretized points, the f_i will depend on the number of points remaining after the culling process in Step 1. Culling points near the intersection of nearby solvent-accessible surface shells will result in a slightly jagged edge. To arrive at converged estimates of the f_i values, we iterate over constructing the SAS and the calculation of ΔG in Equation 2.4. In these series of Step 1 surface constructions, the region-averaged LJ parameters are used to determine new r_{w} distances. In this way, this SAS used to determine the f_i values incorporates the collective structure of the solute molecule.

---

[b]In this void term, the p term can be taken as the negative transfer free energy per unit volume of the solvent, or it can be treated as an adjustable fitting parameter.[128,136] The void volume is, $V_{\mathrm{v}} = V_{\mathrm{sol}} - V_{\mathrm{surf}}$, where V_{sol} is the total solute volume, while V_{surf} is the volume of a molecular structure composed only of the surface atoms.

Optimization of the dispersion potential probe

In order to calculate a particular nonpolar solvation free energy, we must optimize the probe size (σ_p) for the Lennard-Jones field sampling procedure. This is necessary to insure that we pick up the surrounding dispersion interactions properly, and is similar to attractive probe optimization procedures in alternative techniques.[79, 128, 136] While one could optimize the probe size to a large set of target molecules, we decided to start with a single molecule, n-pentane - the middle sized molecule of our linear alkane series. After setting $\epsilon_p = 1$ for convenience, we scanned σ_p values in 0.01 Å increments and sought to minimize the difference between the ΔG using our method and the explicit solvent ΔG for n-pentane. We tested different ϵ_p values and found that the choice of ϵ_p does not change the results. The optimized σ_p value of 0.82 Å turns out to be quite robust, and one can select values within a 0.1 Å window about this midpoint without significantly altering the results below. We attempted optimizing over a larger set of small molecules, but this did not significantly alter σ_p or lead to improvements in the solute ΔG estimations.

Algorithm performance and computational details

There is probably no simple and fair way to compare various methods for computational speed. However, the following provides a good rough estimate. Standard γA approaches are limited by the computational cost of constructing the solvent-accessible surface. These are currently the fastest available methods. Semi-Explicit

Assembly, too, requires construction of the solvent-accessible surface. Additionally, there are the probing step described above which will optimally cost the same as construction of a solvent-accessible surface, and a reconstruction of the solvent-accessible surface with the region-averaged LJ parameters. Thus, in an optimized implementation, the maximum speed of Semi-Explicit Assembly would be about 3-fold slower than γA methods.

The free energy surface of LJ parameters pictured in Figure 2.1 was constructed using explicit solvent free energy calculations of individual spheres in cubic boxes of 1000 TIP3P water molecules at 300 K and 1 atm. The LJ σ values for solute particles in this map cover a range of 0.6 to 7.0 Å linearly in 0.8 Å steps. The LJ ϵ values range from $\sim$0.008 to 4 kcal/mol, where each subsequent ϵ value is two times the previous value.

The free energy calculations were performed using thermodynamic integration with GROMACS 4.0.[55] In thermodynamic integration, the LJ solute particles are reversibly transformed between a fully interacting and non-interacting state over a series of simulation windows, each with their own transformation parameter (λ). Integrating the change in the potential over the change in λ over the full range of λ values gives the free energy difference between these states. A detailed description of the theory behind such calculations can be found elsewhere.[44,71,127] Here, twenty one windows were used for the transformation process, and they spanned $\lambda = 0$ to 1 in even steps of 0.05 units. A soft-core potential was used to minimize integration

error in the transformation process,[124] and the specifics of the actual simulations followed those outlined by Mobley et al.[98] One exception was that the interaction cutoff needed to be longer to accommodate the large particle sizes explored as part of this series. Thus, the LJ cutoff radius was smoothly switched off between 11 and 13 Å. Errors in the free energies were estimated by the limiting value of block averages.[54]

For the polyaromatic hydrocarbon series, explicit solvent free energy calculations were performed on naphthacene, pentacene, hexacene, triphenylene, and perylene because literature values of ΔG were unavailable. The specifics of these calculations are identical to those described above for single LJ spheres, with the exception of larger numbers of water molecules in order to maintain hydration layers thicker than the specified cutoff lengths. The LJ parameters for these molecules were assigned using the general AMBER force field (GAFF).[139]

The Semi-Explicit Assembly nonpolar solvation free energies were averaged values from 40 dot surfaces construction iterations, each using the same set of region-averaged LJ parameters calculated using the initial dot surface. For each of these dot surfaces, spheres of ~300 dots per atom were randomly rotated before culling overlapping points. The ΔG values for all the solutes come from single calculations about the dominant clustered conformation from the explicit solvent simulations. We attempted more detailed configuration analyses for several of the molecules that contained multiple rotatable bonds; however, this led to negligible changes in the final values, so we chose to simply take the dominant conformation as representative of the

whole. With electrostatic effects being much stronger than dispersion, it is likely that the polar part of the free energy is much more sensitive to changes in internal conformations. Calculated error for Semi-Explicit Assembly over the 40 iterations was 0.05 kcal/mol, averaged over all molecules explored in this study. A post-calculation analysis indicated that a similar error can be obtained with fewer than 5 dot surfaces construction iterations.

2.3 Results

For testing our solvation approach, we assume the 'gold standard' right answer solvation free energies are given by experimental data where it exists, or otherwise by all-atom explicit solvent free energy calculations.[93] Here, we compare our predictions to these explicit solvent simulations, to experiments where possible, and to γA values. In supplementary material, we show that this semi-explicit method is also more accurate than other recent approaches.[128,136]

Linear hydrocarbons

The standard first test of solvation models are the linear n-alkanes. Figure 2.3a confirms that the present model agrees with experiments, explicit-water simulations, and standard γA models for these molecules. Interestingly, because of its assumed
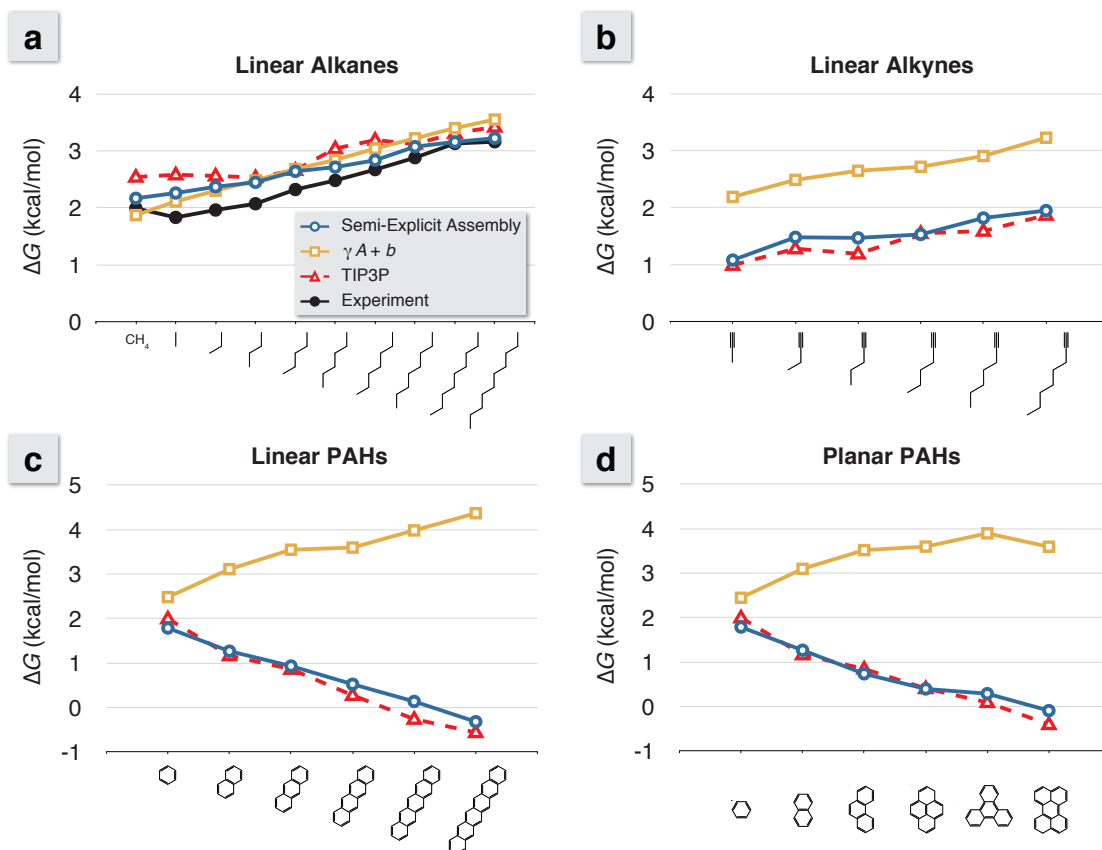
Figure 2.3: The nonpolar solvation free energy for a series of a) linear alkanes, b) linear alkynes, c) polyaromatic hydrocarbons (PAHs) in a linear arrangement, and d) PAHs in a planar arrangement calculated using $\gamma A + b$, Semi-Explicit Assembly, and explicit solvent. For $\gamma A + b$, the traditional $(0.00542 \times \mathrm{SA_{tot}}) + 0.92$ was used,[113] and the TIP3P results are those obtained through explicit free energy calculations.[93] Experimental comparisons to ΔG cannot be drawn with the linear alkynes or PAHs series, because they have a substantial polar term to the overall solvation.
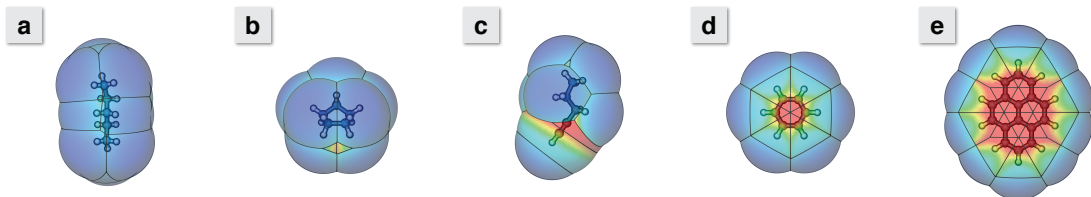
Figure 2.4: Maps of the collective dispersion attraction about the solvent accessible surface (SAS) of a) *n*-pentane, b) cyclopentane, c) pent-1-yne, d) benzene, and e) pyrene. The color of the surface indicates the LJ well-depth, with blue starting at 0 kcal/mol and red lowering to deeper than 5 kcal/mol. Note the red "hot spots" around the triple bond in pent-1-yne and in the center of the benzene and pyrene ring planes. These indicate a significant enhancement of dispersion attraction with the surroundings. As these regions grow with increasing molecule size, these collective dispersion attractions will offset the cost of cavity formation in surrounding solvent. With a simple γA, all these surfaces would be a uniform blue.

linear dependence, typical γA methods give an erroneous prediction for the intercept, $\gamma A + b$, where $b = 0.92$ kcal/mol corresponds to insertion of a solute of near zero size. In reality, the value should be much closer to zero for a solute of zero size. Explicit solvent simulations with TIP3P water give a value of ~ 0.2 kcal/mol. Because our Semi-Explicit approach derives from explicit simulations, our values approximately equal the explicit values.

Figure 2.3b shows solvation free energies for the linear alkynes, from the various models. Alkynes have a carbon-carbon triple bond at the end of the chain. In GAFF,[139] the dispersion interaction well-depth is twice that of carbon-carbon single bonds. Like the explicit simulations, but unlike γA, the Semi-Explicit approach captures the more favorable aqueous solvation of the alkynes relative to the alkanes. Figures 2.4a and 2.4c show that the extra attraction for water of the alkynes is localized near the triple bond.

Hot spots: not all hydrocarbons are the same

Figures 2.4a and 2.4b show the LJ potential surfaces for n-pentane and cyclopentane. Seams between atom surfaces form favorable interaction "hot spot" regions, while methyl end-groups of the alkane chain are a deeper blue and less favorable. The surface area of cyclopentane is less than n-pentane, but this only accounts for a modest decrease of 0.2 kcal/mol in ΔG when using $\gamma A + b$. This modest change is much less than the greater than 1 kcal/mol decrease seen experimentally.[8,9] Semi-Explicit Assembly includes the effects of these "hot spots" and lowers ΔG by an additional 0.4 kcal/mol. The remaining difference between the estimated and the experimental value likely comes from approximations in the Semi-Explicit Assembly approach, such as the void term discussed previously and the incomplete capturing of solvent-solvent interaction enhancement from optimal hydration cages.

Polyaromatic hydrocarbons: linear and non-linear topological effects

The solvation free energies of polyaromatic hydrocarbons (PAHs) provide a more stringent test. Aromatic rings have an important asymmetry. A water molecule at the lateral edge 'sees' one methylene-like group and its two lateral neighbors. But a water molecule centered above or below the plane sees 6 methylene-like groups; see Figures 2.4d and 2.4e. These combined attractions counter cavity formation costs,

resulting in more favorable nonpolar solvation with larger arrangements of aromatic rings; see Figures 2.3c and 2.3d. The γA method errs by predicting that the nonpolar term of PAH molecule solvation should be less favorable with increasing size. Semi-Explicit Assembly correctly captures this non-additivity, and predicts that larger PAH molecules should be more readily hydrated than smaller ones because of the "hot spots" centered above and below the rings.

504 small solute molecules: a variety of molecular shapes

Here, we broaden our comparison to a large diverse test set of solutes. We have calculated ΔG values for the same extensive test set previously studied by Mobley et al.,[93] which is a subset of the molecules explored by Rizzo et al. using various implicit solvent methods.[113] This is a diverse series of compounds that includes a variety of common functional groups in different arrangements.

Figure 2.5a shows the finding of others[93, 113, 128] that γA does not capture the nonpolar or cavity component of the solvation free energies from explicit solvent simulations. The correlation does not improve if a volume term replaces the area term.[93] Figure 2.5b shows that Semi-Explicit Assembly gives much better agreement with the atomically detailed simulations. The correlation coefficient for the latter is 0.91, as compared to 0.15 for the former. The key component in this improvement is accurate calculation of attractive interactions. This correctly lowers the ΔG values for solutes that contain strong attractive elements, like the example cases shown in Figure 2.3.
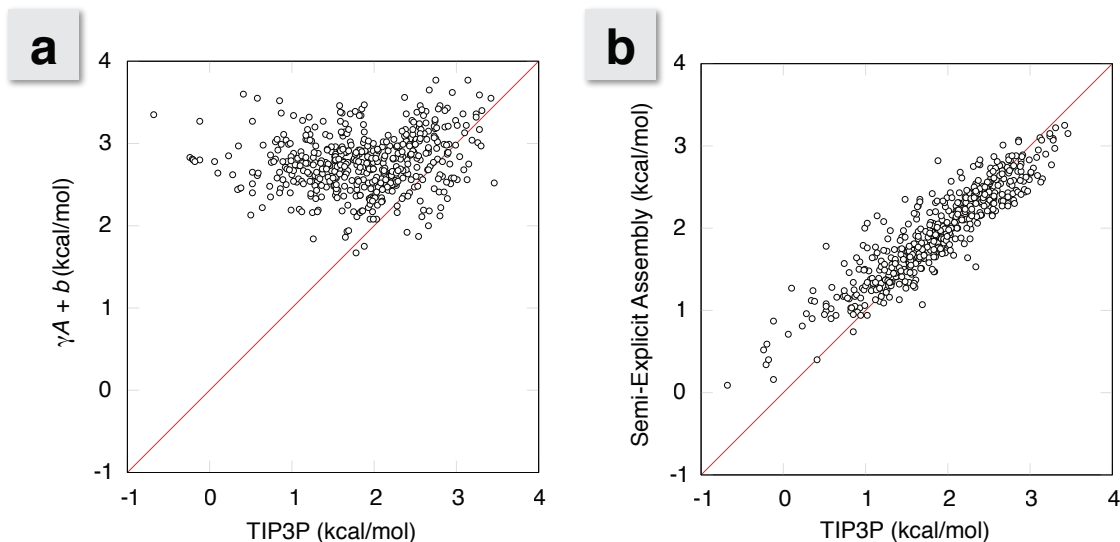
Figure 2.5: Correlation plots of ΔG values comparing a) $\gamma A + b$ and b) our Semi-Explicit Assembly technique with the ΔG values from explicit solvent free energy calculations. A detailed incorporation of dispersion interactions takes what was originally a flat correlation and brings it much more in line with explicit solvent results. This results in a correlation coefficient improvement from 0.15 to 0.91 and an RMS deviation decrease from 1.2 kcal/mol down to 0.3 kcal/mol over the entire set.

Summary

We have described an approach to modeling the solvation free energies of nonpolar solutes in water. We call this approach Semi-Explicit Assembly since its parameters are taken without modification from explicit solvent simulations. The primary computational expense is a pre-computation step in which LJ spheres of various sizes are simulated in explicit water MD calculations. Here, we used the TIP3P water model at 300 K and 1 atm. However, this approach is general and is directly extensible to any explicit-water model, without modification, including to expensive polarizable models for example, and at other temperatures and pressures. These pre-computations intrinsically capture the various structural properties of the sur-

rounding water that are needed to represent solvation free energies. This approach does not require parametrization to large databases of solvation free energies. It goes beyond simpler models in capturing solute shape effects, and not just dependences on solute size. It also goes beyond simpler models in capturing some of the important group non-additivities, but it retains a broader-scale "regional" additivity assumption, so it is nearly as fast to compute as γA methods. Comparisons with explicit solvent simulations of alkynes, branched alkanes, and planar and linear polyaromatic hydrocarbons show that a critical aspect missing from simpler additivity-based models is that some hydrocarbons have hot spots, *i.e.*, regions where one water molecule comes into contact with many carbons at the same time, such as over the centers of aromatic rings. These are regions that contribute to very favorable solvation in water. The results presented show that it is not necessary to sacrifice computational efficiency in order to achieve physically accurate representations of solvation.

2.4 Supplementary Information

Available online at http://pubs.acs.org/doi/suppl/10.1021/ja906399e

Detailed comparisons with enhanced approaches for nonpolar solvation

In the main text, we mentioned comparisons to alternative techniques for treating nonpolar solvation. While these were omitted from the main text, we have included them here for the interested reader. Here we have applied two of the more recent approaches that address the flaws in γA by way of a similar approach, fitting the repulsive and attractive parts of nonpolar solvation separately and recombining these to get an overall picture of the process.[128,136]

Computational Methodology

APBS version 1.0.0 was used for the Wagoner & Baker method calculations.[6] These results were relatively resistant to changes in surface dot density, so those shown are from single calculations at the default density (*sdens*) of 100 points/Å^2. AMBER 10 was used for the Tan, Tan, & Luo method calculations.[19,128] These results are averaged from 5 calculations using dot densities (*maxsph*) of 300, 350, 400 (the AMBER default), 450, and 500 dots per atom. Average errors were 0.04 kcal/mol for weakly attractive molecules (like the *n*-alkanes) and 0.5 kcal/mol for strongly attractive molecules (like the polyaromatic hydrocarbons).
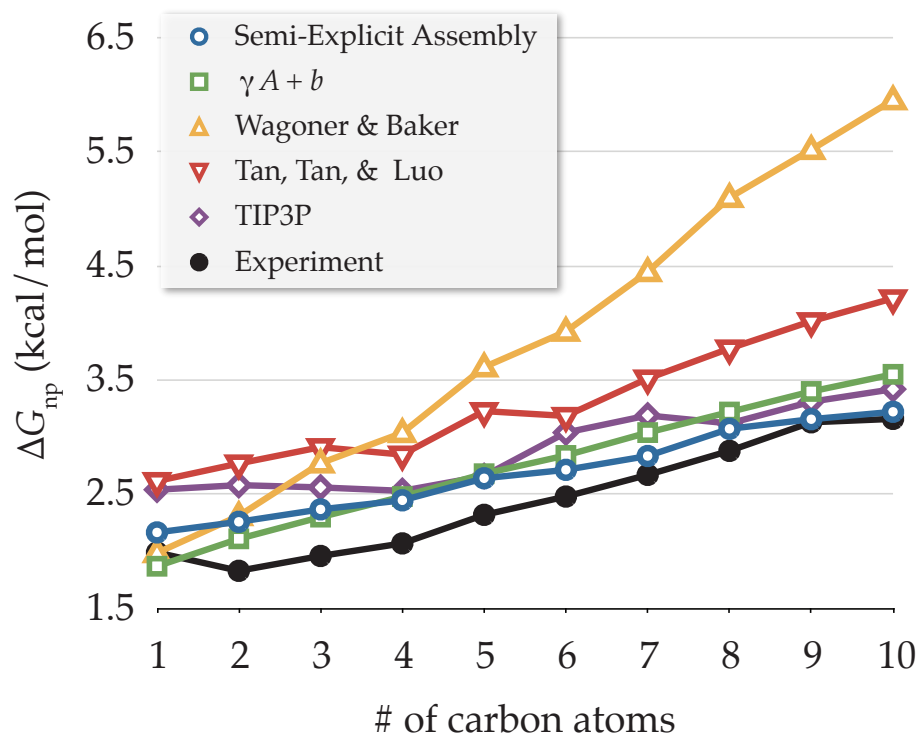
Figure 2.6: The nonpolar transfer free energy for a series of linear alkanes calculated using a variety of different techniques alongside experimental values. See the main paper figure caption and the above Computational Methods for details of the calculation. Note that the y-axis scale has been changed from the main text version to highlight differences in the methods.
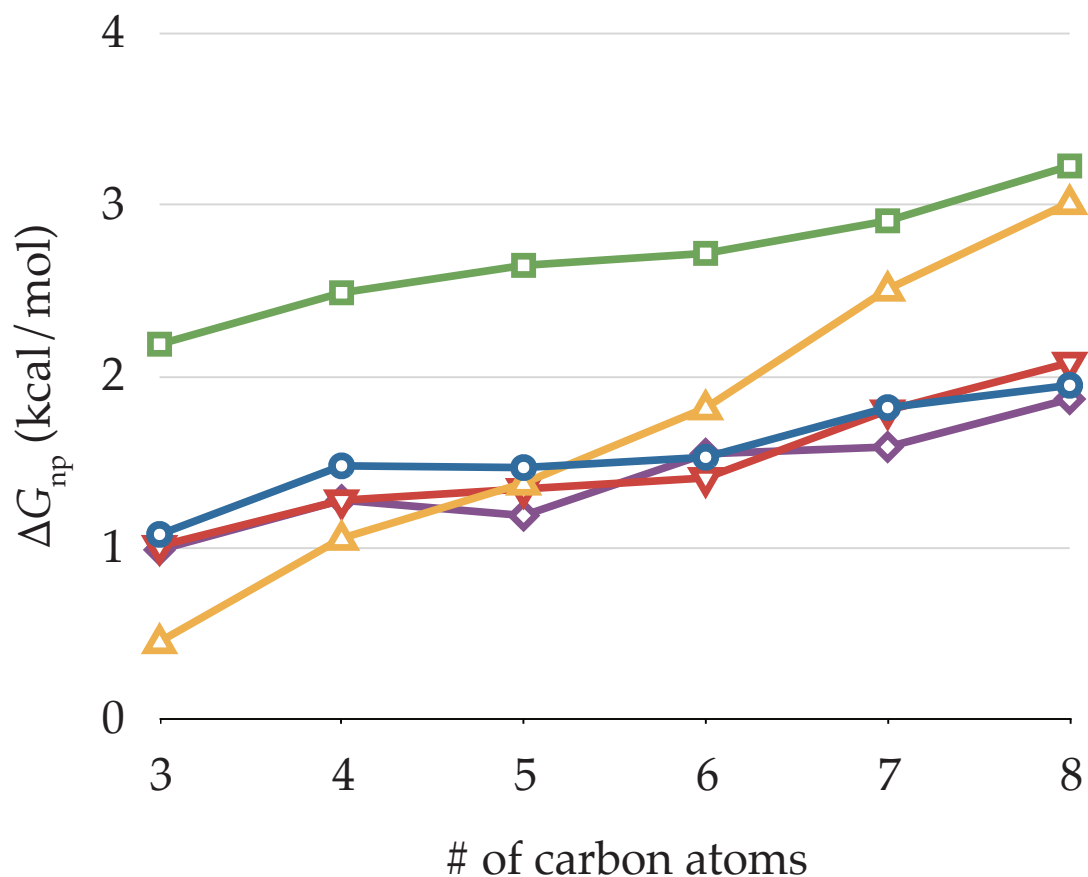
Figure 2.7: The nonpolar transfer free energy for a series of linear alkynes calculated with the same methods used in Fig. 2.6. The triple bond is located at the end of the hydrocarbon chain in all cases. Because the carbon atoms in this bond are given a more attractive dispersion interaction in GAFF,[139] all methods which account for dispersion attraction report lower ΔG_{np} values.
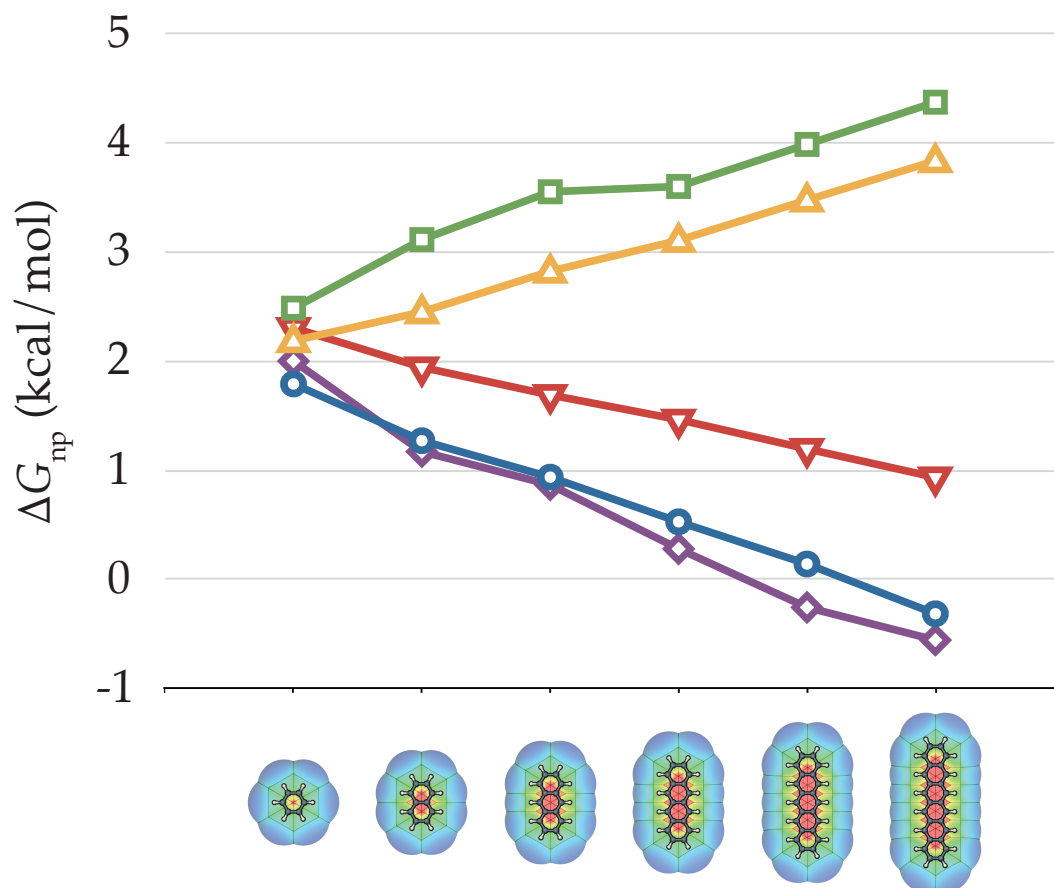
Figure 2.8: The nonpolar transfer free energy for the linear polyaromatic hydrocarbon series ranging from benzene to hexacene calculated using the techniques described previously. In the molecular diagrams along the x-axis, the color indicates relative attractiveness of the molecular LJ potential, with red indicating strongly attractive areas. In explicit solvent calculations, the ΔG_{np} decreases with increasing size, despite the increasing surface area. The more advanced approaches to nonpolar solvation (particularly Semi-Explicit Assembly) do a much better job of capturing this effect than γA.
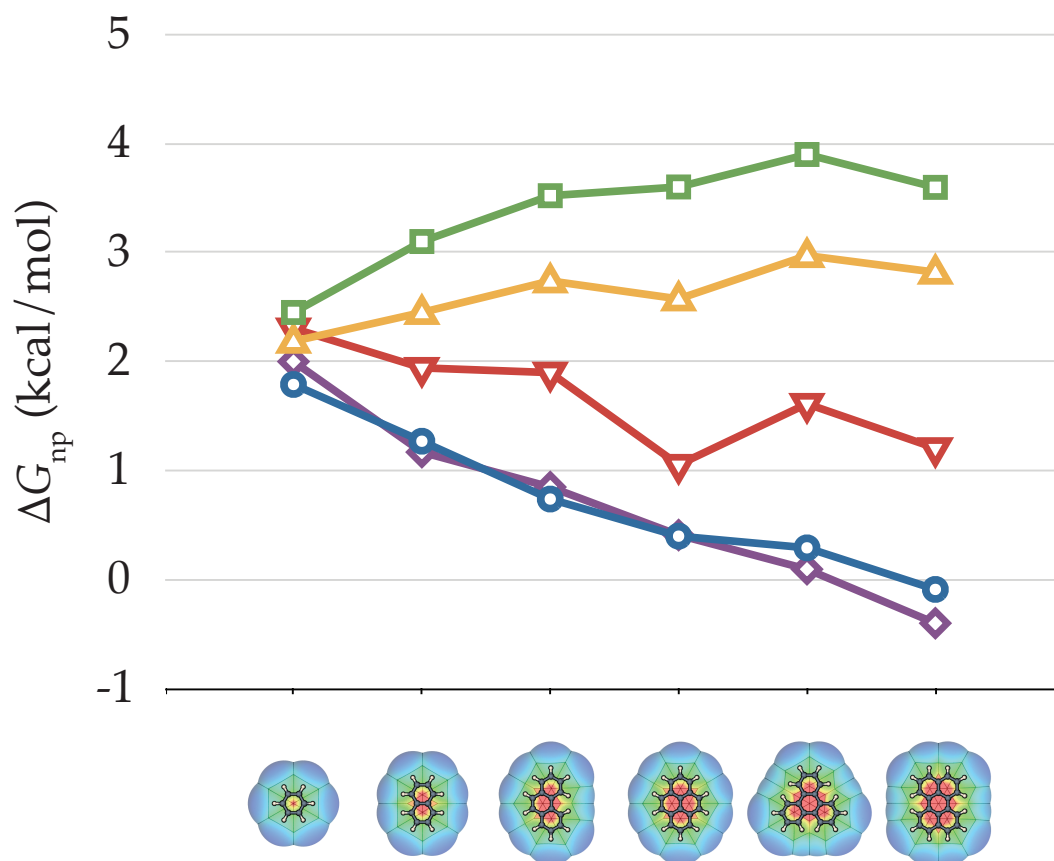
Figure 2.9: The nonpolar transfer free energy for a planar polyaromatic hydrocarbon series ranging from benzene to perylene calculated using the techniques described previoiusly. Like the linear polyaromatic series in Fig. 2.8, the ΔG_{np} decreases with increasing size in explicit solvent, despite the increasing surface area. The more advanced approaches to nonpolar solvation (particularly Semi-Explicit Assembly) do a much better job of capturing this effect than γA.

Comparison results and discussion

The key point that can be taken from these comparisons is that aside from the linear alkanes, these advanced techniques are more successful than γA at incorporating attractive interactions with these solutes. They (like Semi-Explicit Assembly) exhibit more favorable ΔG_{np} values than γA in Figures 2.7, 2.8, and 2.9. In general, the Tan, Tan, & Luo (TTL) method appears to perform better than the Wagoner & Baker (WB) approach, exhibiting slopes closer to explicit solvent and capturing more of the favorable attractive dispersion interactions. This is likely due to the former being fit to a large set of explicit-solvent small molecule transfer free energies, while latter is fit to forces on protein structures.

For further comparisons, we included calculations of the extensive set of 504 small molecules discussed in the main text. Figures 2.10a and 2.10b show the correlation between both the WB and TTL approaches with explicit-solvent respectively. Here, we see that TTL has a tighter correlation with explicit-solvent than WB, resulting in RMSD to explicit-solvent of 0.4 kcal/mol for TTL and 1.0 kcal/mol for WB. Both of these results are less accurate than the 0.3 kcal/mol RMSD exhibited by Semi-Explicit Assembly.

It is important to note that the performance of both WB and TTL could be improved through additional fitting. WB has 3 free fitting parameters, and the optimal form of TTL has 5 free fitting parameters. Achieving the best set of parameters is a
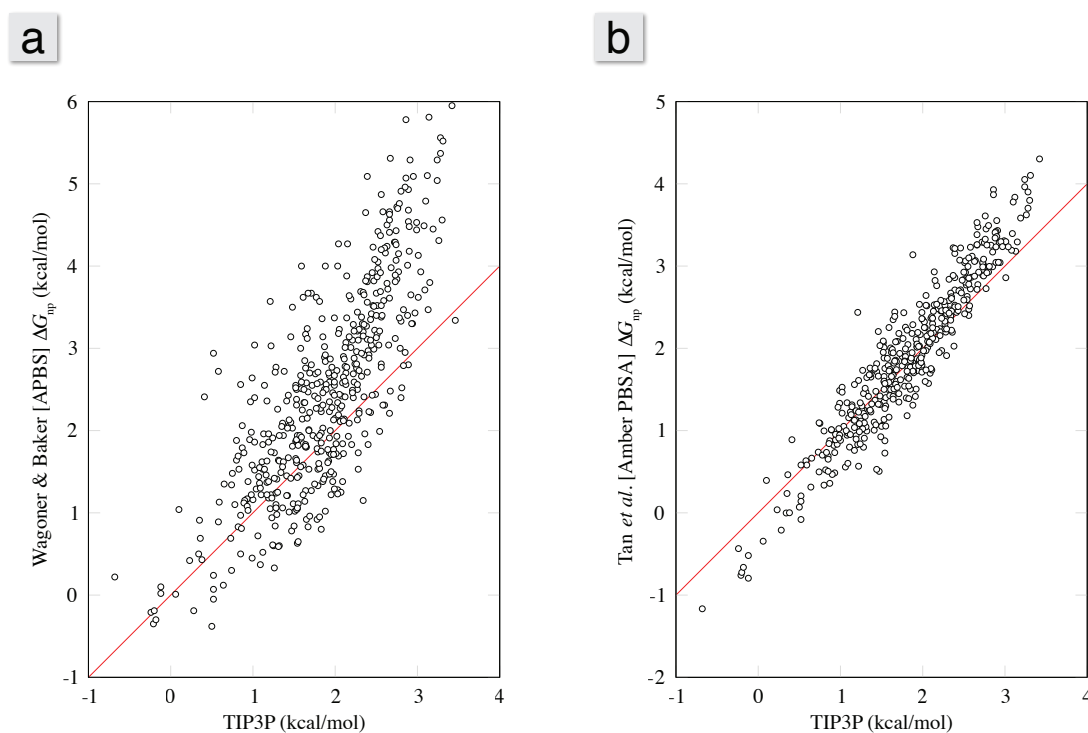
Figure 2.10: The nonpolar transfer free energy for a) the 504 molecule set in Mobley et al.[93] calculated using the Wagoner & Baker method and b) a 445 molecule subset of these molecules calculated using the Tan, Tan, & Luo approach compared to the TIP3P results. The reason for the subset is that 59 of the molecules included atom types outside the allowed parameter space for the 'pbsa' module of AMBER.

challenging task, and it is difficult to insure transferability to molecules outside the optimization set, so it is not surprising that these approaches do not perform as well as Semi-Explicit Assembly in the systems explored here. These two approaches do have the benefit of fewer free parameters than earlier approaches to nonpolar solvation,[28, 46] which assign 2 or 3 free parameters for each atom type. With upwards of 20 available atom types, the parameter space becomes quite large, and the results are more a result of optimal statistical fitting rather than reliance upon the underlying physics. From this perspective, the use of only a single free parameter in Semi-Explicit Assembly is a beneficial attribute. The results obtained using Semi-Explicit Assembly are not as dependent a fitting procedure and are more a result of the actual physics of the microscopic interactions.

Data Tables

Tables 2.1 and 2.2 list the ΔG_{np} and r_{w} values calculated for the series of Lennard-Jones spheres used in this study. The cells of Table 2.1 are colored by the magnitude of the associated ΔG_{np} value: red being less and blue being more favorably solvated.

Table 2.1: Calculated ΔG_{np} values for the base LJ spheres used in this study.

ϵ_{LJ} (kcal/mol)	σ_{LJ} (Å)								
	0.6	1.4	2.2	3.0	3.8	4.6	5.4	6.2	7.0
0.0078125	0.417(3)	0.907(5)	1.66(8)	2.71(1)	4.06(3)	5.69(2)	7.58(4)	9.73(4)	12.09(4)
0.015625	0.412(4)	0.905(5)	1.68(8)	2.73(1)	4.10(2)	5.75(2)	7.59(3)	9.71(4)	12.15(5)
0.03125	0.386(4)	0.878(5)	1.61(9)	2.68(1)	4.01(2)	5.63(3)	7.42(7)	9.57(4)	11.78(8)
0.0625	0.332(4)	0.799(7)	1.51(9)	2.49(1)	3.79(3)	5.26(3)	7.00(4)	8.89(4)	10.85(5)
0.125	0.244(3)	0.632(6)	1.29(8)	2.17(3)	3.26(2)	4.57(3)	6.08(3)	7.65(4)	9.47(5)
0.25	0.088(3)	0.385(6)	0.88(8)	1.59(1)	2.47(2)	3.48(5)	4.63(3)	5.73(5)	6.98(5)
0.5	-0.139(4)	-0.023(7)	0.25(9)	0.67(1)	1.15(4)	1.72(3)	2.16(3)	2.84(5)	3.15(5)
1	-0.504(7)	-0.64(5)	-0.72(1)	-0.75(1)	-0.85(3)	-0.96(3)	-1.32(4)	-1.82(5)	-2.51(9)
2	-1.047(4)	-1.58(6)	-2.18(1)	-2.90(2)	-3.84(2)	-5.11(3)	-6.62(6)	-8.58(6)	-11.06(6)
4	-1.868(6)	-3.00(7)	-4.36(1)	-6.10(2)	-8.30(5)	-11.07(3)	-14.45(4)	-18.62(6)	-23.81(6)

Table 2.2: Extracted r_{w} values for the base LJ spheres used in this study.

ϵ_{LJ} (kcal/mol)	σ_{LJ} (Å)								
	0.6	1.4	2.2	3.0	3.8	4.6	5.4	6.2	7.0
0.0078125	1.98	2.35	2.72	3.08	3.45	3.82	4.19	4.57	4.95
0.015625	2.00	2.37	2.74	3.11	3.49	3.86	4.24	4.61	4.99
0.03125	2.01	2.39	2.76	3.14	3.52	3.90	4.28	4.66	5.04
0.0625	2.02	2.41	2.79	3.17	3.55	3.93	4.31	4.70	5.08
0.125	2.03	2.43	2.81	3.19	3.58	3.96	4.35	4.73	5.12
0.25	2.05	2.45	2.83	3.22	3.61	3.99	4.38	4.77	5.16
0.5	2.07	2.46	2.85	3.24	3.63	4.02	4.41	4.81	5.19
1	2.07	2.48	2.87	3.27	3.66	4.05	4.44	4.84	5.23
2	2.08	2.49	2.89	3.29	3.68	4.08	4.47	4.87	5.26
4	2.10	2.51	2.91	3.31	3.71	4.11	4.50	4.89	5.29

Chapter 3

Modeling aqueous solvation with semi-explicit assembly

Authors: Christopher Fennell, Charles Kehoe, and Ken Dill

Abstract

We describe a computational solvation model called Semi-Explicit Assembly (SEA). SEA water captures much of the physics of explicit-solvent models but with computational speeds approaching those of implicit-solvent models. In this model, we perform explicit-solvent precomputations to create a map of local water solvation behavior around a set of systematically chosen spheres, having different electrostatic and van der Waals parameters. At runtime, the first shell of waters around the full solute molecule is assembled by looking up the properties of first shell waters around the solutes component spheres and distant effects are handled using a solvent continuum. SEA fixes some flaws of implicit-solvent models by incorporating regional additivity to account for curvature and neighbor-interaction effects, by accounting for the discrete nature of microscopic water at a solutes surface, and by treating explicitly waters asymmetrical dipole. SEA does not involve parameter fitting, because the solvent properties are taken directly from the given underlying explicit-solvation model. Extensive tests against five different homologous alkyl series, a set of 504 varied solutes, solutes taken retrospectively from two solvation-prediction events, and a hypothetical polar-solute series show that SEA is about as accurate as explicit solvent simulations, and is about 100-fold faster than Poisson-Boltzmann calculations.

3.1 Introduction

We describe here an approach for computing the free energies of solvation of so-
lutes in water. Aqueous solvation has been modeled at different levels, ranging
from detailed quantum mechanics simulations of few-molecule clusters,[114, 132, 142] to
faster classical simulations of liquid water using up to tens of thousands of ex-
plicit molecules,[25, 30, 56, 78, 80, 84, 92, 93, 98, 99, 117, 118, 134] to very fast models in which water is
treated implicitly as a simple uniform continuous medium.[21, 28, 45, 83, 105, 113, 125, 131] For
large computations, such as those in typical biomolecule simulations, explicit-water
modeling is too slow and expensive, so it is common to use implicit water instead.
However, implicit models require trade-offs in the physics that can limit their accu-
racies:

1. **Continuum.** Water is treated as a continuum, rather than as particulate
   molecules. This can miss important properties of water structuring in first and
   second solvation shells.

2. **Local additivity.** The free energy of solvation is taken to be a sum of free
   energies over solute atoms or small subgroups, whereas, in reality, near-neighbor
   effects and nonadditivities are not uncommon.[32, 85]

3. **Independence of molecular shape.** The nonpolar solvation free energy is
   approximated (in commonly used γA terms) as depending only on the area A
   of solute contact with water, while the free energy can have additional depen-
   dencies on the solute's molecular size or shape.[25, 36, 128, 136]

4. **Water's dipolar asymmetry.** A water molecule is approximately spherical, but it's electrostatic dipole is not coincident with it's center of mass. Water's partial negative charge is on its oxygen, near the center of the molecule, while water's partial positive charges are on its hydrogens, nearer to the surface. Because of this asymmetry, a water molecule is less attracted to a spherical cation than to a spherical anion of identical size.[59,73,112] Implicit-solvation models typically compensate for this effect by an empirical adjustment of cationic radii relative to anionic radii.[73] However, attempting to 'fix' in the solute what is fundamentally a limitation of the water model can lead to non-transferabilites of the parameters in implicit solvent models.

It would be useful to have a computational model of water that is both fast – approaching the speeds of the fastest implicit-solvent models – and that captures the physics and the transferability of explicit-solvent models. Many researchers have contributed to this goal by adding polarization or discretization effects to implicit-solvent models,[42,90,107] developing faster or coarse–grained explicit solvents,[17,20,38,61,81,86] and introducing hybrid models that mix explicit and implicit solvents together.[7,60,69,77,82,141] Here, we take a different approach. We pre-compute solvation properties of water in explicit-solvent simulations that we then capture as approximate one-body terms, which we then sum over assemblies of a solute's component spheres. We call this approach Semi-Explicit Assembly, or SEA water.

3.2 Theory

In the SEA approach, we assume that the free energy of interaction, ΔG, of a solute molecule with its solvating water molecules, is a sum of three terms,

$$\Delta G = \Delta G_{\mathrm{np}} + \Delta G_{\mathrm{pol,surf}} + \Delta G_{\mathrm{pol,bulk}}. \tag{3.1}$$

ΔG_{np} is the free energy of forming the cavity the solute occupies in water and includes the dispersion interactions between the solute and the waters. For purely nonpolar solutes, such as simple hydrocarbons, ΔG_{np} is the only substantial contribution to ΔG. For solute molecules that are polar or charged, the total solvation free energy also has two additional terms (see Equation 3.1): $\Delta G_{\mathrm{pol,surf}}$ describes the electrostatic interactions of the solute with the immediate-neighbor first-shell water molecules that surround it, and $\Delta G_{\mathrm{pol,bulk}}$ describes the electrostatic interactions of the solute with the water molecules that are more distant than the first solvation shell.

For the nonpolar component ΔG_{np}, the SEA approach is described in detail elsewhere.[36] In the present paper, we describe the SEA treatment of the two polar terms. We compute these polar components using two steps: (1) a slow series of pre-simulations for a given solvent and (2) a fast summation of component free energies for a given solute.

SEA Step 1: Pre-calculations on solute spheres in explicit water

SEA treats a solute molecule as a collection of 'atomic' spheres of different types and sizes (*i.e.* different Lennard-Jones parameters and different partial charges). The first step in SEA modeling is to pre-calculate the positions and orientations of explicit-water molecules around those component building-block spheres. SEA can use any given explicit-solvent model. Here, to illustrate the principle, we use the TIP3P water model.[65]

We start by solvating a series of individual spheres in baths of explicit water. Spheres span a range of different sizes and Lennard-Jones parameters (6 ϵ values and 6 σ values) and a range of 11 different partial charges. In total, we do pre-computations on 396 different types of spheres in TIP3P water. Each sphere is simulated for 10 ns at the desired state point, 1 atm and 300 K. From each such simulation, we harvest the statistical properties of water distances and orientations around the sphere. For each type of sphere, we calculate the average distance between the solute and the water (the peak of the solute to water radial distribution function). Calculating radial distribution functions for water around these spheres, we found that the width of the first hydration shell extends out 1 to 3 Å from the van der Waals surface of the spheres, depending on the solute charge. We use a 2 Å cutoff distance from the surface of each solute to define the 'first shell' (Fig. 3.1a) to insure we include most to all the first shell waters without including any 'second shell' water population, which begins to grow beyond this distance around the more highly charged solutes.

Within this region, we calculate the average number of water molecules and generate nearest–neighbor distribution functions. These distribution functions tell us the average distance between first shell waters about a given solute sphere.

Because many of these solute spheres have electrostatic charges, the resultant solvating first-shell water dipoles will have a distribution of orientations with respect to the sphere center. To determine the dipole orientation of an average water molecule, we first align all of the first-shell water molecules onto a common axis normal to the surface of the solute sphere (Fig. 3.1b). The solvent atom orientations with respect to this axis are then binned (Fig. 3.1c). Fig. 3.2 shows a series of these solvent-orientation maps around spheres having charges ranging from -1 to +1. Not unexpectedly, it shows that water dipoles orient most strongly around spheres having the strongest electrostatic potentials, *i.e.* around spheres having the largest absolute charges. The figure shows water's average dipole along the common axis; in the supplementary material, we also discuss a more detailed approximation based instead on water's projected quadrupole moment.

Fig. 3.3 shows that the average first shell water dipole moment has a simple functional relationship *vs.* the strength of the electrostatic field from the solute sphere. This universal relationship is useful because it allows us to replace a simulation lookup with a simple sigmoidal functional form. These sigmoidal functions:

$$f(x) = \frac{1}{c_0 + \exp(c_1 \cdot x)} + c_2, \tag{3.2}$$

46
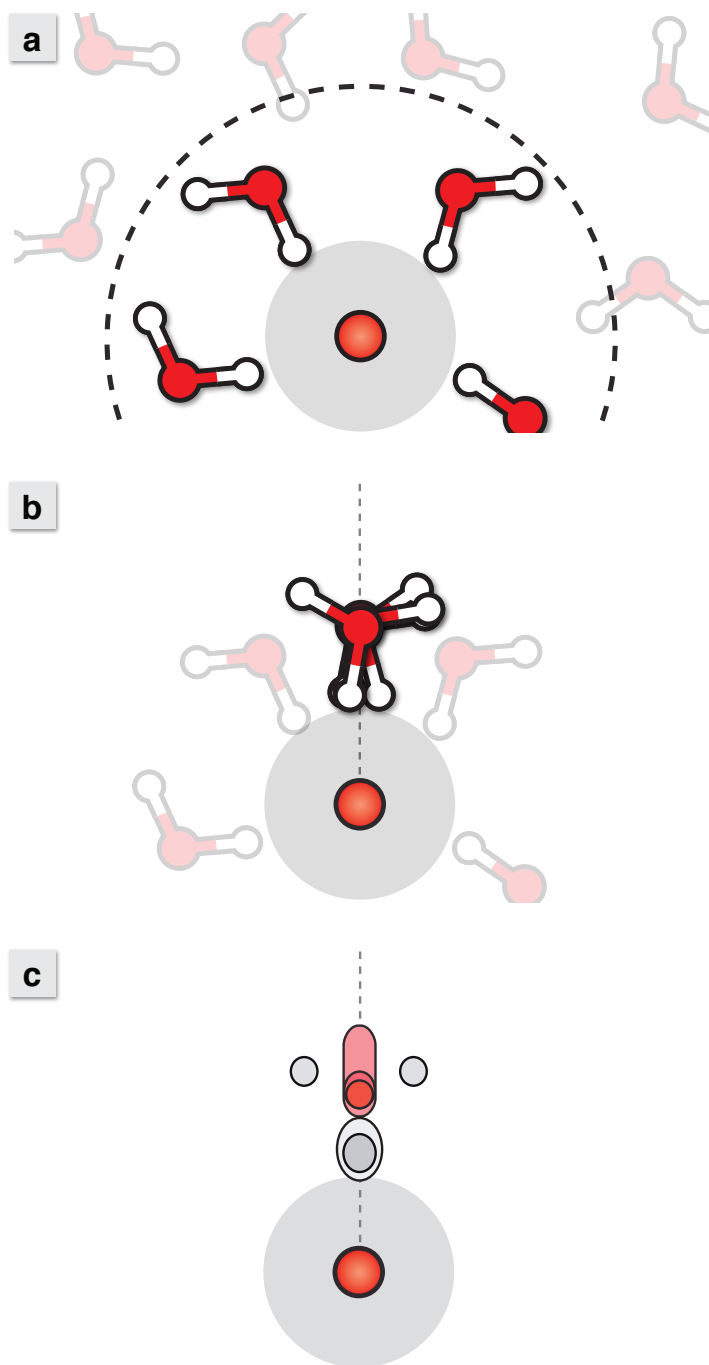
Figure 3.1: The pre-computation step. (a) Simulate the waters around a sphere. (b) Superimpose all the first-shell waters onto a common axis relative to the solute-sphere center. (c) Create an atomic density map.
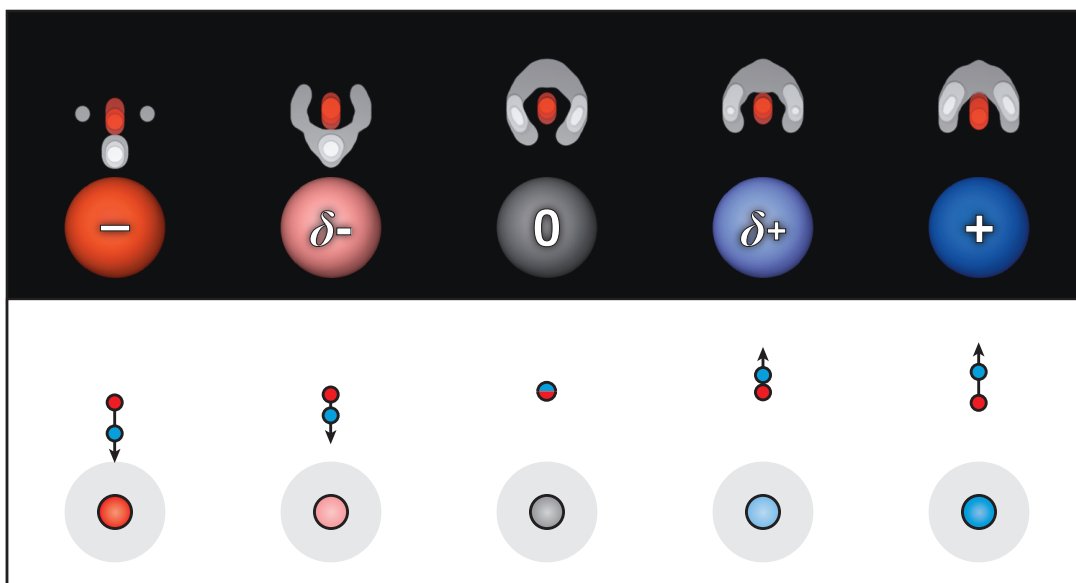
Figure 3.2: Solute spheres having different charges lead to different distributions of first-shell water hydrogen and oxygen atoms, and therefore to different degrees to which water's dipole is oriented relative to the common axis.

have coefficients c_0, c_1, and c_2 which come from separate non-linear curve fits of the positive and negative electrostatic field regions of Fig. 3.3. These component curves have different shapes and limits because of the asymmetry of water's dipole moment.[25,59,73,92,112] Around positive spheres, the limit of the sigmoidal curve coincides with the dipole moment of the TIP3P water model since water's dipole moment points normal to the solute surface in the optimal electrostricted configuration. Around negative spheres, water points one hydrogen towards the solute center in the optimal electrostricted configuration (see Fig. 3.2) and water's dipole is no longer normal to the surface, resulting in a negative curve limit that is smaller than the TIP3P dipole moment.
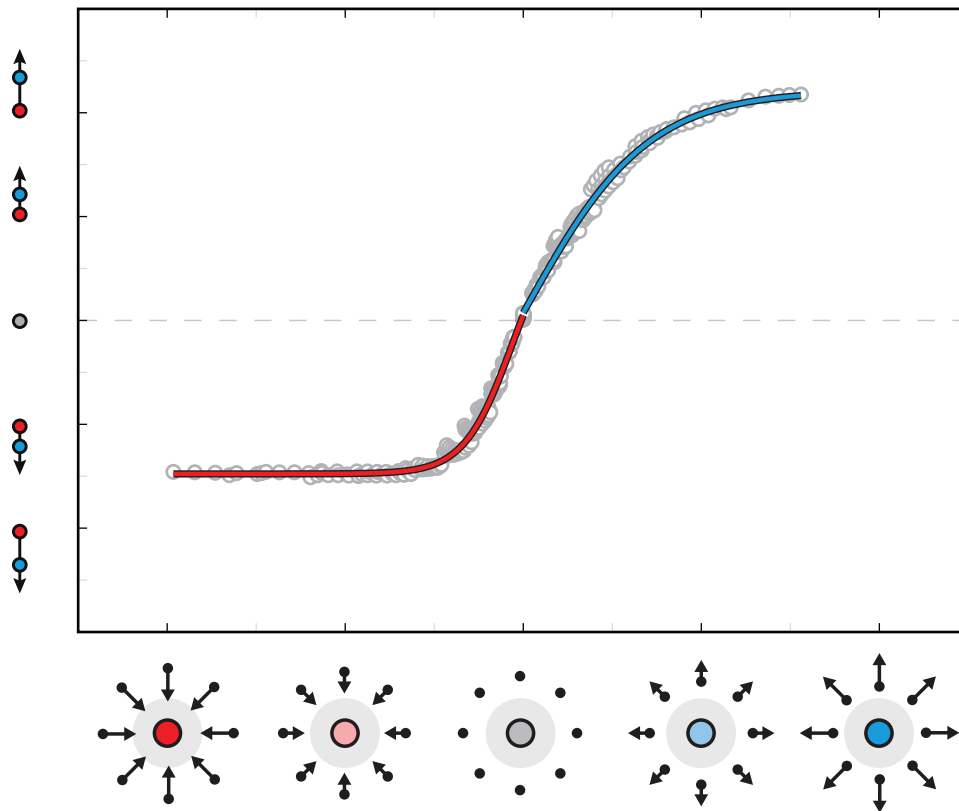
Figure 3.3: The TIP3P water dipole free energy as a function of electric field strength emanating from spheres of different charge and size. The red and blue lines are sigmoidal fits to the data for dipoles pointing toward and away from the surface of the spherical solute respectively.
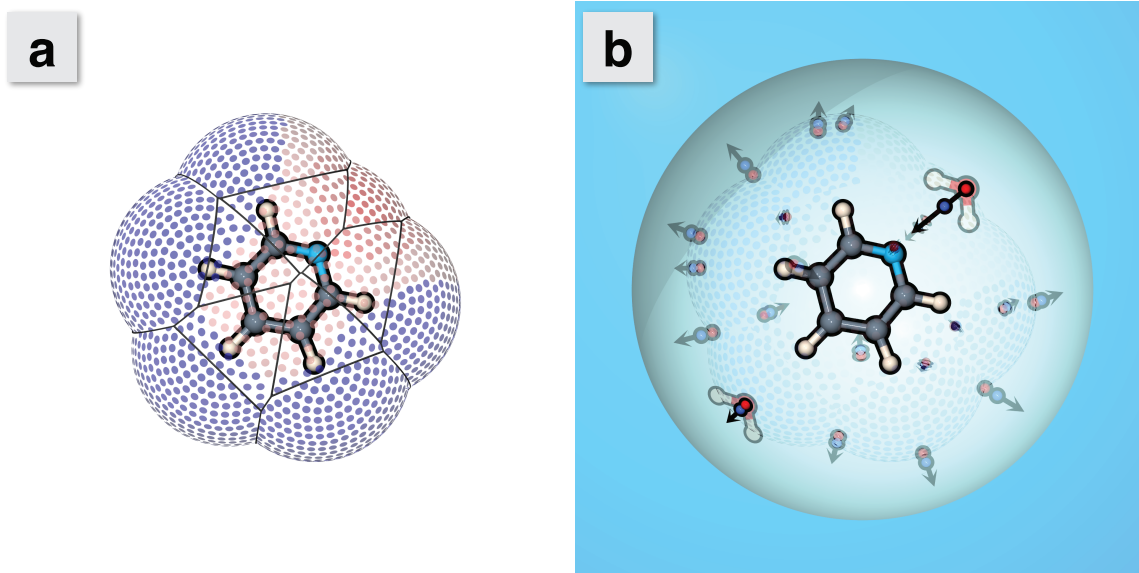
Figure 3.4: The SEA sampling process around pyridine. (a) A solvent-accessible dot surface is constructed using pre-computed water-sphere average distances. (b) Sites are selected on this surface, and dipoles are placed at these sites according to the electric field. Note that sites where the electrostatic field is stronger (*i.e.* near the nitrogen) have longer dipoles than weaker sites, as indicated by the two opaque dipoles. Interactions are accumulated for all the point charges in the cavity to get ΔG_{surf}, and the total dipole in the molecular cavity is used to determine a reaction field term for ΔG_{bulk}. This sampling process is iterated to converge on ΔG_{pol}.

SEA Step 2: Assembling the solvent shell around an arbitrary molecule

Once the pre-computations have been performed, the SEA water model can be used for the rapid computation of the hydration free energy of a given solute. The first step is to determine the surface polarization term, $\Delta G_{\text{pol,surf}}$. This requires placement of surface water dipoles in physically representative locations and orientations. We do this by first generating a solvent accessible dot surface about the solute of interest (Fig. 3.4a). Rather than using the traditional route of 'rolling a hard sphere' of some fixed size over the molecule to generate a Lee–Richards or Connelly surface,[26,74] we

use instead our pre-computed average separations between a water molecule and the sphere center. This strategy captures various physical aspects of solvation, such as the fact that water molecules are held more tightly to solutes having higher charge than to solutes having lower charge.

We select a solvation site at random on this surface and compute the electric field at this point due to all the solute atoms using

$$\mathbf{E} = \sum_{i=1}^{N} \frac{1}{4\pi\varepsilon_0} \frac{q_i}{r_{ij}^2} \hat{\mathbf{r}}_{ij}, \tag{3.3}$$

where q_i is the charge on solute atom i and r_{ij} is the distance between site j and solute atom i. Using the electric field association functions (Fig. 3.3), the appropriate dipole (or quadrupole) is placed at this site along the electric field line. After placing this dipole, surface water sites that are too close are eliminated. The random placement procedure then continues until all possible solvation sites are occupied or eliminated. Fig. 3.4b shows one such configuration about pyridine, where the dipoles are uniformly elongated for clarity. Note that magnitude and direction of the solute electric field is visible through the nature of the placed dipoles. The dipoles are long and point towards the nitrogen atom as it has a large negative partial charge, while the dipoles around the hydrogen atoms are short and point away from the solute surface.

After generating a solvent configuration around a solute, the surface interactions are accumulated into $\Delta G_{\mathrm{pol,surf}}$ using

$$\Delta G_{\mathrm{pol,surf}} = \Delta G_{\mathrm{sol-solv}} + \Delta G_{\mathrm{solv-solv}}, \tag{3.4}$$

where *sol* refers to solute particles and *solv* refers to solvent particles. For a dipolar solvent representation, this can be explicitly written in terms of Coulombic sums as

$$\Delta G_{\mathrm{pol,surf}} = \frac{1}{4\pi\varepsilon_0} \left(\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{q_i q_j}{r_{ij}} + \sum_{i=1,3,5,\ldots}^{M-3} \sum_{j>i+1}^{M} \frac{q_i q_j}{r_{ij}} + \sum_{i=2,4,6,\ldots}^{M-2} \sum_{j>i}^{M} \frac{q_i q_j}{r_{ij}} \right), \tag{3.5}$$

where N is the number of solute atoms and M is the number of solvent partial charges. Equation 3.5 gives a free energy, rather than a potential energy, because the positional and orientational entropies of the water molecules are captured by the pre-computations.

To estimate the total polar component of solvation, we need the first–shell/surface component described above and the bulk component of the electrostatic free energy of solvation. We estimate this quantity using the Onsager reaction field,[5, 104]

$$\Delta G_{\mathrm{pol,bulk}} = \frac{\varepsilon - 1}{2\varepsilon + 1} \cdot \frac{\boldsymbol{\mu}^2}{r_{\mathrm{c}}^3}, \tag{3.6}$$

where $\boldsymbol{\mu}$ is the dipole moment vector of the solute cavity of radius r_{c} and ε is the dielectric constant of the continuum. Here, r_{c} extends from the geometric center of the solute out to the outer boundary of the furthest semi-explicit shell (see the

spherical cavity in Fig. 3.4b). We approximate the cutoff for this reaction field as if it were spherical. This type of reaction field should give a good approximation for neutral polar molecules, which are our primary focus here, but for solutes having a formal net charge, an additional Born term will be needed.[2, 29, 67, 69, 141]

The procedure described above is based on one particular placement of water molecules in the first solvation shell. Clearly, any single such configuration may not be fully representative. So, we repeat the random dipole placement five times to estimate the variance of the ΔG_{pol} as a function of water placement. This variance tells us how many additional sampling iterations are needed to converge an averaged ΔG to a specified standard error.

Finally, we accumulate these solvation-shell snapshots into a Boltzmann-weighted average. We calculate the probability, p_i, of shell-configuration i using,[41]

$$p_i = \frac{e^{-\Delta G_i/k_B T}}{\sum\limits_{j=1}^{N} e^{-\Delta G_j/k_B T}}, \tag{3.7}$$

where N is the total number of solvation configuration samples. The average free energy is the weighted sum over the different shell configurations,

$$\langle \Delta G \rangle = \sum_{i=1}^{N} \Delta G_i \cdot p_i. \tag{3.8}$$

3.3   Computational Methods

Explicit Simulations

The simulations of model spheres were performed using GROMACS 4.0.4.[11,55]   In these calculations, statistics for four spheres were accumulated simultaneously by bonding them in a tetrahedral geometry with 20 Å edge lengths between each particle. The charges of these four particles were assigned such that each tetrahedron is net neutral, and they included values ranging from -1 to +1 in increments of 0.2. Separate simulations were performed for different sets of Lennard–Jones parameters. Model sphere diameters included $\sigma = 1.4$, 2.2, 3.0, 3.8, 4.6, and 5.4 Å, while well–depths included $\epsilon = 0.015625$, 0.03125, 0.0625, 0.125, 0.25, and 0.5 kcal/mol. Each of the resulting 108 tetrahedral arrangements was solvated in a cubic box of TIP3P water with a 12 Å buffer of water between any solute particle and face of the simulation cell.

Each solvated tetrahedron was simulated under periodic boundary conditions in the NPT ensemble at 300 K and 1 atm using the Nose-Hoover thermostat and Parrinello-Rahman barostat with time constants of 1 and 10 ps respectively. The dynamics were propagated with the leapfrog integrator with a timestep of 2 fs. The Lennard-Jones interactions were switched off between 8 and 9 Å, with long-ranged energy and pressure terms included. Smooth PME[34] was used for electrostatics with a real-space

cutoff of 10 Å, spline order of 4, Fourier spacing of 1.2 Å, and relative energy tolerance of 10^{-5}. Trajectory snapshots were saved at 1 ps intervals for the post–processing described in the main text.

Details of the implementation

Solvent-accessible surfaces for SEA calculations were constructed as described in our previous work.[36] The surface detail was set to approximately 80 surface points per solute atom. A minimum of 5 Semi–Explicit solvation configurations were sampled, and this iteration process was continued until the accumulated ΔG_{pol} converged with an error of ~0.05 kcal/mol.

Force fields and conformation sampling

The General Amber Force Field (GAFF)[139] was used to assign Lennard-Jones parameters for all of the small molecules. AM1-BCC partial charges[63] were assigned to initial minimized structures using the ANTECHAMBER program in AMBER.[140] As these structures may not be the ideal representative structures of the molecules in solution, we generated short conformation ensembles for each from 500 ps of MD in explicit solvent using GROMACS 4.0.[11,55] The results were clustered to obtain an average representation of the small molecule, and the partial charges were reassigned using this averaged conformation.

In the cases where conformation plays a role in the calculated transfer free energies, an ensemble of conformations is necessary to obtain an accurate estimation. We generated solute conformation ensembles from 5 ns explicit solvent MD simulations with 2, 20, and 200 ps snapshot intervals for all small molecules investigated in this study. The various conformation sets all gave results that overlapped within error, so we used the ensembles generated from the MD simulations with 200 ps snapshot intervals (25 conformations) for maximum efficiency. The free energies reported for each solute from SEA and the implicit solvent models come from averages over these 25 conformations.

Implicit-solvent calculations

Implicit solvent results were obtained for the Poisson–Boltzmann (PB) and generalized Born (GB) methods using version 1.0 of the Adaptive Poisson–Boltzmann Solver (APBS)[6] and the OBC implementation of GB in AMBER 10,[19,105] respectively. For the APBS calculations, the nonpolar term was computed using the APBS dispersion incorporation method.[136] The PB equation was solved on a 65 x 65 x 65 grid with 0.25 Å spacing to determine the polar component of solvation. In the GB calculations, γA was used for the nonpolar term with a γ of 5 cal mol^{-1} Å^{-2}.

3.4 Predictions and comparisons with experiments

We now compare the SEA water model to experiments, to explicit-solvent simulations, and to PB and GB models of solvation.

(1) Comparing SEA to other methods and to experiments on homologous series of alkyl chains terminated by different functional groups

Fig. 3.5 shows calculated solvation free energies, ΔG, for 5 series of alkyl chains terminated by common organic functional groups. The figures compare SEA to experiments, TIP3P simulations, and GB. First, SEA gives good agreement with the TIP3P results, within 1 kcal/mol in all cases. Because the parameters used in SEA water modeling are identical to those of TIP3P, this shows that SEA's sampling and regional additivity approximation are not degrading the quality of predictions relative to the much more expensive explicit-solvent simulations.

Second, while TIP3P and SEA both give good estimates of experimental ΔG values for acetates, aldehydes, and phenyl groups, both methods predict solvation free energies for the alcohols and amines that are too positive. This indicates that the combination of GAFF and the AM1-BCC partial charges used in the underlying explicit-solvation model needs improvement for these functional groups. Such errors have been noted before in the explicit-solvent models.[84, 93, 134]
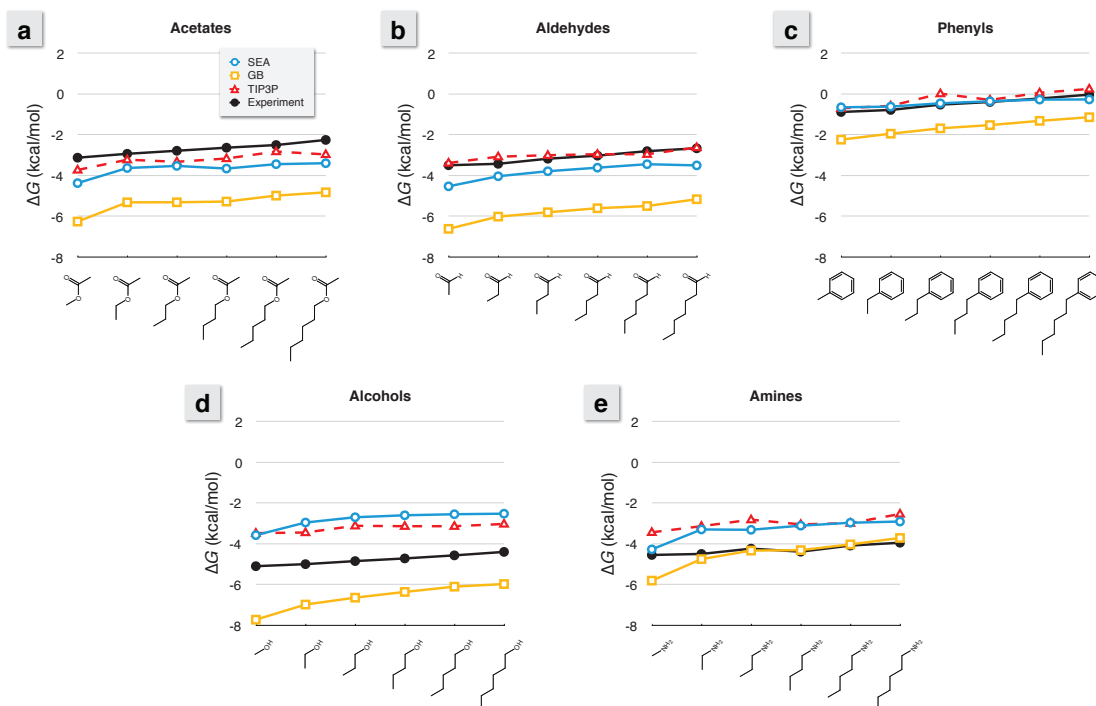
Figure 3.5: Functional group comparisons for the linear alkyl series of a) acetates, b) aldehydes, c) phenyls, d) alcohols, and e) amines. Experimental results come from Ref. 4 and the TIP3P results from Ref. 93. The GB results are from Amber 10 using iGB5 and γA with $\gamma = 5$ cal mol^{-1} Å^{-2}.[105]

Third, we find that GB gives free energies of solvation that are too negative in all cases except for the amines, where it matches experiments quite accurately. The errors in GB can be partly corrected by using the nonpolar SEA term in place of the standard γA term, resulting in a general 0.2–0.4 kcal/mol improvement for these molecules.

(2) Comparing SEA to other methods on a diverse set of 504 small molecules

Next, we explore a larger and more diverse set, containing 504 different small molecules.[15,93,113] This has become a canonical test set for solvation free-energy modeling because of the availability of both experimental data and of extensive TIP3P explicit-solvent simulation results.[93]

Fig. 3.6 compares GB, PB, and SEA with the TIP3P simulation results. Taking TIP3P as the 'gold standard' for this test, the figure shows systematic improvement from GB to PB to SEA. Again, as in the homologous series tests above, SEA is within 1 kcal/mol RMSE relative to TIP3P. Thus, while there is a large gain in computational efficiency from TIP3P explicit-solvent simulations to SEA, there is little loss in predictive accuracy for computing solvation free energies.
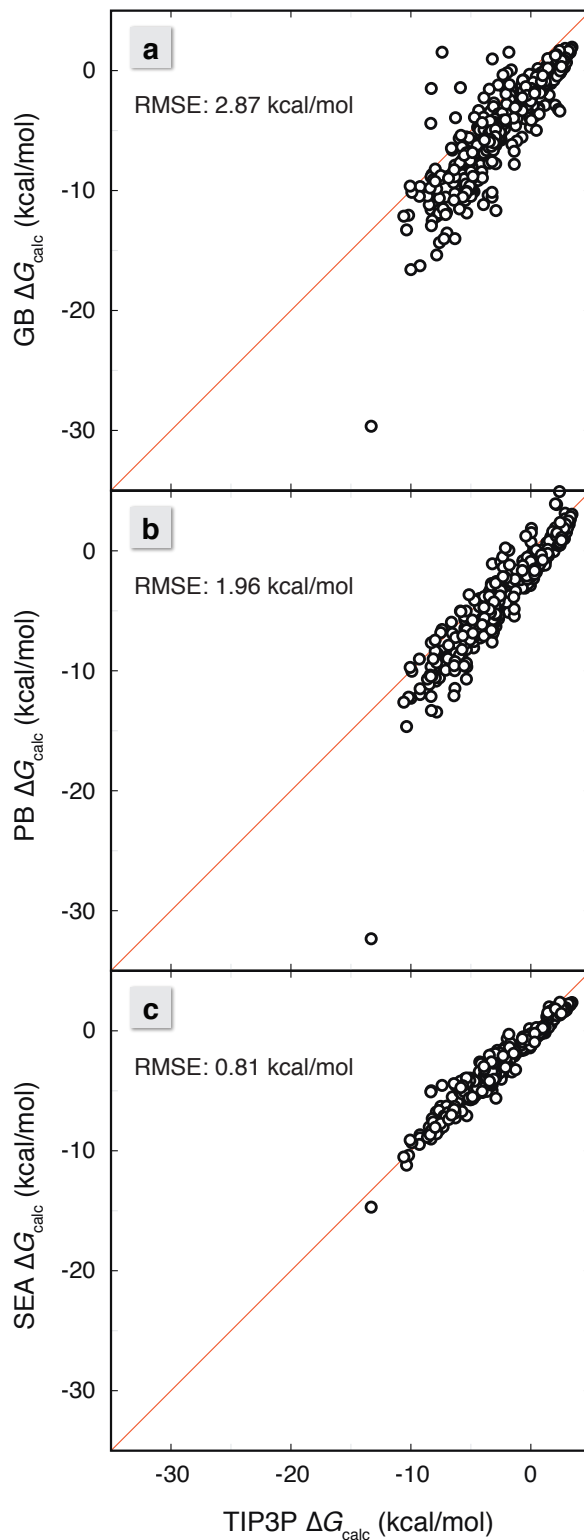
Figure 3.6: Scatter plots of a) GB,[105] b) PB,[6] and c) SEA air-water ΔG transfer values against TIP3P calculations.[93] If the single largest outlier (triacetylglycerol) in the GB and PB calculations is removed, the RMSEs decrease to 2.78 and 1.77 kcal/mol respectively.

The PB and GB implicit-solvent models tend to solvate many of these molecules more favorably than SEA or explicit solvent. There are some exceptions where GB returns significantly less favorable solvation free energies than TIP3P, and these are nearly all molecules with buried amines, like piperazine and triethylamine. Here, the strongly charged nitrogen atoms have limited solvent exposure and do not fully contribute to the overall solvation favorability. The balance of Born radii in such groups has been recognized as a reason behind overly strong salt–bridges in protein simulations using GB, and suggested corrections have included resizing the radii of atoms attached to these nitrogens to fit to explicit solvent energetics.[48] Both PB and SEA appear to avoid suffering from this issue.

Computational performance of SEA vs. TIP3P, PB and GB

Fig. 3.7 compares the performance (computational speed per ligand *vs.* the RMSE free-energy accuracy to experiment) of the various computational methods applied to the test set of 504 small molecules described above. While there are many different implicit-solvent implementations available, our test here is only intended as a general sketch, not an exhaustive and detailed comparison. Briefly, TIP3P is about 10^8-fold slower than GB (the fastest method) in calculating solvation free energies, and that the RMSE errors in GB are about 2.8 kcal/mol while the RMSE errors in TIP3P are about 1.2 kcal/mol. SEA provides an excellent compromise: while $\sim$ 10-fold slower than GB, it is faster than PB and the RMSE errors are smaller than either, about 1.3 kcal/mol.
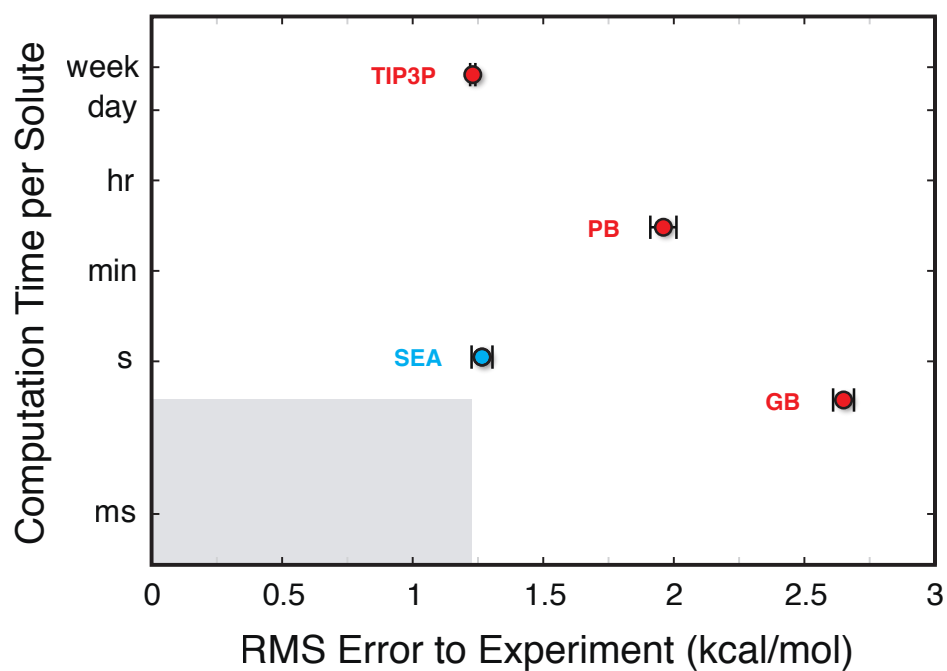
Figure 3.7: The average execution time per solute versus the RMSE to experiment for calculations on the 504 small molecule solute set. Note that the vertical axis is a logarithmic scale. The light-gray region indicates the region of performance faster than GB and more accurate than TIP3P, a target realm for future solvent models.

One advantage of the TIP3P simulations is that they automatically generate multiple conformers of the solute, and given sufficient sampling will generate a properly weighted conformational ensemble. However, most of the small molecules in this test set have few rotatable bonds, leading conformations to only play a minor role in solvation of these solutes. This fact has been used to justify single conformation calculations on larger sets.[15,113] We decided to consider solute conformation effects in Fig. 3.7, and the error bars on the PB, GB, and SEA points show the average standard error per solute over 25 solute conformations. These errors indicate that the average standard deviation of ΔG for each of these methods is 0.2–0.3 kcal/mol. This error comes primarily from solutes with polar groups at different ends of rotatable bonds, like diols. These tend to be some of the most strongly hydrated solutes, so while fluctuation magnitudes can be greater than $k_\mathrm{B}T$ with different conformations of these molecules, the fluctuations tend to be less than 20% of the total solvation free energy.

There are a variety of methods available for generating solute conformation ensembles for implicit solvents, from specialized programs like OpenEye's Omega to general molecular mechanics simulations. Generating a weighted ensemble of conformations often takes longer than individual implicit solvent calculations on a given conformer, and this is certainly the case for the present approach of extracting regular conformations from a 5 ns explicit-solvent simulation. As generating conformations would dominate the computation time, we would not be able to distinguish between the computational costs of PB, GB, and SEA, so we neglect the time preparing conformations in Fig. 3.7.

Random configuration sampling vs. Boltzmann weighting

The solvation shell sampling process in SEA is initially random, and we re-weight the average ΔG according to equation 3.8. How much of an effect does this have on the final energies? We compared the random average and Boltzmann-weighted average ΔG values for the 504 small molecule set and found the error to TIP3P and to experimental results to be identical. Such is expected for non-polar molecules, where the first–shell water dipoles are weak and contribute little to favorable solvation, so we ordered the results by their dipole moment as a measure of relative polarity. There was no observable difference between these averages for any arbitrary subset of polar molecules. These results indicate that the first–shell assembly procedure is successfully producing thermally relevant solvent configurations simply using steric exclusion of neighboring solvation sites.

(3) Retrospective analysis of SAMPL blind-test solvation predictions

Recently, a community-wide blind prediction event for small-molecule solvation free energies, called SAMPL, has been run by OpenEye Software.[50,99] The SEA method that we describe in this paper has not yet been tested in that event. (We tested an incomplete version without water dipoles or solute conformational sampling in SAMP09). Here, as a second-best alternative to a blind test, we retrospectively test the present SEA approach predictions on the solute molecules from both prior SAMPL events.

As we found in the other two comparisons above, the SEA model performs at about the same level of accuracy as the TIP3P model on which our model is based. Fig. 3.8a shows the calculated ΔG versus experiment for the 56 SAMPL-1 molecules. The resulting scatter is similar to that observed from explicit solvent simulations.[96] The RMSE of SEA $vs.$ experiments is 4.1 kcal/mol, compared to 3.8 kcal/mol for TIP3P, using AM1-BCC partial charges. Both SEA and explicit solvent have the same problems: they don't handle well molecules containing sulfur or phosphorus, a difficulty that is likely due to the GAFF force field parameters for these atom types.[96]

Fig. 3.8b shows 23 SAMPL09 molecules, and the RMSE for this small set is under 2 kcal/mol. The two largest outliers are d-xylose and d-glucose. For these sugars, SEA predicts solvation free energies that are 4 and 5 kcal/mol less favorable than experiments. Taken together with the alcohol comparisons in Fig. 3.5d, it suggests a problem in the force field parameters for hydroxyl groups. While it would be unfair to compare our current retrospective results with the prospective tests made by other groups in SAMPL, nevertheless the present tests indicate that SEA is comparable to the best current prediction methods. Because the SEA method does not involve the fitting of free parameters, it is reasonable to expect that SEA should perform with about the same RMSE accuracy as TIP3P simulations in future such tests.
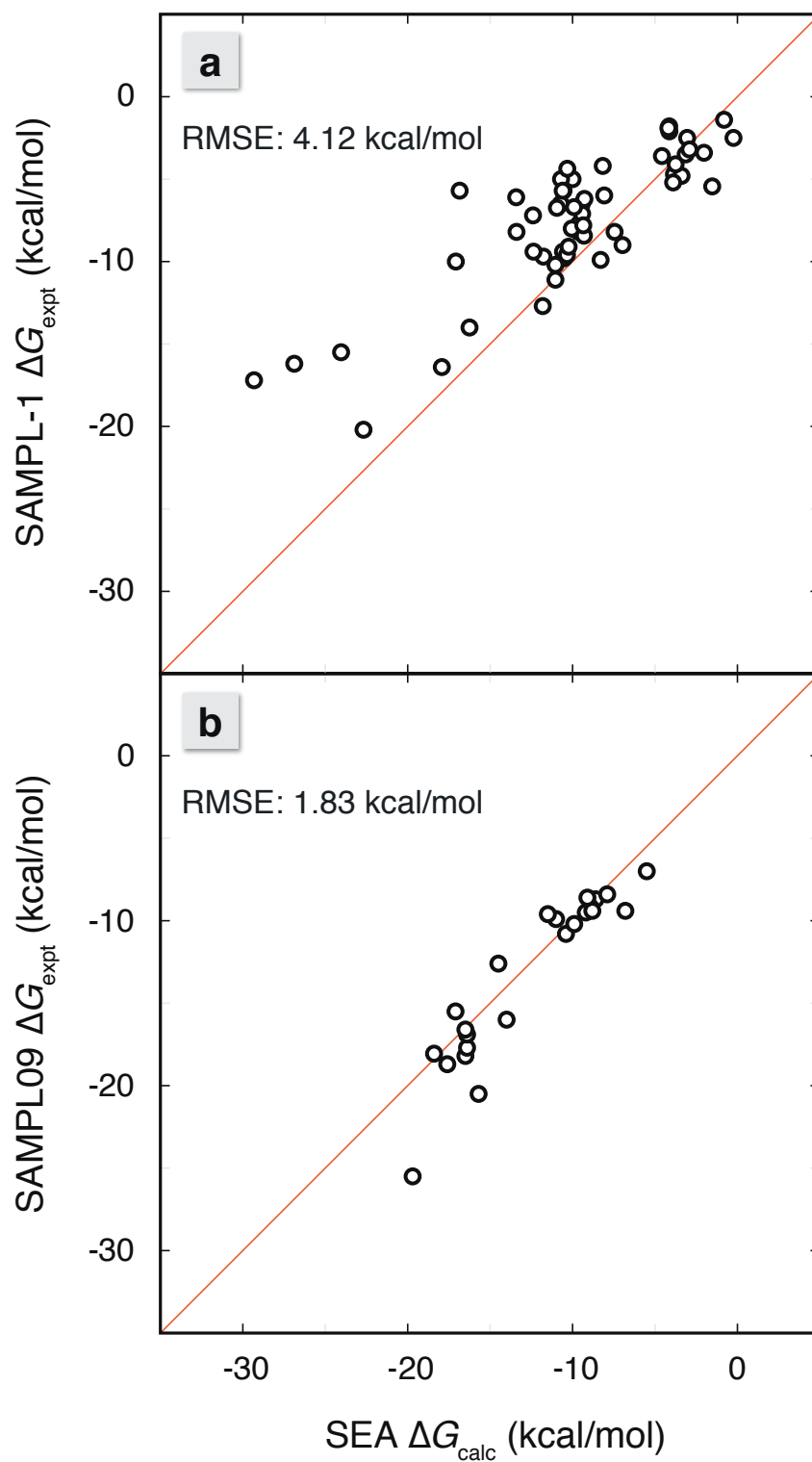
Figure 3.8: The calculated ΔG from SEA versus experimental values for the a) 56 blind prediction molecules of the SAMPL-1 event and the b) 23 blind prediction molecules in the SAMPL09 event.

(4) SEA fixes some physical flaws in implicit-solvent models

Because water's dipole is asymmetric, it means that positively charged atoms of a given radius solvate less favorably than negatively charged atoms of the same radius.[25,59,73,92,110,112] Such asymmetries in the physics are not directly captured in implicit-solvent models. In implicit-solvent models, this asymmetry is handled instead by empirical readjustment of solvation radii. To understand this asymmetry more quantitatively, Mobley *et al.* created a series of polar molecules that they solvated in explicit solvent.[92] These fictitious benzene-like 'bracelet' molecules were created to have either positive heads or negative heads, and were otherwise net neutral. The difference in free energy of solvation of the positive-head bracelet and negative-head bracelet was found to be about 10 kcal/mol. Zero difference in free energy would have been expected from implicit-solvent models. This is a useful test set of molecules for exploring whether SEA water modeling properly captures water's dipolar asymmetry.

Fig. 3.9 shows the results for bracelets of different sizes and shapes. It shows that while SEA does capture the physics of the asymmetry, it underestimates the free energy difference of the asymmetry. We find, however, that if we go beyond just using dipolar representation of water in SEA modeling, and use a quadrupolar representation instead, the agreement with TIP3P simulations is much better. A quadrupole gives a better fit to the water response map around atoms having a formal negative charge (see Fig. 3.2), like the head atom on the negative bracelets.
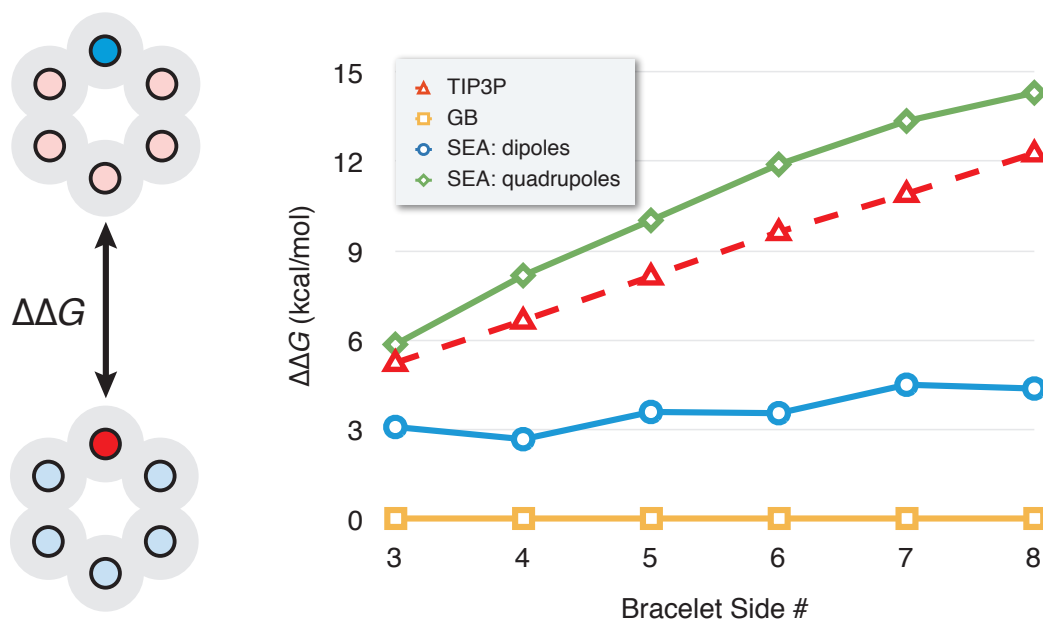
Figure 3.9: The solvation asymmetry ($\Delta\Delta G$) upon partial charge inversion for model polygons studied in Ref. 92. In this case, one atom of each polygon has a formal charge, while the neutralizing counter charge is distributed evenly among the remaining atoms, as shown in the 6 sided bracelet illustration on the left.

3.5   Conclusions

We have described a method, called SEA, for computing the free energies of solvation of solutes in water. SEA uses pre-computations with some chosen explicit-solvation model and then assembles solvation shells of water in order to compute solvation free energies of arbitrary solutes. Here, we performed these pre-computations using TIP3P water, though more complex models, such as polarizable water models, or non-aqueous solvents would also be practical. SEA aims to correct some of the physical deficits of implicit-solvent models, such as water's asymmetrical dipole, the lack of particulate water in the first solvation shell, the local additivity approximation, and the assumption that the free energy of cavity creation depends only on solute surface area and not on its shape. We have shown that SEA water predicts the solvation free energies for a range of non-polar and polar small molecules at approximately the same accuracy as in TIP3P computer simulations, but it runs much faster than explicit simulations, much closer to the speeds of implicit solvent models.

3.6   Supplementary Information

Available online at http://www.pnas.org/content/108/8/3234/suppl/DCSupplemental

Nearest–neighbor distribution function analysis

From the set of precomputation simulations, we generate simplified free energy representations of the local solvent molecules about single spheres. To assemble the solvation behavior around a given arbitrary solute, it is important that these solvent representations are placed on the surface in a physically reasonable manner, basically in a manner that mimics how explicit solvent occupies the surface of the solute. If they are too close to one another, the solvent–solvent interaction term may be too strong, leading to artificially unfavorable solvation. To obtain more physical spatial solvent arrangements, we place the dipoles or quadrupoles in accordance with the surface solvent nearest–neighbor distribution functions.

Weak Field
(less electrostriction)

Strong Field
(more electrostriction)
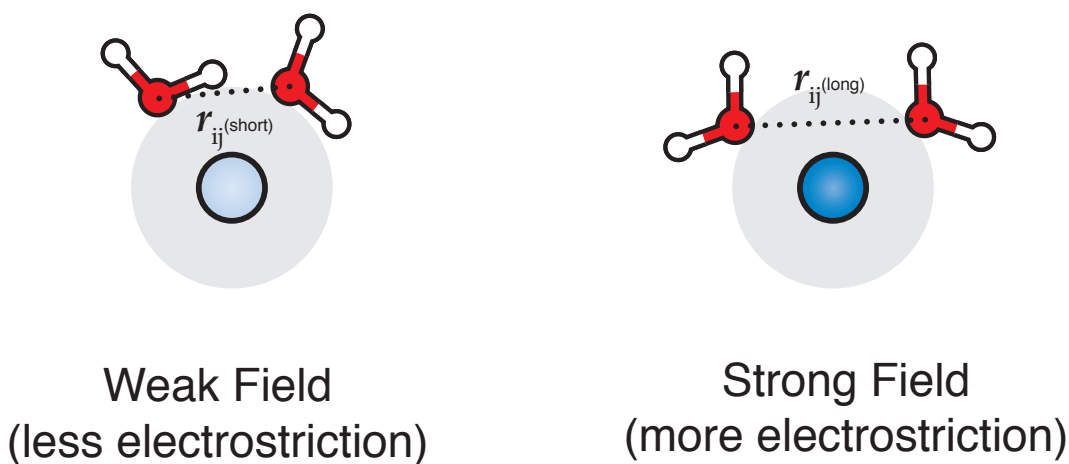
Figure 3.10: With weaker solute–solvent interactions (weak field), surface water is more free to form hydrogen bonds, leading to separation distances (r_{ij}) similar to bulk water (2.7 Å). With strong solute–solvent interactions (strong field), a surface water molecule is unable to form hydrogen bonds with neighboring surface waters, leading to longer separation distances.

The idea behind the surface solvent nearest–neighbor distribution function (P_{NN}) is to determine from explicit solvent simulations a solvent arrangement connection to the degree of electrostriction about a solute. Fig. 3.10 illustrates the observed behavior of water around weakly and strongly charged solutes. As the solute–solvent interactions become stronger (with increasing charge magnitude), the water becomes more strongly pinned to the solute and less able to form optimal hydrogen bonds with neighboring surface waters. Unable to form favorable interactions, surface waters interact in a more repulsive fashion, leading to longer separation distances. To determine these separation distances, we go to the extracted first–shell explicit solvent trajectories discussed in the main text, and bin water–water separations between nearest–neighbor waters. This can be determined as

$$P_{\text{NN}}(r_{\text{ij}}) = \frac{1}{M \cdot \langle N_{\text{AB}} \rangle} \sum^{M} \left(\sum_{i \in A}^{N_{\text{A}}} \sum_{j \in B}^{N_{\text{B}}} \delta \left(\min r_{\text{ij}} \right) \right). \tag{3.9}$$

Here, M is the number of configuration frames, $\langle N_{\text{AB}} \rangle$ is the average number of water molecules in frames where there exist at least 2 molecules (one termed A and the other B), and $\delta(\min r_{\text{ij}})$ is a delta function for each minimum r_{ij} value. Numerically, Eq. 3.9 is equal to 1, but the shape of P_{NN} as a function of r_{ij} is what we are interested in. The curve is peaked at the most populated surface solvent separation distance. Thus, for a given solute electric field strength, we know the optimal water spacing about a possible surface site.

In practice, we generate a table (see Tab. 3.2) of these P_{NN} peak values as a function of solute charge and size (*i.e.* local curvature). After placing a water representation dipole on the solute surface, we eliminate all possible subsequent sites that have distances shorter than those composing this peak. This eliminates unphysically close solvent placement.

Dipolar and quadrupolar free energy representations of solvent

One of the key goals of Semi-Explicit Assembly is to reproduce the surface water polarization response from explicit solvent simulations. This is accomplished through the creation of free energy maps of the surface solvent molecules. These maps then need to be placed about the solute surface accordingly so that we can accumulate the polar solvation free energy due to surface water. In an effort to keep SEA conceptually simple, we decided to start with the most straightforward electrostatic representation, a simple dipole. As the electric field about the simple spherical solutes in our explicit solvent simulations is always normal to the surface, we decided to condense these maps onto a single normal vector to generate our 'free energy dipoles'.

Fig. 3.11 illustrates the process used to perform this transformation. Charge density is binned along the normal vector dimension to determine the local charge distribution relative to each spherical solute surface. To generate a dipole representation, the partial charge of the explicit water model's oxygen is assigned to the location of average negative charge density, while a neutralizing positive partial charge is assigned to the location of the average positive charge density.
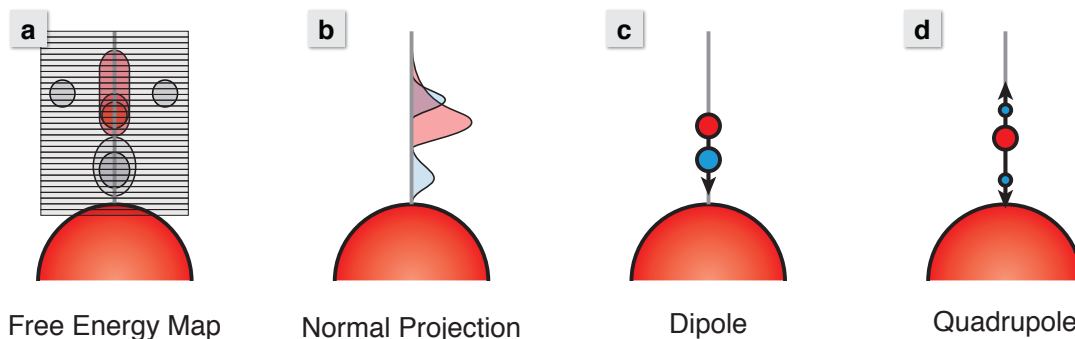
Figure 3.11: For each spherical solute simulation, the a) resulting free energy map is binned along the vector normal to the solute to generate a b) projection of the charge density along this vector. This charge density projection can be simplified to a c) dipole or d) quadrupole by calculating the locations of the average density of the positively and negatively charged distributions in the normal projection.

In the main text, we discussed a quadrupolar representation that could be used in place of the simple dipole representation. A quadrupole comes naturally from the free energy maps for water, because the hydrogen atoms easily separate out into two distinct distributions, one near to and the other far from the solute surface. When generating the free energy map, we actually classify these two hydrogens to separate near and far distributions, and place water hydrogen partial charges at the average distance of these distributions from the solute. This results in a linear quadrupole as seen in Fig. 3.11d. This is easily represented as two dipoles, and we can apply separate electric field curve fits to each. The benefit to a quadrupole representation is apparent around negatively charged solutes, as one of the solvent partial positive charges can get considerably closer to the solute. Drawbacks of a quadrupole representation include increased complexity and performance degradation in the interaction accumulation loop due the additional solvent site.

Electric field fits and data tables

The explicit solvent pre-simulations provide us with simple measures of surface solvent response. These measures include radial water distances, free energy dipole magnitudes in electric fields near a solutes surface, and water–water separation distances. We have also accumulated mean molecular occupancy values (and occupancy distribution variances) for the first solvation shell; however, these rarely get used as the water–water separation distances already help limit the number of "waters" that solvate each solute atom. While we have already accumulated and continue to accumulate all these quantities for other solvent types, we include below the relevant plots and tables for TIP3P at 300 K and 1 atm, the solvent and state point used for the SEA calculations in the main text.
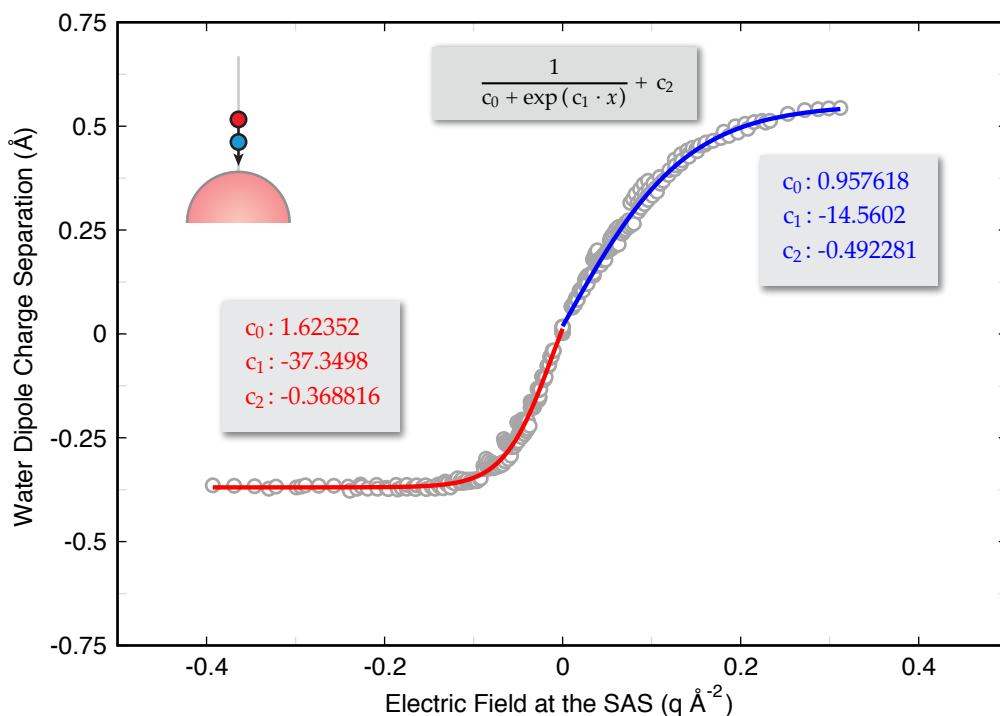
Figure 3.12: Sigmoidal curve fits to the positive and negative field regions of the dipole charge separation vs. surface electric field plots. The sign indicates direction of the dipole – with positive the partial positive charge is further than the negative from the solute and with negative it is nearer to the solute. These fits are done in internal units rather than SI units for ease in semi-explicit configuration assembly. One can multiply by 4.006 to convert charge separation to Debye, and multiply the electric field by $16.02/(4\pi\epsilon_0)$ to convert $q\text{Å}^{-2}$ to N/C.
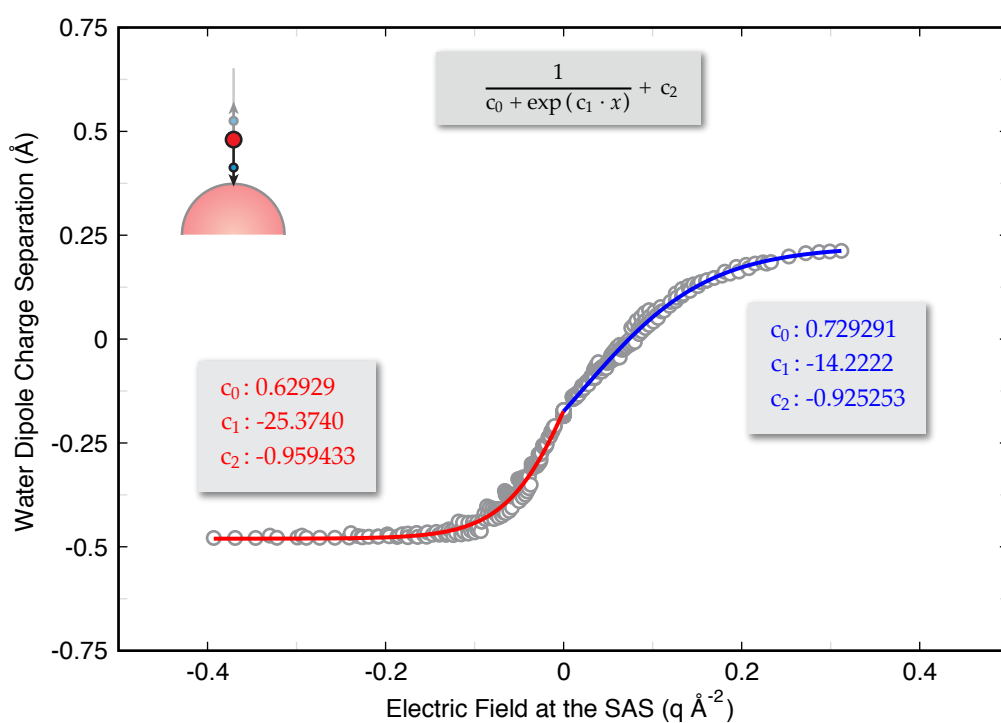
Figure 3.13: Sigmoidal curve fits to the positive and negative field regions of the inner dipole charge separation of the quadrupole vs. surface electric field plots.
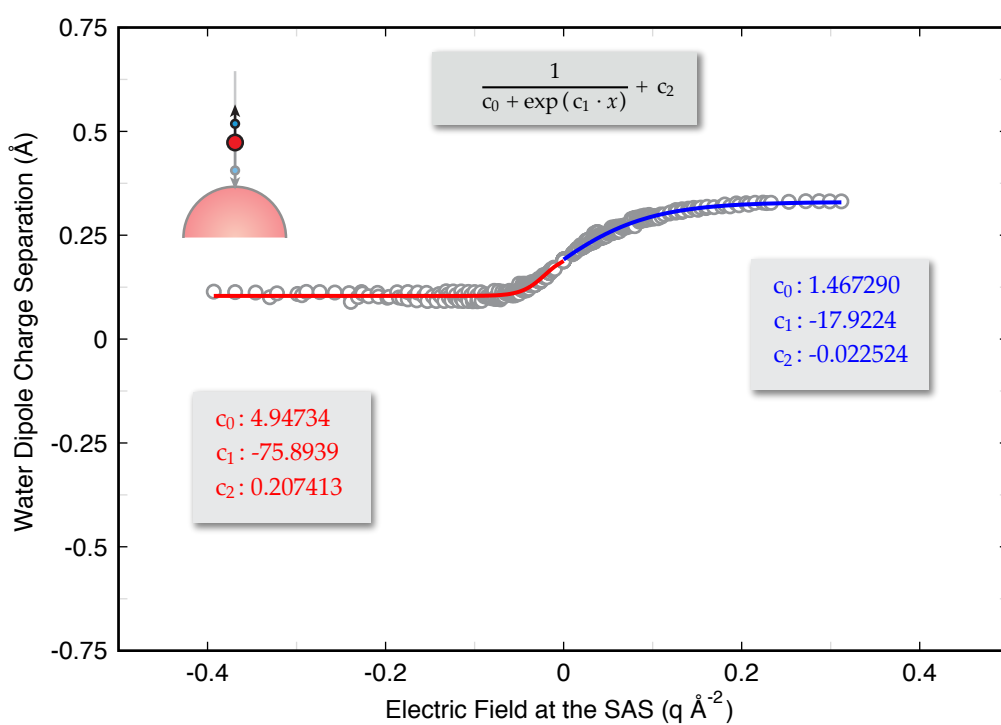
Figure 3.14: Sigmoidal curve fits to the positive and negative field regions of the outer dipole charge separation of the quadrupole vs. surface electric field plots.

Table 3.1: Extracted r_{w} values from the base LJ sphere pre-simulations used in this study. These come from water-oxygen/solute radial distribution functions, and are used in assembling the solvent accessible dot surface for iterative sampling to calculate ΔG_{pol}.

ϵ_{LJ} (kcal/mol)	σ_{LJ} (Å)					
	1.4	2.2	3.0	3.8	4.6	5.4
-1						
0.015625	1.74	2.04	2.39	2.91	3.41	3.88
0.03125	1.60	2.08	2.46	2.98	3.47	3.94
0.0625	1.65	2.14	2.56	3.05	3.54	4.01
0.125	1.70	2.19	2.65	3.13	3.61	4.08
0.25	1.76	2.26	2.73	3.21	3.68	4.14
0.5	1.83	2.33	2.82	3.29	3.76	4.21
-0.8						
0.015625	1.65	2.01	2.49	3.00	3.50	3.97
0.03125	1.66	2.10	2.56	3.07	3.57	4.03
0.0625	1.71	2.18	2.64	3.15	3.63	4.09
0.125	1.76	2.25	2.72	3.22	3.70	4.15
0.25	1.82	2.33	2.81	3.30	3.77	4.21
0.5	1.88	2.40	2.88	3.37	3.83	4.27
-0.6						
0.015625	1.62	2.07	2.62	3.15	3.63	4.07
0.03125	1.69	2.15	2.69	3.21	3.68	4.12
0.0625	1.79	2.23	2.77	3.28	3.74	4.18
0.125	1.85	2.32	2.84	3.34	3.80	4.23
0.25	1.91	2.40	2.92	3.40	3.85	4.28
0.5	1.97	2.48	2.99	3.46	3.91	4.33
-0.4						
0.015625	1.74	2.26	2.82	3.31	3.75	4.16
0.03125	1.80	2.34	2.89	3.37	3.79	4.20
0.0625	1.86	2.42	2.95	3.42	3.84	4.25
0.125	1.93	2.49	3.01	3.46	3.89	4.29
0.25	2.00	2.56	3.06	3.51	3.93	4.34
0.5	2.08	2.63	3.12	3.56	3.97	4.38
-0.2						
0.015625	2.00	2.58	3.04	3.44	3.83	4.22
0.03125	2.08	2.63	3.08	3.48	3.87	4.26
0.0625	2.15	2.68	3.11	3.51	3.91	4.30
0.125	2.21	2.72	3.15	3.55	3.94	4.34
0.25	2.27	2.76	3.18	3.58	3.98	4.37
0.5	2.32	2.79	3.21	3.61	4.01	4.41
0						
0.015625	2.37	2.74	3.11	3.49	3.86	4.24
0.03125	2.39	2.76	3.14	3.52	3.90	4.28
0.0625	2.41	2.79	3.17	3.55	3.93	4.31
0.125	2.43	2.81	3.19	3.58	3.96	4.35
0.25	2.44	2.83	3.22	3.61	3.99	4.38
0.5	2.46	2.85	3.24	3.63	4.02	4.42

ϵ_{LJ} (kcal/mol)	σ_{LJ} (Å)					
	1.4	2.2	3.0	3.8	4.6	5.4
+0.2						
0.015625	2.25	2.67	3.08	3.46	3.84	4.22
0.03125	2.29	2.71	3.10	3.50	3.88	4.26
0.0625	2.32	2.74	3.14	3.53	3.92	4.30
0.125	2.35	2.77	3.17	3.56	3.95	4.34
0.25	2.38	2.80	3.20	3.59	3.99	4.38
0.5	2.41	2.83	3.22	3.62	4.02	4.41
+0.4						
0.015625	2.04	2.52	2.98	3.40	3.80	4.19
0.03125	2.09	2.57	3.02	3.44	3.84	4.23
0.0625	2.15	2.63	3.06	3.48	3.88	4.27
0.125	2.20	2.67	3.11	3.52	3.92	4.32
0.25	2.26	2.72	3.15	3.56	3.96	4.36
0.5	2.30	2.76	3.18	3.59	4.00	4.39
+0.6						
0.015625	1.93	2.35	2.84	3.30	3.72	4.13
0.03125	1.97	2.42	2.90	3.35	3.78	4.19
0.0625	2.02	2.49	2.96	3.40	3.83	4.23
0.125	2.07	2.55	3.02	3.46	3.88	4.28
0.25	2.13	2.61	3.07	3.51	3.92	4.33
0.5	2.19	2.67	3.12	3.55	3.96	4.37
+0.8						
0.015625	1.85	2.23	2.70	3.18	3.64	4.06
0.03125	1.89	2.31	2.77	3.25	3.70	4.12
0.0625	1.93	2.38	2.85	3.32	3.76	4.18
0.125	1.98	2.45	2.92	3.38	3.82	4.24
0.25	2.03	2.51	2.99	3.44	3.87	4.29
0.5	2.10	2.58	3.05	3.50	3.92	4.34
+1						
0.015625	1.79	2.19	2.57	3.06	3.54	3.99
0.03125	1.83	2.25	2.65	3.14	3.61	4.06
0.0625	1.87	2.31	2.74	3.22	3.68	4.12
0.125	1.92	2.37	2.82	3.29	3.75	4.18
0.25	1.99	2.43	2.90	3.37	3.82	4.24
0.5	2.09	2.50	2.97	3.44	3.88	4.30

Table 3.2: Extracted solvent–solvent spacing distances from the base LJ sphere pre-simulations used in this study. These are the peak locations of the first shell nearest–neighbor distribution functions, and are used to eliminate dot surface sampling sites that are too close to occupied solvent sites. As seen below, this is more of an issue around smaller negatively charged solute atoms. Also, when the field gets extremely strong, solute–solvent interactions dominate the spacing by trying to pack as many first shell waters in as possible (lowering the solvent spacing back to the bulk r_{OO} distance).

ϵ_{LJ}	σ_{LJ} (Å)					
(kcal/mol)	1.4	2.2	3.0	3.8	4.6	5.4
-1						
0.015625	2.50	2.83	3.01	3.07	2.78	2.76
0.03125	2.55	2.78	3.04	3.10	2.77	2.77
0.0625	2.62	2.84	3.11	3.10	2.77	2.78
0.125	2.70	2.90	3.07	3.19	2.79	2.78
0.25	2.77	2.96	3.09	3.22	2.79	2.77
0.5	2.85	3.01	3.07	3.23	2.79	2.79
-0.8						
0.015625	2.59	2.98	3.08	2.78	2.77	2.76
0.03125	2.64	2.94	3.10	2.81	2.78	2.76
0.0625	2.70	2.91	3.18	2.81	2.79	2.77
0.125	2.77	2.98	3.20	2.83	2.79	2.77
0.25	2.85	3.03	3.28	2.82	2.78	2.78
0.5	2.92	3.07	3.24	2.80	2.79	2.80
-0.6						
0.015625	2.71	3.09	3.17	2.78	2.77	2.76
0.03125	2.80	3.16	3.30	2.78	2.79	2.76
0.0625	2.83	3.19	3.27	2.79	2.78	2.77
0.125	2.89	3.21	3.30	2.79	2.78	2.78
0.25	2.95	3.24	3.25	2.80	2.77	2.77
0.5	3.01	3.17	3.34	2.79	2.77	2.77
-0.4						
0.015625	2.88	3.34	2.80	2.79	2.76	2.75
0.03125	2.96	3.30	2.80	2.78	2.77	2.76
0.0625	3.04	3.40	2.80	2.79	2.77	2.76
0.125	3.13	3.38	2.80	2.78	2.77	2.76
0.25	3.16	3.37	2.82	2.79	2.76	2.76
0.5	3.21	3.36	2.80	2.77	2.79	2.77
-0.2						
0.015625	3.42	2.78	2.77	2.75	2.75	2.75
0.03125	3.37	2.80	2.78	2.76	2.76	2.74
0.0625	3.21	2.77	2.77	2.77	2.76	2.75
0.125	3.10	2.79	2.79	2.76	2.75	2.75
0.25	3.03	2.77	2.77	2.76	2.75	2.76
0.5	2.86	2.79	2.76	2.76	2.76	2.76
0						
0.015625	2.76	2.75	2.74	2.75	2.73	2.75
0.03125	2.75	2.75	2.75	2.75	2.75	2.75
0.0625	2.78	2.75	2.75	2.76	2.75	2.75
0.125	2.78	2.76	2.76	2.75	2.75	2.75
0.25	2.76	2.75	2.76	2.75	2.75	2.76
0.5	2.77	2.76	2.77	2.76	2.76	2.75

ϵ_{LJ}	σ_{LJ} (Å)					
(kcal/mol)	1.4	2.2	3.0	3.8	4.6	5.4
+0.2						
0.015625	2.74	2.74	2.74	2.73	2.74	2.74
0.03125	2.74	2.73	2.73	2.74	2.74	2.73
0.0625	2.75	2.74	2.74	2.74	2.75	2.75
0.125	2.76	2.74	2.74	2.74	2.74	2.75
0.25	2.73	2.76	2.74	2.75	2.74	2.74
0.5	2.72	2.76	2.73	2.74	2.74	2.76
+0.4						
0.015625	3.13	2.77	2.74	2.73	2.73	2.74
0.03125	3.16	2.74	2.74	2.75	2.73	2.73
0.0625	3.16	2.74	2.74	2.74	2.73	2.73
0.125	3.21	2.75	2.73	2.73	2.73	2.74
0.25	3.30	2.73	2.76	2.74	2.74	2.74
0.5	3.24	2.76	2.74	2.73	2.74	2.74
+0.6						
0.015625	2.97	3.17	2.74	2.73	2.73	2.73
0.03125	3.00	3.16	2.74	2.73	2.73	2.73
0.0625	3.04	3.18	2.74	2.74	2.73	2.73
0.125	3.08	3.05	2.74	2.74	2.73	2.73
0.25	3.13	2.76	2.74	2.73	2.74	2.73
0.5	3.10	2.79	2.74	2.74	2.74	2.73
+0.8						
0.015625	2.87	3.04	3.15	2.74	2.74	2.73
0.03125	2.91	3.06	2.80	2.74	2.73	2.73
0.0625	2.96	3.09	2.77	2.73	2.73	2.73
0.125	3.01	3.10	2.77	2.74	2.73	2.73
0.25	3.04	3.13	2.74	2.74	2.73	2.73
0.5	3.04	3.14	2.74	2.74	2.73	2.74
+1						
0.015625	2.80	2.91	3.10	2.76	2.74	2.72
0.03125	2.85	2.95	3.12	2.76	2.73	2.74
0.0625	2.90	3.00	3.11	2.74	2.73	2.73
0.125	2.94	3.04	3.14	2.76	2.73	2.74
0.25	2.97	3.07	3.09	2.74	2.74	2.73
0.5	2.87	3.09	3.10	2.73	2.73	2.73

Chapter 4

Testing the semi-explicit assembly solvation model in the SAMPL3 community blind test

Authors: Charles Kehoe, Christopher Fennell, and Ken Dill

Abstract

We report here a test of the Semi-Explicit Assembly (SEA) model in the solvation free energy category of the SAMPL3 blind prediction event (summer 2011). We tested how dependent the SEA results are on the chosen force field by performing calculations with both the General Amber and OPLS force fields. We compared our SEA results with full molecular dynamics simulations with explicit solvent. Of the 20 submissions, our SEA/OPLS results gave the second smallest RMS errors in free energies compared to experiments. SEA gives results that are very similar to those of its underlying force field and explicit solvent model. Hence, while the SEA water modeling approach is much faster than explicit solvent simulations, its predictions appear to be just as accurate.

4.1   Introduction

Motivated by the need for improved computational models of water and aqueous solutions, we recently developed a solvation model called Semi-Explicit Assembly (SEA) water.[39,40] In this paper, we report a test of the SEA model through our participation in the solvation-free-energy category of the SAMPL3 blind prediction challenge.

4.2 Description of the SEA-water method

SEA has been described in detail elsewhere,[39,40] so here we give only a brief summary. As shown in figure 4.1, SEA divides the calculation of the solvation free energy of a solute into two parts. First, there is a pre-computation stage which samples the interaction of various test spheres (having different radii, charges and van der Waals properties) with a chosen explicit-solvent model of water, such as TIP3P. Among other properties, these pre-simulations provide the average axial dipole moment of first-shell waters as a function of the local electric field around the different solute spheres. Second, at runtime, SEA models the solute under study as an assembly of those pre-computed spheres. We can then rapidly compute a solvation free energy as a sum over the properties of all the water molecules around the solute. In this way, the SEA model captures many of the physical and structural properties of the explicit-water model on which it was parameterized, and also captures the particulate and electrostatic properties of first-shell waters. Yet the calculations at runtime are nearly as fast to compute as those of implicit-solvent models.

SAMPL3 small molecules: a test of chlorinated molecule solvation

The SAMPL event, developed and run by OpenEye Software, is a community-wide blind test of computational chemistry prediction methods, including the prediction of solvation free energies.[47,51,100] At the start, participants are provided with a
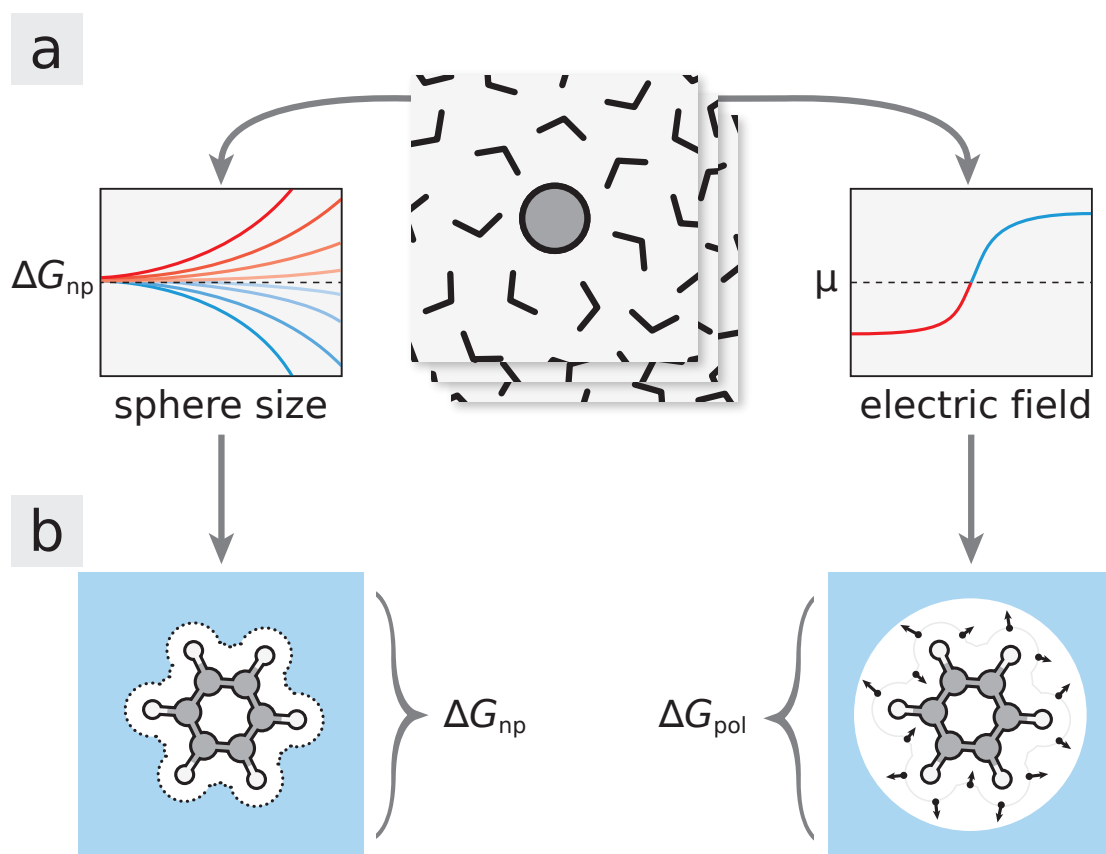
Figure 4.1: (a) The SEA solvation model involves pre-simulations of neutral and charged spheres in explicit water to generate maps of the ΔG_{np} vs. sphere size and solvent dipole response vs. surface electric field. (b) After these one-time calculations, we can rapidly calculate the ΔG_{np} and ΔG_{pol} terms of the total solvation free energy for an arbitrary solute molecule.

set of varied solutes for which they do not know the solvation free energies. Each group then uses its particular methodology to compute solvation free energies for these compounds. After these predictions are submitted, the SAMPL organizers then provide experimental solvation free energies to compare against. This year's event, SAMPL3, included 36 different solute molecules, roughly evenly divided among ethanes, biphenyls, and dibenzo-p-dioxins. The molecules in each class varied only in the number and location of substituted chlorines. This systematic approach allowed for an analysis of trends across these classes of molecules, but it also had the potential to exaggerate errors for any methods that are challenged by these particular molecule types or by chlorinated solutes. In this work, we show our results for predictions on these 36 molecules, and explore how the systematic changes in the chemical structure affect the details of solvation with SEA, explicit, and implicit calculations.

4.3 Simulation Methods

The SEA method relies on underlying pre-simulations using an atomically detailed force field. For comparisons of computational solvation methods, we performed SEA and explicit-solvent calculations with the General Amber Force Field (GAFF) and the OPLS force field.[66, 139] We used AM1-BCC partial charges[62] for the GAFF calculations. As an additional implicit solvent comparison, we performed Poisson-Boltzmann (PB) calculations with the resultant GAFF solute topologies. Our GAFF parameters and AM1-BCC charges came from the ANTECHAMBER program in AmberTools

1.4,[138] and we used the TIP3P water model[64] to solvate these structures, following the practices of Mobley et al.[95] OPLS parameters were from Desmond 2.4.2.1,[16] and we solvated these structures using the SPC water model,[10] as was done in a published study of OPLS small-molecule solvation free energy calculations.[120] For each force field, we also submitted two sets of results: one set with minimally relaxed, single conformations, and another with multiple conformations per molecule, sampled from explicit water simulations.

We submitted the SEA and PB calculations to the SAMPL3 event as blind predictions, and performed the explicit calculations for the present study. We also compared results from solutes put into a single (dominant) configuration against results that sampled some of the solute's conformational options.

Setting up molecules with Amber-based and OPLS-based force fields

We performed all molecular dynamics (MD) calculations with GROMACS 4.0.4.[12,55] We switched off Lennard-Jones interactions between 8 and 9 Å and applied long-ranged energy and pressure corrections. We used smooth particle-mesh Ewald for long-ranged electrostatics accumulation,[35] this with a real-space cutoff of 10 Å, a spline order of 6, fourier spacing of 1 Å, and a real-space energy tolerance parameter of 10^{-6} kJ/mol. We used the SETTLE algorithm to constrain the geometry of water

molecules in explicit-solvent simulations,[91] and the LINCS algorithm to constrain covalent bonds to hydrogen atoms on solute molecules.[53]

We solvated the target solutes in a rhombic dodecahedral box with a 12 Å buffer of explicit water between the atoms of the solute and the edges of the box. We relaxed these systems in two steps, first with 1000 steps of steepest descent minimization followed by 10 ps of constant energy MD with a 1 fs timestep. All timesteps after this relaxation phase were 2 fs. For the single-conformation calculations with SEA, we used the final solute structure following 100 ps of Langevin Dynamics (LD) equilibration of these relaxed structures in explicit solvent. For the multi-conformation sets, we followed relaxation with 100 ps of constant temperature LD (300 K), 100 ps of constant pressure (1 atm) equilibration using the Berendsen thermostat, and 700 ps of constant pressure dynamics using the Parrinello-Rahman barostat. We rescaled the simulation box to the average volume from last 500 ps of this constant pressure trajectory, and equilibrated this system with 500 ps of constant volume simulation before our 1 ns LD production run. We selected 10 conformations at equal time intervals from the production run calculations with SEA. Our final multi-conformation results are arithmetic averages of the SEA calculations on these 10 solute conformers.

SEA solvation free energy calculations

The SEA solvation method uses an extensive set of pre-simulations that were calculated using Lorentz-Berthelot mixing rules for Lennard-Jones (LJ) interactions. These resulting property tables are compatible with Amber force fields, but not directly compatible with OPLS force fields because OPLS uses a geometric mean for determining the LJ size (σ) parameters between dissimilar atom types. To perform OPLS force field calculations with SEA, we converted the OPLS LJ parameters to an arithmetic mean equivalent,

$$\sigma_{\text{new}} = 2\left(\sqrt{\sigma_{\text{OPLS}} \cdot \sigma_{\text{wat}}}\right) - \sigma_{\text{wat}} \tag{4.1}$$

where σ_{wat} is the LJ σ value for the explicit water model of interest, in this case SPC water.[10] Note that converting atom parameters in this way will result in differences in solute intramolecular LJ interactions. This does not directly affect the SEA calculations as there is no solute intramolecular interaction contribution in its estimation of the solvation free energy.

Poisson-Boltzmann solvation free energy calculations

We performed our PB tests by combining a polar term and a non-polar term. For the polar term, we solved the linearized Poisson equation provided in APBS.[6] We set the

dielectric boundary as the molecular surface and used a solvent dielectric constant of 78. For the non-polar term, we used the standard expression

$$\Delta G_{np} = 0.00542 \times SASA + 0.92, \qquad (4.2)$$

where SASA is the solvent-accessible surface area in Å^2 to give ΔG_{np} values in kcal/mol.[113] Our PB calculations were performed on the multi-conformer GAFF structures.

Explicit solvent solvation free energy calculations

For our explicit solvent comparisons, we performed solvation free energy calculations as others have done in previous studies.[97,98] Here, we used thermodynamic integration to transform the solute molecules between the relevant state-points. For the ΔG_{chg} term, we turned off the partial charges of the solute in TIP3P or SPC water (for the GAFF and OPLS calculations respectively) over λ windows of $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Each of these windows involved a 5 ns simulation with GROMACS 4.0.4 using the same simulation protocols described previously for the conformation sampling calculations. In order to determine the absolute ΔG_{pol} term for transfer from air to water, we subtracted out the internal Coulombic and conformation distribution contributions to ΔG_{chg}. We calculated these by determining the ΔG_{chg} term in vacuum, using λ window values of $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

For the ΔG_{np} term calculation, we converted the uncharged solute to a soft-core representation[124] over the set of λ window values {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}. We used the standard trapezoid rule for integration, and accumulated errors via the limiting value of block averages.[54]

4.4   Results and Discussion

Figure 4.2 shows an overview of our SAMPL3 submissions. For each method, we show the RMS difference in free energy between the predicted solvation free energy and the experimental value. After we parameterized the molecules and sampled their conformations, SEA generated each prediction in less than one second on a 3 GHz Intel Core 2 E8400 processor. As shown, SEA was one of the more accurate predictors in the event, despite its modest computational cost. In terms of absolute performance relative to the rest of the predictions, it should be noted that this is a small set of very related solutes, so it is difficult to make definitive assessments. We perform additional comparisons below to explore these SEA results in more detail.

As an implicit solvent reference, we submitted PB results using the GAFF and TIP3P Lennard-Jones parameters to define the molecular surface and solvent accessible surface area along with the same AM1-BCC partial charges used in the SEA calculations. This combination of parameters did not perform particularly well, about 1.7 kcal/mol higher RMSE than the analogous SEA submission. Most of this difference comes from
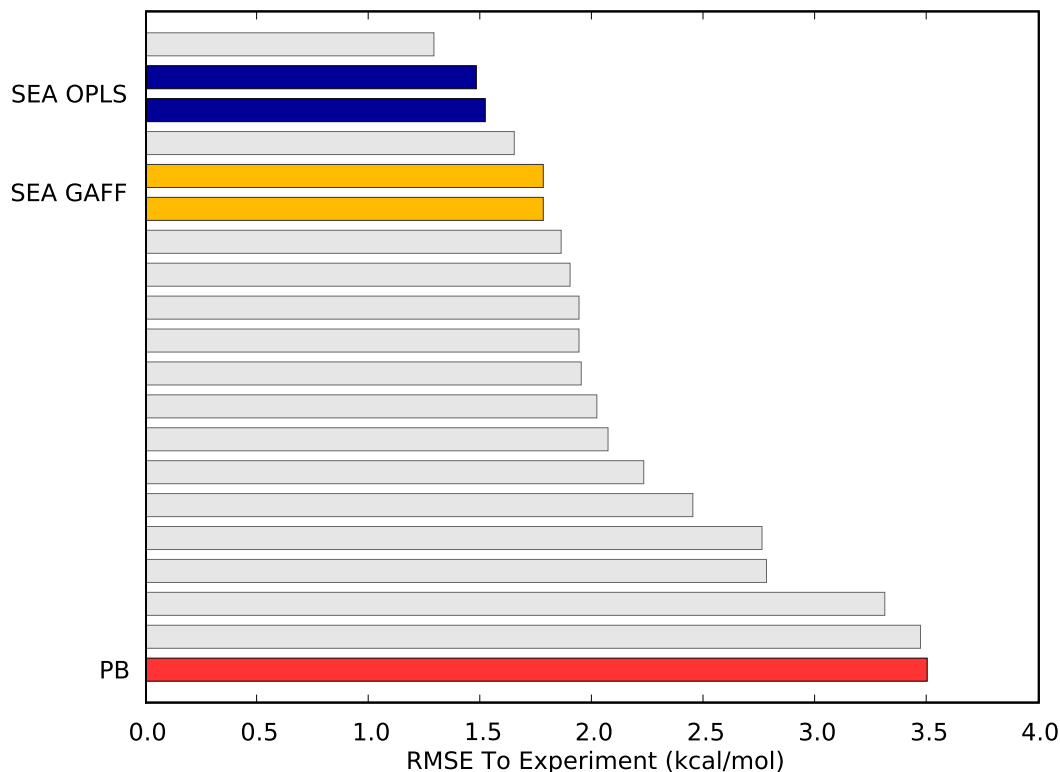
Figure 4.2: Overall performance of the 20 submissions to SAMPL3. Average RMSE of each method; best predictions are at the top. Our submissions are highlighted.

the use of equation 4.2 as opposed to a more microscopically physical approach to ΔG_{np}. If the SEA ΔG_{np} term is used in place of equation 4.2, this application of PB would have given an only 0.2 kcal/mol higher RMSE than SEA.

Multiple conformations do not alter the solvation free energies

The 36 compounds in the SAMPL3 test were all fairly rigid. Not surprisingly, we found that the free energies of the conformational ensembles of the solutes are accurately approximated, in this case, by the single dominant conformer. The differences

90

are within the errors of our calculations, only a few hundredths of a kcal/mol, which is similar to the deviations between runs on identical input. This is markedly different than for previous SAMPL events, in which some solute molecules (like the sugars in SAMPL2) involved multiple energetically important conformations.

After the SAMPL meeting, it came to our attention that three of the biphenyl molecules, 1,2,4,5-tetrachloro-3-(3,4-dichlorophenyl) benzene, 1,2,3,4-tetrachloro-5-(3,4-dichlorophenyl) benzene, and 1,2,3,4-tetrachloro-5-(3,4,5-trichlorophenyl) benzene, have a large dihedral barrier between two stereochemically unique conformations that might affect the resulting solvation free energies. While the multi-conformation calculations failed to show any added uncertainty for these molecules, no barrier crossings were identified over the course of a 10 ns MD conformation calculation.

To test if these unsampled conformations would play any role in the resulting solvation free energies, we performed single-point SEA calculations about minimized structures on either side of this dihedral barrier. Table S1 shows the results of these calculations. The differences between these conformations, though not completely negligible, are in most cases much less than the difference between the force field's predictions and the experimental results.

The SEA model is an accurate mimic of the underlying explicit water model

One key finding here is that the SEA model captures quite accurately the much more expensive explicit force field simulations on which it was parameterized; see figure 4.3. That is, SEA/GAFF gives very similar results to GAFF explicit solvent calculations, and SEA/OPLS gives very similar results to OPLS explicit solvent calculations. The advantage is that the SEA methods are six orders of magnitude faster to compute.

So, if the force field parameters are flawed, the corresponding SEA models show the same errors. For example, for the GAFF topologies, the most heavily chlorinated biphenyl and dioxin solutes were the most problematic for both the SEA and explicit simulations. This was true for the OPLS topologies as well, but in a different way. With the GAFF topologies, SEA and explicit solvent systematically *under*-solvated these molecules. With the OPLS topologies, SEA and explicit solvent systematically *over*-solvated these molecules. The reason for this is shown in figure 4.4. While the LJ parameters are similar for the aromatic carbon and substituted chlorine atoms, the partial charges are roughly 7 times greater with OPLS than with GAFF/AM1-BCC. The resulting difference in ring electric fields is why we see an 8 kcal/mol difference between GAFF in OPLS in both the SEA and explicit calculations. While it is unclear which set of partial charges gives a better representation of the actual electric field (the GAFF calculations are closer to experiment in this case, but not in others), SEA and explicit solvent respond to the topologies in the same way.
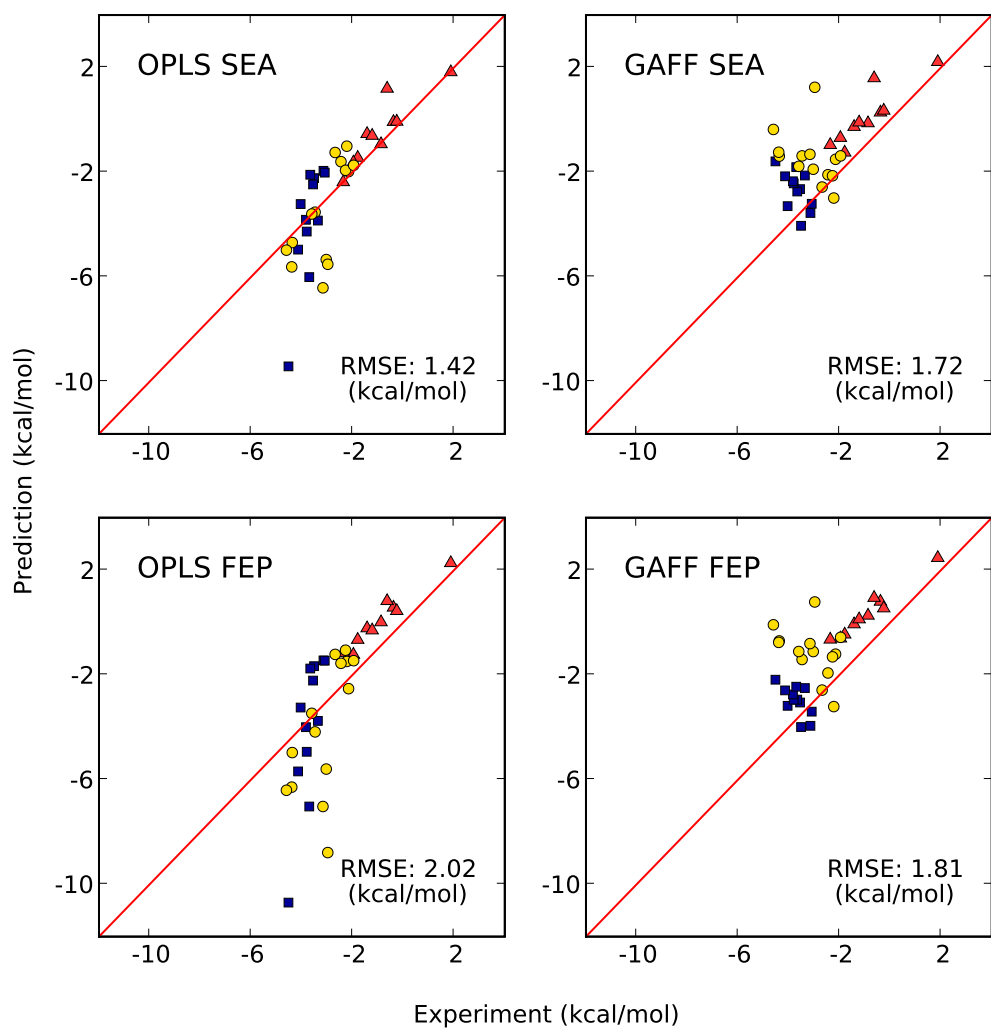
Figure 4.3: The performance of SEA/GAFF, SEA/OPLS, GAFF explicit and OPLS explicit vs. experimental data for all the solutes in SAMPL3. Ethanes appear as triangles, biphenyls as circles, and dioxins as squares.
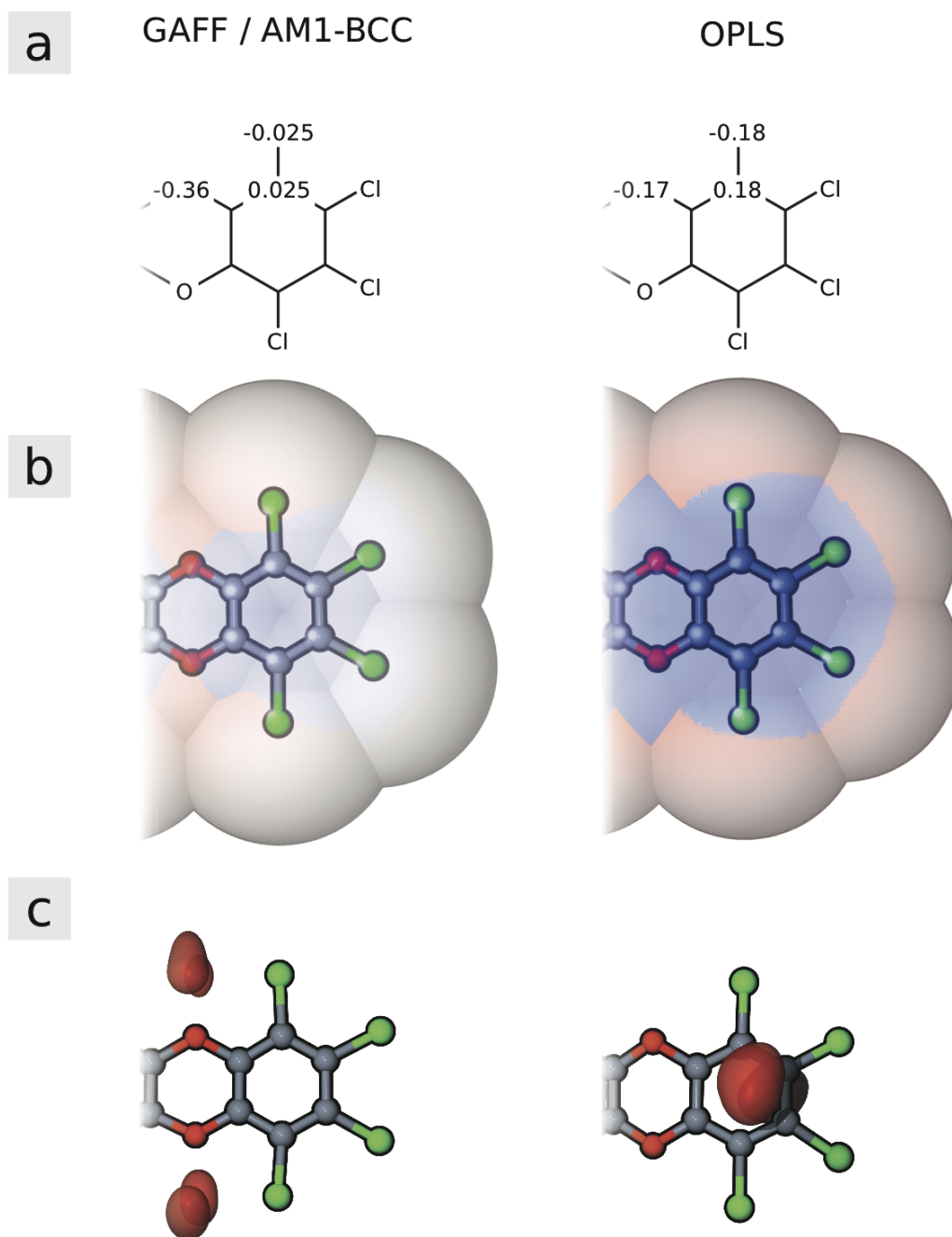
Figure 4.4: A comparison of the aromatic ring region of octachlorodibenzo-p-dioxin with the GAFF/AM1-BCC and OPLS topologies. The (a) partial charges are larger in OPLS, resulting in a (b) stronger field seen by SEA at the solvent-accessible surface. This difference in electric fields also affects explicit solvent, with (c) a greater water occupancy probability interacting with the ring with the OPLS partial charges.

We also performed retrospective tests on past SAMPL solutes using SEA

As a further test, we performed SEA calculations on all the previous SAMPL solute sets; our results are shown in Figure 4.5. We used the same simulation protocol as that used for the SAMPL3 calculations. We compared the multi-conformer results because previous SAMPL events had some solutes with more flexible dihedrals.

OPLS performed poorly in SAMPL2, while GAFF performed poorly in SAMPL1. In SAMPL2, OPLS performed poorly on cyclic nitrogen compounds. When these compounds are removed from the dataset, OPLS performance improves markedly, while GAFF performance degrades slightly. Similarly in SAMPL1, GAFF (and OPLS) perform poorly on sulfur compounds. Removing the sulfur compounds reduces the problem. Results from both forcefields also improve when we skip the amides from SAMPL0.

These results illustrate a key challenge of force field development: parameters the work better for some molecules may be worse for others. SEA provides a tool for diagnosing these problems. It is notable that for every SAMPL, one of either SEA/GAFF or SEA/OPLS gives predictions that are among the very best predictions of that year.
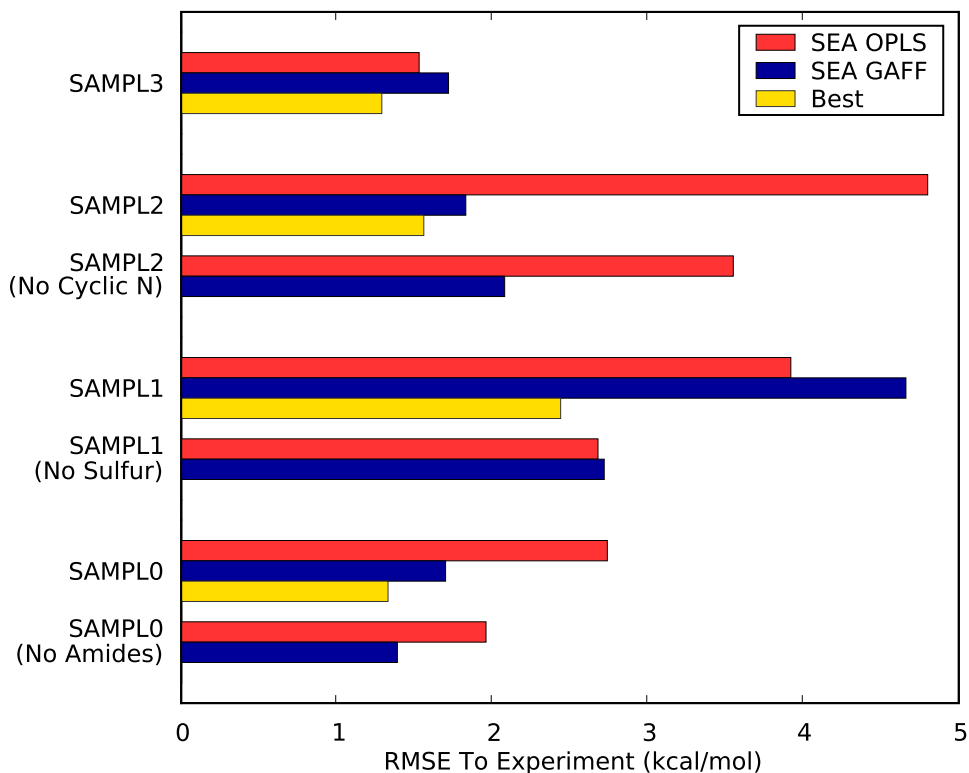
Figure 4.5: The performance of SEA/GAFF and SEA/OPLS on all SAMPL events. Which force field performs better depends heavily on the set of molecules in question. In each case, however, SEA is able to produce results rivaling the top performer from that year, given a sufficiently accurate force field. We are also able to identify some force field weaknesses: both forcefields have trouble with the sulfur compounds in SAMPL1, and the amides in SAMPL0. OPLS also improves markedly if we skip the cyclic nitrogen rings from SAMPL2. SEA's accuracy and speed can greatly facilitate force field tuning in cases like these.

4.5 Conclusions

We have tested the SEA water method of computational solvation in the blind test SAMP3 event on 36 solute molecules. We submitted SEA calculation results using both GAFF and the OPLS force field. SEA/OPLS performed second best among 20 submissions. For these rigid solutes, we confirmed that sampling solute conformations added no further value. We compared our SEA approach with explicit solvent free-energy calculations, and found that both methods give quantitatively similar solvation free energies when based on the same underlying force field. The advantage of the SEA method over explicit solvation simulations is that (in this test) it is six orders of magnitude faster to compute. Hence SEA water promises to be faster than explicit solvation and more accurate than implicit solvation.

4.6 Supplementary Information

Available online at

http://www.springerlink.com/content/553821460730g10l/supplementals

Table S1 SEA solvation energies for alternate conformers of three biphenyls from SAMPL3. The two conformers of these molecules were separated by particularly high energy barriers, so we tested each conformer explicitly to see if they solvated differently. All values are in kcal/mol.
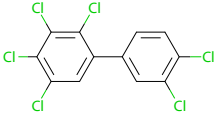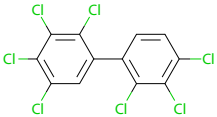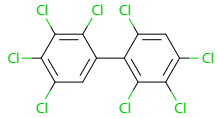
Molecule	SEA GAFF	SEA OPLS	Experiment
	-1.58 / -1.71	-5.74 / -5.53	-3.04
	-1.19 / -1.15	-5.77 / -6.01	-4.40
	-0.34 / -0.28	-5.19 / -4.98	-4.61

Table S2a Our SAMPL3 hydration free energy submissions. All values are in kcal/mol. The first four columns are SEA predictions, using the indicated force-field. Multi-conformer SEA submissions are marked as "MC", single-conformers as "SC". The last column contains multi-conformer PB results.
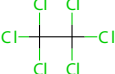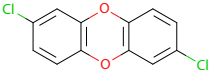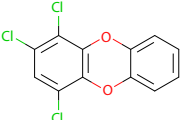
Molecule	GAFF MC	GAFF SC	OPLS MC	OPLS SC	PB GAFF
—	2.21	2.25	1.82	1.93	1.66
	0.28	0.45	-0.08	-0.02	0.42
	-0.13	-0.25	-0.93	-0.93	0.19
	-1.25	-1.16	-1.45	-0.42	-0.64
	0.35	0.47	-0.07	-0.07	0.81
	-0.69	-1.84	-1.63	-0.82	-0.30
	-0.27	-0.40	-0.54	-0.68	0.41
	-0.96	-0.84	-2.38	-2.55	-0.27
	-0.10	-0.03	-0.61	-0.68	0.81
	1.59	1.49	1.19	1.12	2.02

99

Table S2b Our SAMPL3 hydration free energy submissions, continued. All values are in kcal/mol. The first four columns are SEA predictions, using the indicated forcefield. Multi-conformer SEA submissions are marked as "MC", single-conformers as "SC". The last column contains multi-conformer PB results.
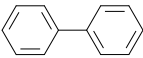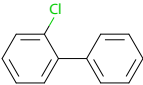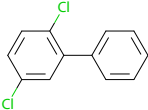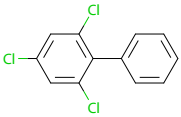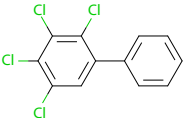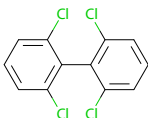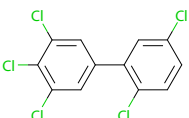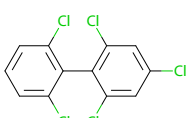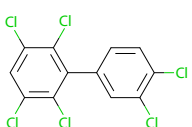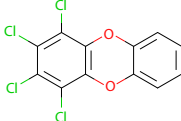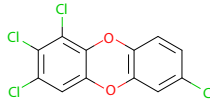
Molecule	GAFF MC	GAFF SC	OPLS MC	OPLS SC	PB GAFF
	-2.99	-2.76	-1.01	-0.91	-1.41
	-2.57	-2.41	-1.25	-1.32	-0.98
	-2.10	-1.64	-1.60	-1.31	-0.43
	-1.51	-1.60	-1.98	-2.22	0.11
	-1.38	-1.80	-3.53	-3.49	0.46
	-2.14	-2.09	-1.93	-1.89	-0.46
	-1.78	-1.62	-3.60	-3.82	0.55
	-1.38	-1.42	-1.74	-1.74	0.33
	-1.39	-1.46	-4.69	-4.72	0.95

Table S2c Our SAMPL3 hydration free energy submissions, continued. All values are in kcal/mol. The first four columns are SEA predictions, using the indicated forcefield. Multi-conformer SEA submissions are marked as "MC", single-conformers as "SC". The last column contains multi-conformer PB results.
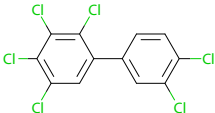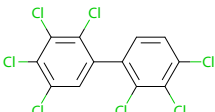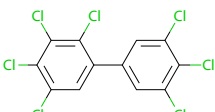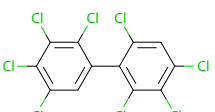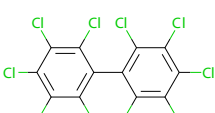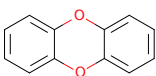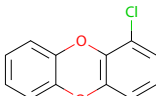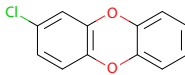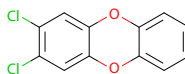
Molecule	GAFF MC	GAFF SC	OPLS MC	OPLS SC	PB GAFF
	-1.89	-1.73	-5.34	-5.39	0.82
	-1.24	-1.22	-5.62	-5.98	1.17
	-1.32	-1.09	-6.42	-5.93	1.42
	-0.37	-0.40	-4.98	-4.76	1.92
	1.24	1.31	-5.52	-5.97	3.11
	-3.56	-3.43	-1.95	-2.16	-1.34
	-4.05	-3.77	-2.23	-2.16	-1.33
	-3.21	-3.60	-2.02	-2.15	-0.91
	-2.65	-2.96	-2.47	-2.53	-0.45

Table S2d Our SAMPL3 hydration free energy submissions, continued. All values are in kcal/mol. The first four columns are SEA predictions, using the indicated forcefield. Multi-conformer SEA submissions are marked as "MC", single-conformers as "SC". The last column contains multi-conformer PB results.
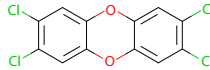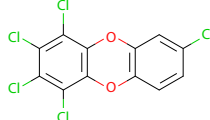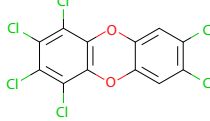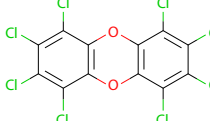
Molecule	GAFF MC	GAFF SC	OPLS MC	OPLS SC	PB GAFF
	-2.74	-2.69	-2.10	-2.05	-0.48
	-3.30	-3.23	-3.22	-3.35	-0.44
	-2.43	-2.37	-4.27	-4.29	0.15
	-2.35	-2.43	-3.82	-3.86	0.24
	-2.13	-2.18	-3.85	-4.01	0.49
	-2.16	-2.16	-4.96	-4.92	0.64
	-1.81	-1.68	-6.01	-6.30	1.07
	-1.59	-1.62	-9.42	-9.20	1.65

Table S3a Explicit solvent hydration free energy calculations. All values are in kcal/mol. Our errors are 0.10 or less, except as indicated. Experimental values, as reported by the SAMPL3 organizers, are included for reference.
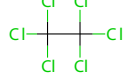
Molecule	GAFF FEP	OPLS FEP	Experiment
—	2.47	2.27	1.87
	0.80	0.57 ± 0.14	-0.39
	0.26	0.01	-0.88
	-0.46	-0.66	-1.80
	0.54	0.44	-0.26
	-0.61 ± 0.16	-1.23 ± 0.27	-1.97
	-0.07	-0.22	-1.43
	-0.65 ± 0.11	-1.16 ± 0.18	-2.37
	0.12	-0.30	-1.23
	0.94	0.82	-0.64

Table S3b Explicit solvent hydration free energy calculations, continued. All values are in kcal/mol. Our errors are 0.10 or less, except as indicated. Experimental values, as reported by the SAMPL3 organizers, are included for reference.
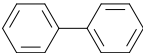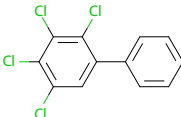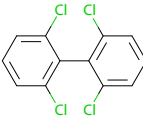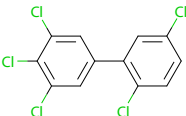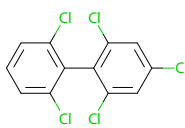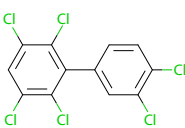
Molecule	GAFF FEP	OPLS FEP	Experiment
	-3.21	-1.49 ± 0.42	-2.23
	-2.58	-1.22	-2.69
	-1.93	-1.56	-2.46
	-1.21	-2.53	-2.16
	-1.42	-4.18 ± 0.13	-3.48
	-1.31	-1.06 ± 0.12	-2.28
	-1.11	-3.47	-3.61
	-0.57	-1.46 ± 0.11	-1.96
	-0.71	-4.97	-4.38

Table S3c Explicit solvent hydration free energy calculations, continued. All values are in kcal/mol. Our errors are 0.10 or less, except as indicated. Experimental values, as reported by the SAMPL3 organizers, are included for reference.
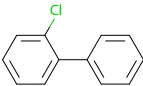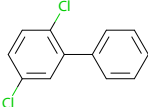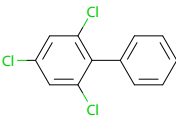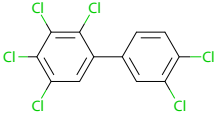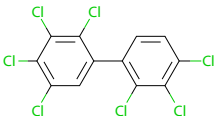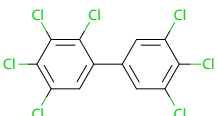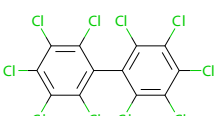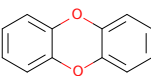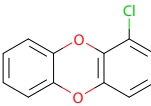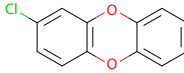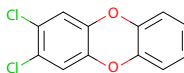
Molecule	GAFF FEP	OPLS FEP	Experiment
	-1.11	-5.60	-3.04
	-0.76	-6.29	-4.40
	-0.81	-7.03	-3.17
	-0.09	-6.41	-4.61
	0.78	-8.79	-2.98
	-3.95	-1.45	-3.15
	-4.00	-1.67	-3.52
	-3.41	-1.46	-3.10
	-3.06	-2.22	-3.56

Table S3d Explicit solvent hydration free energy calculations, continued. All values are in kcal/mol. Our errors are 0.10 or less, except as indicated. Experimental values, as reported by the SAMPL3 organizers, are included for reference.

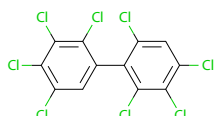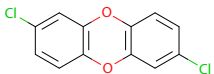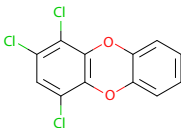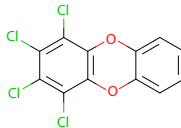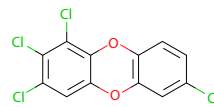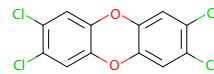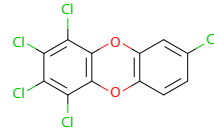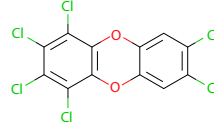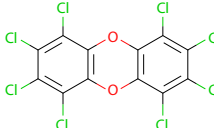| Molecule	GAFF FEP	OPLS FEP	Experiment
	-2.95	-1.76	-3.67
	-3.18	-3.25	-4.05
	-2.95	-4.94	-3.81
	-2.75	-4.00	-3.84
	-2.51	-3.76	-3.37
	-2.60	-5.69	-4.15
	-2.45	-7.03	-3.71
	-2.19	-10.70	-4.53

Table S4a Multi-conformer SEA predictions, using two different forcefields, for molecules from the SAMPL2 challenge. All values are in kcal/mol. Experimental values are included for comparison.

Molecule	SEA/GAFF	SEA/OPLS	Experiment
5-bromouracil	-16.29	-10.74	-18.39
5-chlorouracil	-16.18	-10.61	-17.82
5-flurouracil	-16.20	-12.12	-17.01
5-iodouracil	-17.53	-10.85	-18.57
5-trifluoromethyluracil	-16.85	-9.76	-15.46
6-chlorouracil	-13.72	-9.36	-16.00
acetylsalicylic acid	-10.86	-8.79	-12.33
butyl paraben	-8.34	-5.75	-8.74
caffeine	-13.72	-7.98	-12.82
cyanuric acid	-18.76	-12.18	-18.39
d-glucose	-20.12	-18.43	-25.44
d-xylose	-16.44	-13.81	-20.50
diflunisal	-6.95	-9.98	-7.63
ethyl paraben	-8.68	-6.29	-9.20
flurbiprofen (racemic)	-7.90	-7.34	-8.68
ibuprofen (racemic)	-5.58	-5.54	-7.01
ketoprofen (racemic)	-10.20	-8.45	-10.83
methyl paraben	-9.11	-6.66	-9.52
naproxen	-9.75	-7.94	-10.35
phthalimide	-11.40	-9.18	-9.73
propyl paraben	-8.66	-6.14	-9.35
sulfolane	-9.09	-10.12	-9.17
uracil	-16.08	-10.36	-16.06

Table S4b Multi-conformer SEA predictions, using two different forcefields, for molecules from the SAMPL1 challenge. All values are in kcal/mol. Experimental values are included for comparison.

Molecule	SEA/GAFF	SEA/OPLS	Experiment
1-amino-4-anilino-anthraquinone	-11.41	-9.12	-7.44
1-amino-anthraquinone	-11.38	-6.99	-7.97
1,2-dinitroxypropane	-4.70	-4.76	-4.95
1,4,5,8-tetramino-anthraquinone	-21.28	-15.69	-8.94
2-butyl nitrate	-1.48	-1.01	-1.82
4-amino-4-nitroazobenzene	-11.70	-7.85	-11.24
4-dimethylamino-azobenzene	-6.65	-2.47	-6.66
alachlor	-8.11	-6.21	-8.21
aldicarb	-7.70	-7.70	-9.84
ametryn	-10.06	-5.58	-7.65
azinphosmethyl	-14.94	-8.74	-10.03
benefin	-1.29	-4.43	-3.51
bensulfuron	-29.06	-28.16	-17.17
bromacil	-12.11	-7.03	-9.73
butyl nitrate	-1.39	-1.27	-2.09
captan	-7.25	-6.22	-9.01
carbaryl	-8.87	-5.80	-9.45
carbofuran	-9.62	-5.67	-9.61
carbophenothion	-14.18	-5.91	-6.50
chlordane	-0.35	-1.38	-3.44
chlorfenvinphos	-8.36	-5.74	-7.07
chlorimuronethyl	-25.21	-22.20	-14.01
chloropicrin	0.41	-2.45	-1.45
chlorpyrifos	-11.13	-8.01	-5.04
dialifor	-20.32	-8.69	-5.74
diazinon	-9.26	-6.85	-6.48
dicamba	-8.32	-7.74	-9.86
dichlobenil	-4.23	-4.43	-4.71
dinitramine	-9.45	-7.08	-5.66
dinoseb	-5.56	-7.21	-6.23
endosulfan alpha	-8.30	-9.32	-4.23

Table S4c Multi-conformer SEA predictions, using two different forcefields, for molecules from the SAMPL1 challenge (continued). All values are in kcal/mol. Experimental values are included for comparison.

Molecule	SEA/GAFF	SEA/OPLS	Experiment
endrin	-3.50	-3.48	-4.82
ethion	-16.80	-8.48	-6.10
ethyleneglycol mononitrate	-6.72	-6.67	-8.18
fenuron	-10.62	-6.57	-9.13
heptachlor	-0.27	-1.22	-2.55
isobutyl nitrate	-1.21	-1.19	-1.88
isophorone	-3.69	-2.30	-5.18
lindane	-1.54	-4.02	-5.44
malathion	-13.64	-8.11	-8.15
methomyl	-9.53	-6.22	-10.65
methyparathion	-10.35	-7.46	-7.19
metsulfuronmethyl	-26.21	-27.66	-15.54
nitralin	-9.04	-12.85	-7.98
nitroglycol	-4.86	-5.89	-5.73
nitroxyacetone	-6.91	-6.74	-5.99
oxamyl	-13.14	-9.10	-10.18
parathion	-8.77	-6.96	-6.74
pebulate	-3.81	-2.32	-3.64
phorate	-10.82	-5.11	-4.37
pirimor (pirimicarb)	-12.32	-6.23	-9.41
profluralin	-1.19	-3.99	-2.45
prometryn	-9.66	-4.62	-8.43
propanil	-9.55	-8.15	-7.78
pyrazon	-17.68	-10.60	-16.43
simazine	-10.65	-5.39	-10.22
sulfometuron-methyl	-24.61	-28.63	-20.25
terbacil	-11.22	-6.83	-11.14
terbutryn	-10.12	-4.51	-6.68
thifensulfuron	-26.25	-29.09	-16.23
trichlorfon	-12.15	-9.59	-12.74
trifluralin	-1.39	-4.10	-3.25
vernolate	-3.47	-2.16	-4.13

Table S4d Multi-conformer SEA predictions, using two different forcefields, for molecules from the SAMPL0 challenge. All values are in kcal/mol. Experimental values are included for comparison.

Molecule	SEA/GAFF	SEA/OPLS	Experiment
1,1-diacetoxyethane	-6.81	-3.15	-4.97
1,1-diethoxyethane	-2.00	-1.78	-3.28
1,2-diethoxyethane	-3.03	-1.78	-3.54
1,4-dioxane	-3.05	-1.69	-5.05
benzyl bromide	-1.60	-2.91	-2.38
benzyl chloride	-1.34	-1.31	-1.93
bis-2-chloroethyl ether	-3.36	-2.05	-4.23
diethyl propanedioate	-6.74	-3.79	-6.00
diethyl sulfide	-0.17	-0.43	-1.43
dimethoxymethane	-4.37	-2.18	-2.93
ethylene glycol diacetate	-7.97	-3.61	-6.34
glycerol triacetate	-10.76	-5.81	-8.84
imidazole	-8.38	-7.42	-9.81
m-bis(trifluoromethyl)benzene	-0.81	-0.40	1.07
N,N-4-trimethylbenzamide	-8.38	-4.11	-9.76
N,N-dimethyl-p-methoxybenzamide	-6.75	-4.83	-11.01
phenyl formate	-5.54	-2.75	-3.82
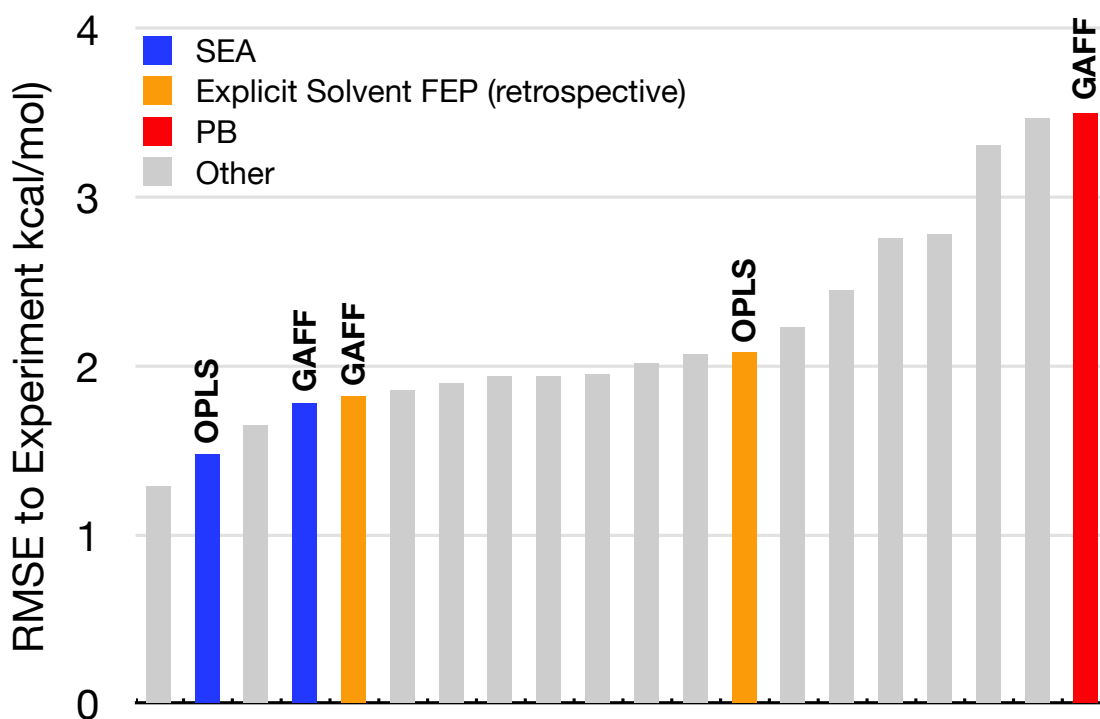
Figure 4.6: **Fig. S1** Overall performance of the SAMPL3 submissions, as compared to our retrospective explicit solvent FEP results. Multi-conformer SEA submissions have been omitted for clarity. RMSEs are bootstrapped numbers for SAMPL submissions, and averages for FEP calculations. For our SEA submissions, the bootstrap vs. average difference was less than a tenth of a kcal/mol.

Chapter 5

SurfMap: Mapping the Solvation Energy Landscape

Authors: Charles Kehoe, Christopher Fennell, and Ken Dill

Not yet published

Abstract

Here we present a new solvation visualization tool, aimed at showing the energetic behavior of water on the surface of a solute. Dubbed SurfMap, the system can be used with any solvation model capable of producing the relevant energetic and geometric information. We show SurfMap results with our Semi-Explicit Assembly (SEA) water model, along with some Poisson-Boltzmann results for comparison. SurfMap illustrates how SEA provides more detailed and accurate information about solvation, for both small and large molecules. SurfMap, like the SEA water model, is fast and widely applicable. SEA visualization provides dramatically faster performance and more information than similar explicit-solvent-based methods.

5.1   Introduction

Accurate solvation is one of the most important and difficult problems in molecular simulation. Many water models are available, varying widely in speed and accuracy. Because of the inherent complexity of solvation, water models often work well on some classes of molecules, and cause serious problems with others. It can often be very difficult to understand why this happens or how it might be fixed.

To address these and other problems, we developed a solvation visualization tool named SurfMap, which generates three-dimensional colored maps showing where sol-

vent molecules are on a solute surface, and how energetically favorable their placement is. This makes it immediately apparent how and where the solvent's interaction with the molecule results in a molecular solvation energy. Using SEA, SurfMap can generate maps for any molecule using any combination of forcefield and explicit water model. It is also compatible with any implicit solvent model that generates per-atom energetic information; we have adapted it to map results from the PB solvation model, for example. This allows direct comparisons between solvation models, as we will show below.

Because solvation is so important, this system has many other potential applications as well. Both ligand and protein interactions often depend heavily on solvent behavior, which can be difficult and expensive to quantify. Being able to directly see water's behavior in these situations could be of immense value. This graphical system makes it easier to understand, assess, and learn from solvent models. Energetics are displayed visually, in up to sub-atomic resolution, providing much more information than typical residue- or molecule-based assessments. Map construction is nearly as fast as the implicit solvent modeling itself.

Other approaches for mapping the behavior of water in molecular systems exist, such as Schrodinger's WaterMap system.[3] However, WaterMap typically entails several million steps of Monte Carlo simulation in full explicit solvent, followed by a complex statistical analysis. In the protein systems to which it is often applied, this can be quite computationally expensive. Furthermore, the end product is not a full distri-

bution of water behavior, as provided by our system, but a few highlighted spots in which waters are alleged to localize at various energy levels. Focusing on a few small, fixed waters sites may be conceptually convenient, but it is not as realistic or informative as a full-surface analysis of molecular solvation at sub-atomic resolution. This is the result that SEA and SurfMap, working together, can provide.

5.2   Methods

Structure Preparation

All small molecules were parameterized with GAFF and AM1-BCC charges, using the ANTECHAMBER program in AmberTools 1.4.[138] For a more realistic solvation surface, zero-size hydrogens were given an atomic radius of 1.2 Å . Molecules were first relaxed in explicit solvent with the TIP3P water model,[64] following the practices of Mobley et al.[95]

We performed all molecular dynamics (MD) calculations with GROMACS 4.0.4.[12,55] We switched off Lennard-Jones interactions between 8 and 9 Å and applied long-ranged energy and pressure corrections. We used smooth particle-mesh Ewald for long-ranged electrostatics accumulation,[35] this with a real-space cutoff of 10 Å, a spline order of 6, fourier spacing of 1 Å, and a real-space energy tolerance parameter of 10^{-6} kJ/mol. We used the SETTLE algorithm to constrain the geometry of water molecules in explicit-solvent simulations,[91] and the LINCS algorithm to constrain covalent bonds to hydrogen atoms on solute molecules.[53]

We solvated the target solutes in a rhombic dodecahedral box with a 10 Å buffer of explicit water between the atoms of the solute and the edges of the box. We relaxed these systems in two steps, first with 1000 steps of steepest descent minimization followed by 10 ps of constant energy MD with a 1 fs timestep. All timesteps after this relaxation phase were 2 fs. We followed relaxation with 100 ps of constant temperature LD (300 K), 100 ps of constant pressure (1 atm) equilibration using the Berendsen thermostat, and 700 ps of constant pressure dynamics using the Parrinello-Rahman barostat. We rescaled the simulation box to the average volume from last 500 ps of this constant pressure trajectory, and equilibrated this system with 500 ps of constant volume simulation before recording our fully relaxed structure. We stripped the waters from this structure and fed it into SEA for analysis.

We obtained the BphC structure from RCS PDB structure 1DHY. We cut out just domain 1 (residues 1-135) and restored missing structure using Swiss-PdbViewer 4.0.4.[49] The pdb2gmx program from GROMACS 4.0.4 then assigned GAFF forcefield parameters, using ffamber.[31,122]

Solvation

For the small molecules, we ran 100 iterations of SEA using a surface segmentation detail of 5. For BphC, we used only 10 iterations.

We performed our PB tests by combining a polar term and a non-polar term. For the polar term, we solved the linearized Poisson equation using APBS.[6] We used a probe radius of 0.65 Å , a solvent dielectric constant of 78, and a 65 x 65 x 65 grid with 0.25 Å spacing. We calculated a Weeks-Chandler-Andersen nonpolar term using APBS, with $\gamma = 0.0085$ kJ per mol-Å .

We then needed to convert our PB results to a water-energy point cloud of the type that SEA outputs. To begin, we generated molecular hydration surfaces in the same way SEA does, with a surface segmentation detail of 12, but at a uniform distance of 1.4 Å from the Lennard-Jones surface. APBS provides an electrostatic potential distribution, as well as nonpolar contributions and exposed surface areas for each atom. We calculated the polar solvation energy contributed by each atom by multiplying its charge by the (interpolated) electrostatic potential at the atomic center. To this we added the atom's nonpolar contribution. But we still wanted a per-water energy, as opposed to a per-atom measurement.

To solve this problem, we recalled from our work with SEA that the first solvation shell lies within approximately 2 Å of the atomic surface. Using the exposed surface area for each atom, we can estimate the volume of the partial solvation shell. The shell volume for a fully exposed atom with radius r_i angstroms would be

$$V_{i0} = \frac{4}{3}\pi((r_i + 2)^3 - r_i^3) = 8\pi(r_i^2 + 2r_i + \frac{4}{3})$$

This corresponds to a fully exposed surface, with area

$$A_{i0} = 4\pi r_i^2$$

We can then estimate the volume of the partial solvation shell by taking the portion of the volume indicated by the exposed area A_i:

$$V_i = \frac{A_i}{A_{i0}}V_{i0} = \frac{A_i}{r_i}(2r_i + 4 + \frac{8}{3r_i})$$

Multiplying by the density ρ_0 of bulk water (at 300K and 1 atm) then gives an estimate of the number of waters solvating the atom. We use this estimate to scale the PB per-atom solvation energy to a per-water number.

Mapping

For each solvation method, we collected all water locations generated to locate the solvation surface. We estimated surface normals by fitting a plane to the 30 nearest neighbors of each point. We then ran these sets of points through the Poisson Surface Reconstruction algorithm[68] as implemented in the Computer Graphics Algorithms Library.[1] This produced an implicit surface function, which CGAL then tessellated to generate a triangle mesh. Tessellation used a maximum triangle radius of 0.2 Å, a maximum surface approximation error of 0.5 Å, and a minimum triangle angle of 20 degrees.

Because of the way the surface reconstruction works, the mesh vertices we get out are not usually related to the surface points we put in. We are then left with the task of assigning colors to all the mesh vertices to represent energies. To do this, we locate the 5 nearest original points p_i for each mesh vertex v, and estimate the water energy at v by weighting the energies of the nearby p_i according to their inverse square distance from v. Each vertex is then assigned a color to represent the water energy at that point, and our 3D viewer blends the colors across the surface. For reference, we make the solvation surface partially transparent, and display the solute atoms as sphere (using neutral colors) underneath.

5.3   Results

Our results show how SEA is able to model water behavior similar to explicit solvent, resulting in more correct solvation energies than other lightweight solvent models. SEA provides high-resolution information about water energies and water-solute distances. It also shows asymmetric charge behavior, and can solvate large molecules.

SEA provides localized, realistic solvation behavior

Figure 5.1 shows solvations of phenol, illustrating how SEA produces correct, detailed water behavior, and our solvation visualizer clearly shows what various solvent models

119
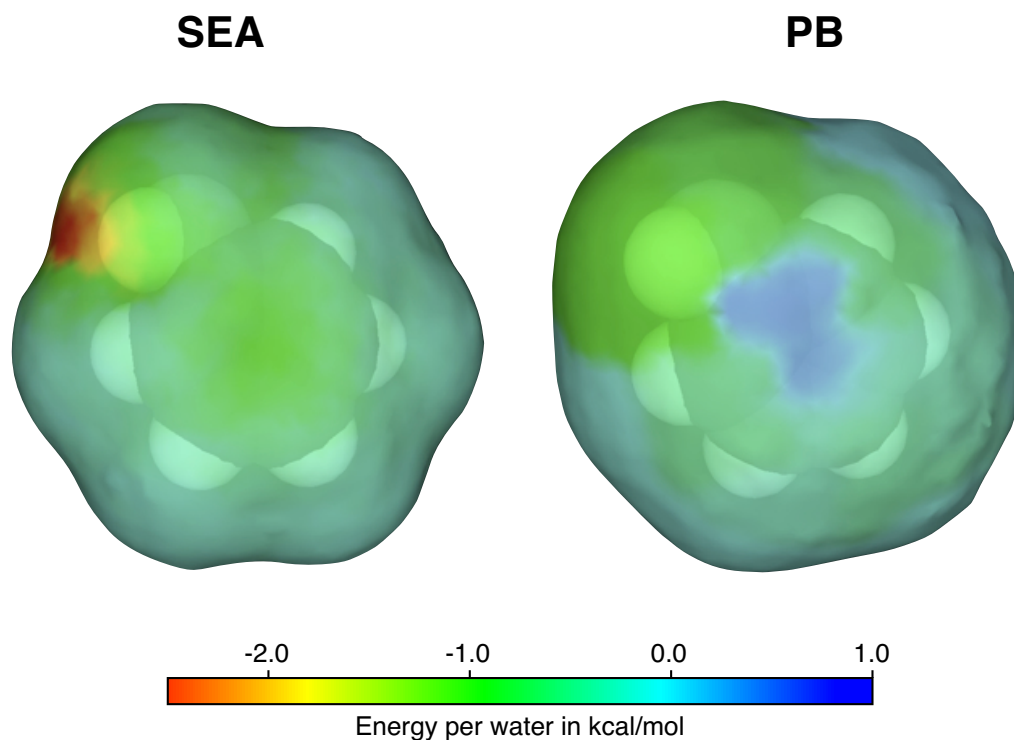
Figure 5.1: Phenol as solvated by SEA and PB. SEA highlights the exposed proton on the hydroxyl group, the center of the aromatic plane, and the back of the hydroxyl oxygen. PB, however, is unable to provide detail any more fine-grained than individual atoms, so we get only a general indication that the hydroxyl is favorable. However, this is juxtaposed against a strikingly unfavorable spot on the carbon nearest the hydroxyl group. This unphysical behavior underscores PB's inability to provide detailed solvation information, and may help to explain why SEA's solvation free energies are around 1 kcal/mol closer to explicit solvent (and experimental) results for this molecule.

are doing. SEA is able to highlight very specific spots that contribute to favorable solvation, such as the exposed protons on hydroxyl groups. Notice also the favorable spots in the center of the aromatic planes, and behind the hydroxyl oxygens, where lone pairs would attract hydrogen bonds.

Figure 5.2 shows how SEA is able to capture the non-additive solvation behavior exhibited by the isomers of dihydroxybenzene.[37] These isomers have sharply different solvation free energies, due to the relative positioning of their hydroxyl groups. A simple additive solvation approach would not pick up on these differences. However, SEA clearly shows how the water has full access to some hydroxyls, and only partial access to others. This difference has important energetic consequences.

SEA adjusts water distances for accurate solvation

Figure 5.3 shows two solvations of dimethylamine, from both SEA and PB. Again, SurfMap reveals vivid geometric and energetic differences between solvation models. SEA provides information about how close waters get to solute atoms, while PB does not. As a result, SEA calculates a solvation free energy much closer to that of explicit solvent.

-8.0 kcal/mol   -6.2 kcal/mol

Energy per water in kcal/mol
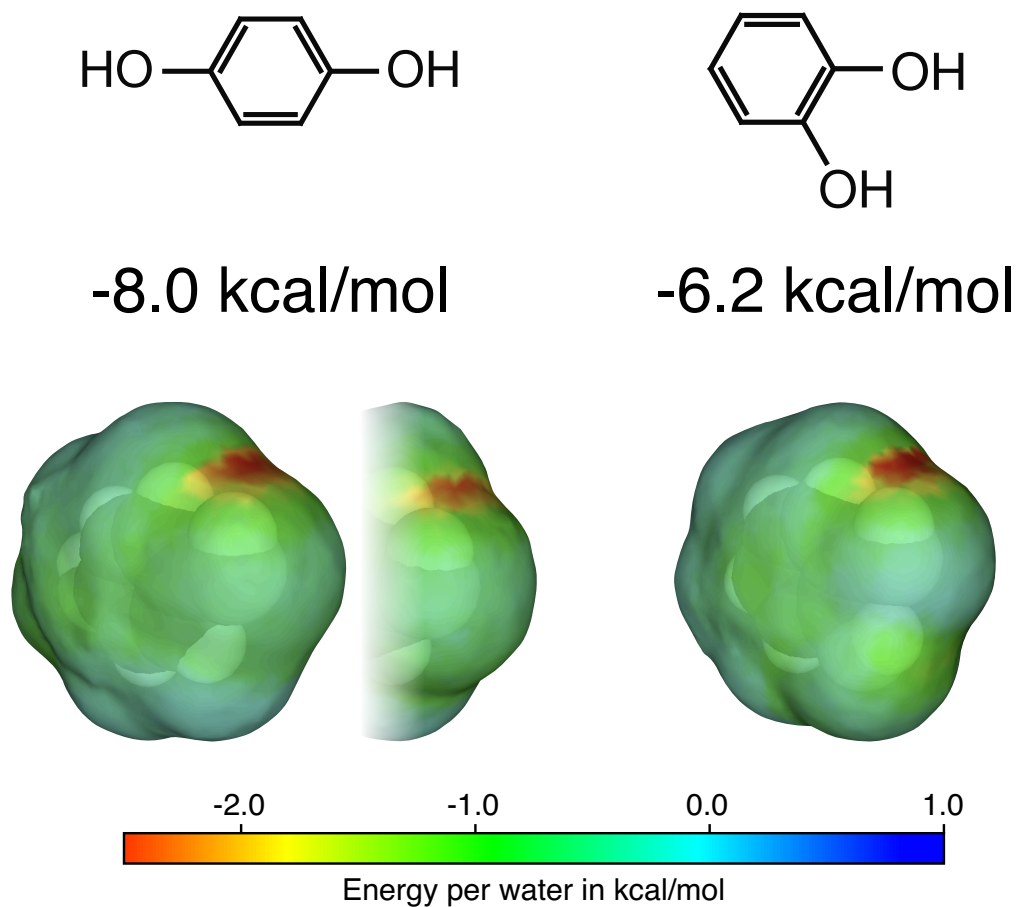
Figure 5.2: 1,4-dihydroxybenzene and 1,2-dihydroxybenzene as solvated by SEA. The 1,4 isomer has hydroxyl groups on opposite ends of the molecule, as shown, and SEA highlights the favorable tips of both of them. In the 1,2 isomer, however, the first hydroxyl proton is exposed to solvent while the second is occluded by the first. Accordingly, SEA highlights the first much more strongly than the second. The favorable regions on the aromatic plane and backs of the oxygens are still present.

SEA: -2.2 kcal/mol

PB: -0.1 kcal/mol

Explicit: -2.7 kcal/mol

Energy per water in kcal/mol
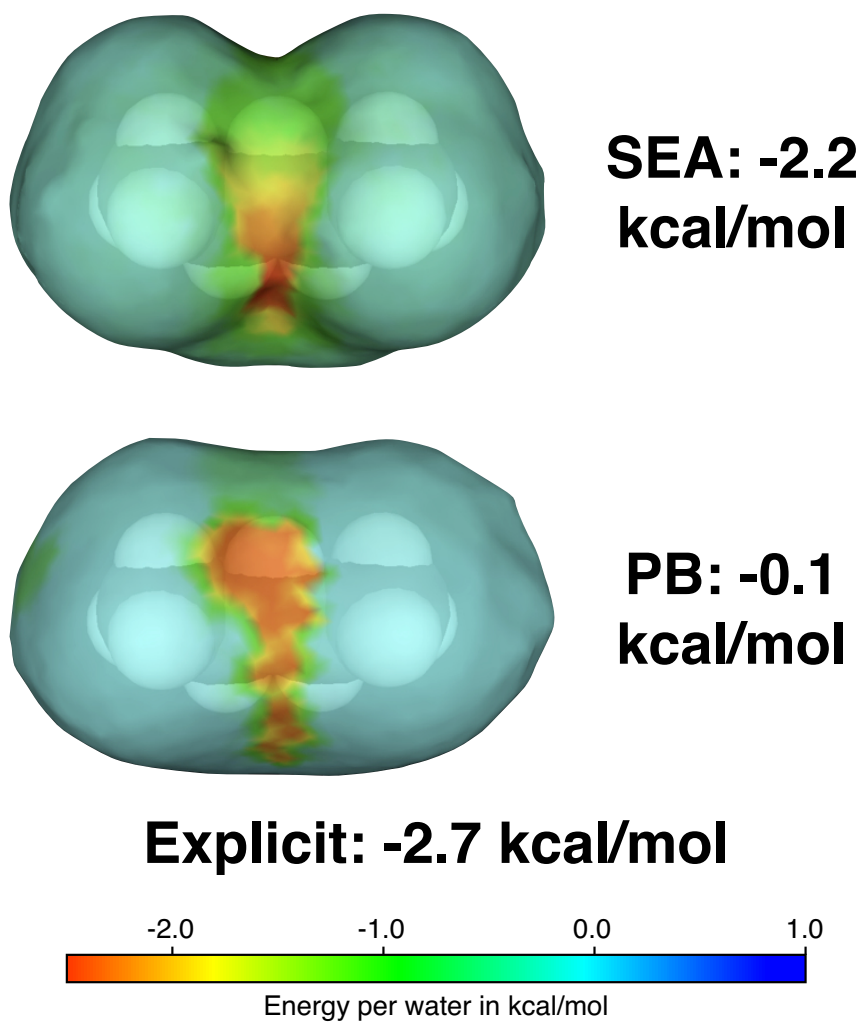
Figure 5.3: Diethylamine as solvated by SEA and PB. SEAs solvation plot shows the waters tightening around the nitrogen and its proton, in response to their strong charges. But PB is unable to report anything but a generic solvation (probe) radius, so the waters attraction is not as evident in the center of the molecule. This results in much less accurate results for the solvation free energy.
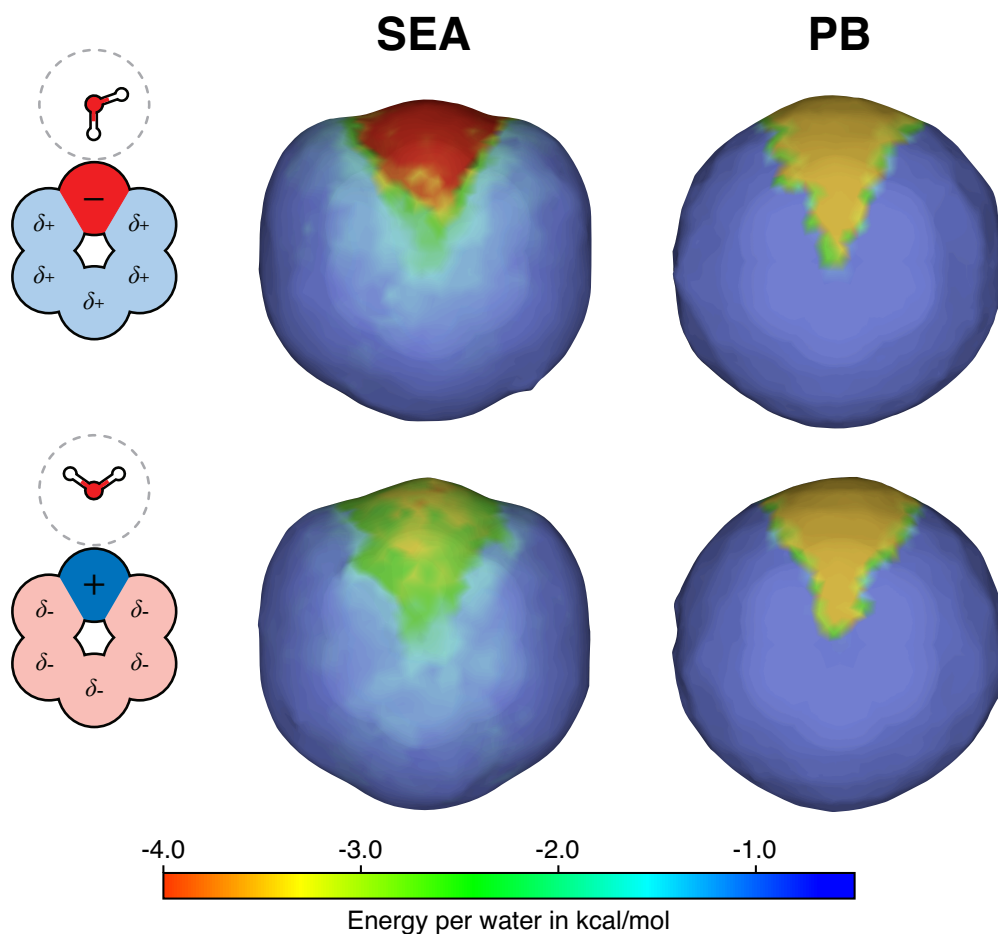
Figure 5.4: SEA solvation of charge asymmetry "bracelets". These molecules, shown schematically on the left, are identical, except that all charges are reversed. As SEA correctly predicts, waters protons are able to get closer to the negative charge than its oxygen can get to the positive one, so the negative bracelet is more favorably solvated. This difference can also be seen in explicit solvent, but most implicit models (like PB) treat these molecules identically.

SEA reproduces water asymmetry

The two molecules in Figure 5.4 do not really exist, but they vividly illustrate the serious charge-asymmetry problem suffered by many implicit solvation models. We have studied these hexagonal bracelets before:[92] they consist of a full formal charge bonded to five partial, opposing charges in a planar hexagonal ring, forming a neutral molecule. While most implicit solvent models will calculate identical solvation free energies for these molecules, SEA and explicit are able to uncover several kcal/mol of differences caused by water's asymmetric charge behavior.

SEA handles large molecules

Figure 5.5 shows a SEA solvation of a 135-residue domain from the protein BphC. This map took several minutes to produce, but provides as much solvation information as a multi-day explicit solvent simulation. BphC folds into two separate domains, whose inner hydrophobic faces then stick together. A comparison to a map of residue hydrophobicities shows SEA's accuracy, and the additional information it provides over coarser approaches.
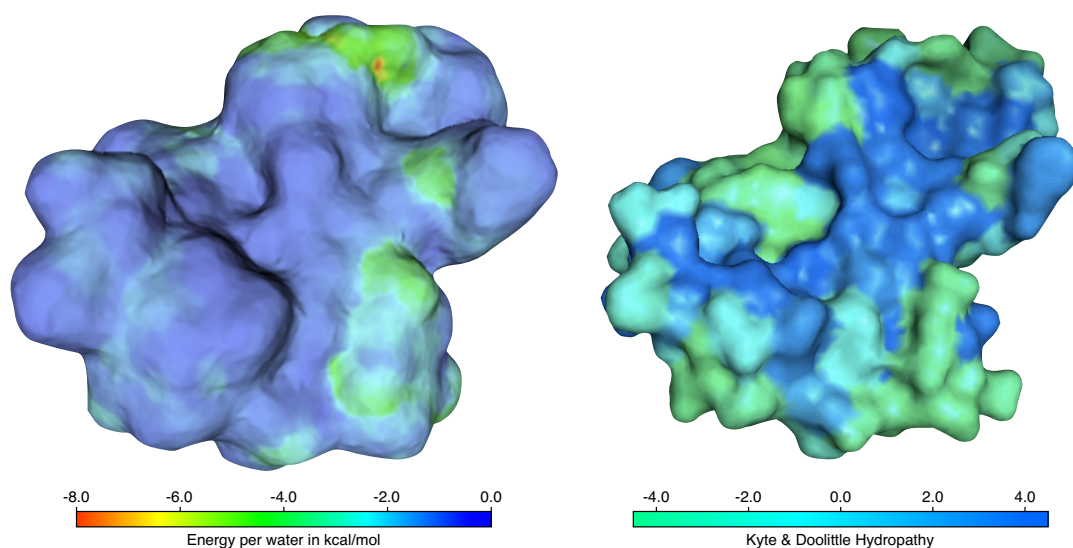
Figure 5.5: Pictured are two maps of the hydrophobic (dimerizing) side of BphC, domain 1. On the left is the SEA solvation map, showing detailed solvent behavior at sub-atomic resolution. Hydrophobic faces and pockets are clearly highlighted, as are solvent-facing edge regions. In the upper-right groove, surrounding charges collude to create an extremely favorable spot for water, or possibly salt bridging. On the right is a simpler solvation map, with residues colored by the Kyte & Doolittle hydropathy.[72] The map generally concurs with SEA, including several brighter spots on the solvent-facing edges, but clearly cannot match SEA's level of detail.

5.4 Conclusions

We have shown results from SurfMap, a new solvation visualization tool, which is able to show solvation energetics and geometry in sub-atomic detail. We have used it to compare and contrast solvation model performance, and to understand factors that may lend accuracy to one model or another. We have been able to see how the SEA water model applies explicit solvent physics at implicit solvent speeds. SurfMap is fast and easy to apply to molecules large and small, from individual ligand and functional groups to full proteins. Since solvent effects play such an important role in all molecular behavior and interactions, SurfMap could prove to be of immense value to the molecular modeling and design community.

Bibliography

[1] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[2] Johan Åqvist. Ion–water interaction potentials derived from free energy perturbation slmulations. *J. Phys. Chem.*, 94(21):8021–8024, 1990.

[3] Robert Abel, Noeris K. Salam, John Shelley, Ramy Farid, Richard A. Friesner, and Woody Sherman. Contribution of explicit solvent effects to the binding affinity of small-molecule inhibitors in blood coagulation factor serine proteases. *ChemMedChem*, 6(6):1049–1066, 2011.

[4] Michael H. Abraham, Gary S. Whiting, Richard Fuchs, and Eric J. Chambers. Thermodynamics of solute transfer from water to hexadecane. *J. Chem. Soc. Perkin Trans. 2*, pages 291–300, 1990.

[5] M. P. Allen and D. J. Tildesley. *Computer Simulations of Liquids*. Oxford University Press, New York, 1987.

[6] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *PNAS*, 98(18):10037–10041, August 2001.

[7] Dmitrii Beglov and Benoît Roux. Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. *J. Chem. Phys.*, 100(12):9050–9063, 1994.

[8] A. Ben-Naim and Y. Marcus. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.*, 81(4):2016–2027, 1984.

[9] Arieh Ben-Naim. *Solvation Thermodynamics*. Plenum Press, New York, 1987.

[10] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. Simple point charge water. In B. Pullman, editor, *Intermolecular Forces*, pages 331–342. Reidel, Dordrecht, 1981.

[11] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.

[12] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.

[13] Kallol M. Biswas, Daniel R. DeVido, and John G. Dorsey. Evaluation of methods for measuring amino acid hydrophobicities and interactions. *J. Chromatogr. A*, 1000:637–655, 2003.

[14] Par Bjelkmar, Per Larsson, Michel A. Cuendet, Berk Hess, and Erik Lindahl. Implementation of the charmm force field in gromacs: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *Journal of Chemical Theory and Computation*, 6(2):459–466, 2010.

[15] A. J. Bordner, C. N. Cavasotto, and R. A. Abagyan. Accurate transferable model for water, n-octanol, and n-hexadecane solvation free energies. *J. Phys. Chem. B*, 106(42):11009–11015, 2002.

[16] Kevin J. Bowers, Edmond Chow, Huafeng Xu, Ron O. Dror, Michael P. Eastwood, Brent A. Gregersen, John L. Klepeis, Istvan Kolossvary, Mark A. Moraes, Federico D. Sacerdoti, John K. Salmon, Yibing Shan, and David E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, SC '06, New York, NY, USA, 2006. ACM.

[17] D. Bratko, L. Blum, and A. Luzar. A simple model for the intermolecular potential of water. *J. Chem. Phys.*, 83(12):6367–6370, 1985.

[18] Sergio Cabani, Paolo Gianni, Vincenzo Mollica, and Luciano Lepori. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Solution Chem.*, 10(8):563–595, 1981.

[19] D. A. Case, T. A. Darden, III T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F.Wong, F. Paesani, J. Van-

icek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, and P. A. Kollman. *AMBER 10*. University of California, San Francisco, 2008.

[20] Aviel Chaimovich and M. Scott Shell. Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy. *Phys. Chem. Chem. Phys.*, 11:1901–1915, 2009.

[21] Candee C. Chambers, Gregory D. Hawkins, Christopher J. Cramer, and Donald G. Truhlar. Model for aqueous solvation based on class IV atomic charges and first solvation shell effects. *J. Phys. Chem.*, 100(40):16385–16398, 1996.

[22] D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437:640–647, 2005.

[23] David Leonard Chapman. Li. a contribution to the theory of electrocapillarity. *Philosophical Magazine Series 6*, 25(148):475–481, 1913.

[24] J. Chen and C. L. Brooks III. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.*, 10:471–481, 2008.

[25] I. Chorny, K. A. Dill, and M. P. Jacobson. Surfaces affect ion pairing. *J. Phys. Chem. B*, 109(50):24056–24060, 2005.

[26] Michael L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.

[27] J. Cornette, K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, 195:659–685, 1987.

[28] Christopher J. Cramer and Donald G. Truhlar. An scf solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Science*, 256(5054):213–217, 1992.

[29] Tom Darden, David Pearlman, and Lee G. Pedersen. Ionic charging free energies: Spherical versus periodic boundary conditions. *J. Chem. Phys.*, 109(24):10921–10935, 1998.

[30] Yuqing Deng and Benoît Roux. Hydration of amino acid side chains: Nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules. *J. Phys. Chem. B*, 108(42):16567–16576, 2004.

[31] Allison J. DePaul, Erik J. Thompson, Sarav S. Patel, Kristin Haldeman, and Eric J. Sorin. Equilibrium conformational dynamics in an rna tetraloop from massively parallel molecular dynamics. *Nucleic Acids Research*, 38(14):4856–4867, 2010.

[32] K. A. Dill. Additivity principles in biochemistry. *J. Biol. Chem.*, 272(2):701–704, 1997.

[33] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

[34] U. Essman, L. Perela, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8592, 1995.

[35] U. Essman, L. Perela, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103(19):8577–8592, 1995.

[36] C. J. Fennell, C. Kehoe, and K. A. Dill. Oil/water transfer is partly driven by molecular shape, not just size. *J. Am. Chem. Soc.*, 132(1):234–240, 2010.

[37] Christopher Fennell and Ken Dill. Physical modeling of aqueous solvation. *Journal of Statistical Physics*, 145:209–226, 2011. 10.1007/s10955-011-0232-9.

[38] Christopher J. Fennell and J. Daniel Gezelter. On the structural and transport properties of the soft sticky dipole and related single-point water models. *J. Chem. Phys.*, 120(19):9175–9184, 2004.

[39] Christopher J. Fennell, Charles W. Kehoe, and Ken A. Dill. Modeling aqueous solvation with semi-explicit assembly. *Proc. Natl. Acad. Sci. USA*, 108(8):3234–3239, February 2011.

[40] Christopher J. Fennell, Charlie Kehoe, and Ken A. Dill. Oil/Water transfer is partly driven by molecular shape, not just size. *J. Am. Chem. Soc.*, 132(1):234–240, January 2010.

[41] Jens Fennen, Andrew E. Torda, and Wilfred F. van Gunsteren. Structure refinement with molecular dynamics and a boltzmann-weighted ensemble. *J. Biomol. NMR*, 6:163–170, 1995.

[42] Jan Florián and Arieh Warshel. Langevin dipoles model for ab initio calculations of chemical processes in solution: Parametrization and application to hydration free energies of neutral and ionic solutes and conformational analysis in aqueous solution. *J. Phys. Chem. B*, 101(28):5583–5595, 1997.

[43] F. Floris and J. Tomasi. Evaluation of the dispersion contribution to the solvation energy. a simple computational model in the continuum approximation. *J. Comp. Chem.*, 10(5):616–627, 1989.

[44] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications.* Academic Press, New York, 1996.

[45] E. Gallicchio and R. M. Levy. Agbnp: An analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comp. Chem.*, 25(4):479–499, 2004.

[46] E. Gallicchio, L. Y. Zhang, and R. M. Levy. The sgb/np hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comp. Chem.*, 23(5):517–529, 2002.

[47] Matthew T. Geballe, A. Geoffrey Skillman, Anthony Nicholls, J. Peter Guthrie, and Peter J. Taylor. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput. Aided Mol. Des.*, 24:259–279, 2010.

[48] Raphaël Geney, Melinda Layten, Roberto Gomperts, Viktor Hornak, and Carlos Simmerling. Investigation of salt bridge stability in a generalized born solvent model. *J. Chem. Theory Comput.*, 2(1):115–127, 2006.

[49] Nicolas Guex and Manuel C. Peitsch. Swiss-model and the swiss-pdb viewer: An environment for comparative protein modeling. *ELECTROPHORESIS*, 18(15):2714–2723, 1997.

[50] J. Peter Guthrie. A blind challenge for computational solvation free energies: Introduction and overview. *J. Phys. Chem. B*, 113(14):4501–4507, 2009.

[51] J. Peter Guthrie. A blind challenge for computational solvation free energies: Introduction and overview. *J. Phys. Chem. B*, 113(14):4501–4507, 2009.

[52] E. Helfand, H. Reiss, H. L. Frisch, and J. L. Lebowitz. Scaled particle theory of fluids. *J. Chem. Phys.*, 33(5):1379–1385, 1960.

[53] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[54] Berk Hess. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.*, 116(1):209–217, 2002.

[55] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[56] Berk Hess and Nico F. A. van der Vegt. Hydration thermodynamic properties of amino acid analogues: A systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B*, 110(35):17616–17626, 2006.

[57] D. M. Huang, P. L. Geissler, and D. Chandler. Scaling of hydrophobic solvation free energies. *J. Phys. Chem. B*, 105(28):6704–6709, 2001.

[58] G. Hummer, S. Garde, A. E. García, A. Pohorille, and L. R. Pratt. An information theory model of hydrophobic interactions. *Proc. Natl. Acad. Sci. USA*, 93:8951–8955, 1996.

[59] Gerhard Hummer, Lawrence R. Pratt, and Angel E. García. Free energy of ionic hydration. *J. Phys. Chem.*, 100(4):1206–1215, 1996.

[60] Wonpil Im, Simon Bernèche, and Benoît Roux. Generalized solvent boundary potential for computer simulations. *J. Chem. Phys.*, 114(7):2924–2937, 2001.

[61] Sergei Izvekov and Gregory A. Voth. Multiscale coarse-graining of mixed phospholipid/cholesterol bilayers. *J. Chem. Theory Comput.*, 2(3):637–648, 2006.

[62] Araz Jakalian, B. L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. *Journal of Computational Chemistry*, 21(2), January 2000.

[63] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. method. *J. Comput. Chem.*, 21(2):132–146, 2000.

[64] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[65] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[66] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[67] Mika A. Kastenholz and Philippe H. Hünenberger. Computation of methodology-independent ionic solvation free energies from molecular simulations. II. the hydration free energy of the sodium cation. *J. Chem. Phys.*, 124(22):224501, 2006.

[68] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[69] Gregory King and Arieh Warshel. A surface constrained allatom solvent model for effective simulations of polar solutions. *J. Chem. Phys.*, 91(6):3647–3661, 1989.

[70] Peter Kolb, Rafaela S Ferreira, John J Irwin, and Brian K Shoichet. Docking and chemoinformatic screens for new ligands and targets. *Current Opinion in Biotechnology*, 20(4):429 – 436, 2009. ¡ce:title¿Protein technologies / Systems and synthetic biology¡/ce:title¿.

[71] Peter Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, 93(7):2395–2417, 1993.

[72] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–132, 1982.

[73] Wendell M. Latimer, Kenneth S. Pitzer, and Cyril M. Slansky. The free energy of hydration of gaseous ions, and the absolute potential of the normal calomel electrode. *J. Chem. Phys.*, 7(2):108–111, 1939.

[74] B. Lee and R. M. Richards. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55(3):379–400, 1971.

[75] Jooyoung Lee, Sitao Wu, and Yang Zhang. Ab initio; protein structure prediction. In Daniel John Rigden, editor, *From Protein Structure to Function with Bioinformatics*, pages 3–25. Springer Netherlands, 2009.

[76] Juyong Lee, Jin-Soo Kim, and Chaok Seok. Cooperativity and specificity of cys2his2 zinc finger protein-dna interactions: A molecular dynamics simulation

study. *The Journal of Physical Chemistry B*, 114(22):7662–7671, 2010. PMID: 20469897.

[77] M. S. Lee, F. R. Salsbury, Jr., and M. A. Olson. An efficient hybrid explicit/implicit solvent method for biomolecular simulations. *J. Comput. Chem.*, 25(16):1967–1978, 2004.

[78] Ronald M. Levy and Emilio Gallicchio. Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.*, 49(1):531–567, 1998.

[79] Ronald M. Levy, Linda Y. Zhang, Emilio Gallicchio, and Anthony K. Felts. 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):9523–9530, 2003.

[80] Jens P. Linge, Mark A. Williams, Christian A. E.M. Spronk, Alexandre M. J. J. Bonvin, and Michael Nilges. Refinement of protein structures in explicit solvent. *Proteins: Struct. Funct. Genet.*, 50(3):496–506, 2003.

[81] Yi Liu and Toshiko Ichiye. Soft sticky dipole potential for liquid water: A new model. *J. Phys. Chem.*, 100(7):2723–2730, 1996.

[82] Valère Lounnas, Susanna K. Lüdemann, and Rebecca C. Wade. Towards molecular dynamics simulation of large proteins with a hydration shell at constant pressure. *Biophys. Chem.*, 78:157–182, 1999.

[83] R. Luo, L. David, and M. K. Gilson. Accelerated poisson-boltzmann calculations for static and dynamic systems. *J. Comput. Chem.*, 23(13):1244–1253, 2002.

[84] Justin L. MacCallum and D. Peter Tieleman. Calculation of the water–cyclohexane transfer free energies of neutral amino acid side-chain analogs using the opls all-atom force field. *J. Comput. Chem.*, 24(15):1930–1935, 2003.

[85] Alan E. Mark and Wilfred F. van Gunsteren. Decomposition of the free energy of a system in terms of specific interactions: Implications for theoretical and experimental studies. *J. Mol. Biol.*, 240:167–176, 1994.

[86] Siewert J. Marrink, Alex H. de Vries, and Alan E. Mark. Coarse grained model for semiquantitative lipid simulations. *J. Phys. Chem. B*, 108(2):750–760, 2004.

[87] Marc A. Mart-Renom, Ashley C. Stuart, Andrs Fiser, Roberto Snchez, Francisco Melo, and Andrej ali. Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):291–325, 2000.

[88] T. Mavromoustakos, S. Durdagi, C. Koukoulitsa, M. Simcic, M. G Papadopoulos, M. Hodoscek, and S. Golic Grdadolnik. Strategies in the rational drug design. *Current Medicinal Chemistry*, 18(17):2517–2530, 2011.

[89] Julien Michel and Jonathan Essex. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *Journal*

of Computer-Aided Molecular Design, 24:639–658, 2010. 10.1007/s10822-010-9363-3.

[90] S. Miertuš, E. Scrocco, and J. Tomasi. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.*, 55(1):117–129, 1981.

[91] S. Miyamoto and P. A. Kollman. SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water models. *J. Comput. Chem.*, 13(8):952–962, 1992.

[92] D. L. Mobley, A. E. Barber II, C. J. Fennell, and K. A. Dill. Charge asymmetries in hydration of polar solutes. *J. Phys. Chem. B*, 112(8):2405–2414, 2008.

[93] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.*, 5(2):350–358, 2009.

[94] David Mobley. Lets get honest about sampling. *Journal of Computer-Aided Molecular Design*, 26:93–95, 2012. 10.1007/s10822-011-9497-y.

[95] David Mobley, Shaui Liu, David Cerutti, William Swope, and Julia Rice. Alchemical prediction of hydration free energies for sampl. *Journal of Computer-Aided Molecular Design*, pages 1–12. 10.1007/s10822-011-9528-8.

[96] David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, and Ken A. Dill. Predictions of hydration free energies from all-atom molecular dynamics simulations. *J. Phys. Chem. B*, 113(14):4533–4537, 2009.

[97] David L. Mobley, Christopher I. Bayly, Matthew D. Cooper, Michael R. Shirts, and Ken A. Dill. Small molecule hydration free energies in explicit solvent: An extensive test of Fixed-Charge atomistic simulations. *J. Chem. Theory Comput.*, 5(2):350–358, February 2009.

[98] David L. Mobley, É. Dumont, John D. Chodera, and Ken A. Dill. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. *J. Phys. Chem. B*, 111(9):2242–2254, 2007.

[99] Anthony Nicholls, David L. Mobley, J. Peter Guthrie, John D. Chodera, Christopher I. Bayly, Matthew D. Cooper, and Vijay S. Pande. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.*, 51(4):769–779, 2008.

[100] Anthony Nicholls, David L. Mobley, J. Peter Guthrie, John D. Chodera, Christopher I. Bayly, Matthew D. Cooper, and Vijay S. Pande. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.*, 51(4):769–779, 2008.

[101] Frank Noé, Christof Schütte, Eric Vanden-Eijnden, Lothar Reich, and Thomas R. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, 106(45):19011–19016, 2009.

[102] Yasuhiko Nozaki and Charles Tanford. The Solubility of Amino Acids and Two Glycine Peptides in Aqueous Ethanol and Dioxane Solutions. ESTABLISH-

MENT OF A HYDROPHOBICITY SCALE. *J. Biol. Chem.*, 246(7):2211–2217, 1971.

[103] Asim Okur and Carlos Simmerling. Chapter 6 hybrid explicit/implicit solvation methods. volume 2 of *Annual Reports in Computational Chemistry*, pages 97 – 109. Elsevier, 2006.

[104] Lars Onsager. Electric moments of molecules in liquids. *J. Am. Chem. Soc.*, 58(8):1486–1493, 1936.

[105] A. Onufriev, D. Bashford, and D.A. Case. Modification of the generalized born model suitable for macromolecules. *J. Phys. Chem. B*, 104(15):3712–3720, 2000.

[106] Tatsuo Ooi, Motohisa Oobatake, George Nemethy, and Harold A. Scheraga. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA*, 84:3086–3090, 1987.

[107] Arno Papazyan and Arieh Warshel. Effect of solvent discreteness on solvation. *J. Phys. Chem. B*, 102(27):5348–5357, 1998.

[108] Robert A. Pierotti. The solubility of gases in liquids. *J. Phys. Chem.*, 67:1840–1845, 1963.

[109] J. W. Pitera and W. F. van Gunsteren. The importance of solute-solvent van der waals interactions with interior atoms of biopolymers. *J. Am. Chem. Soc.*, 123(13):3163–3164, 2001.

[110] Enrico O. Purisima and Traian Sulea. Restoring charge asymmetry in continuum electrostatics calculations of hydration free energies. *J. Phys. Chem. B*, 113(24):8206–8209, 2009.

[111] Anna Radzicka and Richard Wolfenden. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27(5):1664–1670, 05 1988.

[112] Sowmianarayanan Rajamani, Tuhin Ghosh, and Shekhar Garde. Size dependent ion hydration, its asymmetry, and convergence to macroscopic behavior. *J. Chem. Phys.*, 120(9):4457–4466, 2004.

[113] Robert C. Rizzo, Tiba Aynechi, David A. Case, and Irwin D. Kuntz. 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):128–139, 2006.

[114] Willian R. Rocha, Kaline Coutinho, Wagner B. de Almeida, and Sylvio Canuto. An efficient quantum mechanical/molecular mechanics monte carlo simulation of liquid water. *Chem. Phys. Lett.*, 335(1-2):127–133, February 2001.

[115] K. Sharp, A. Nicholls, R. Fine, and B. Honig. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science*, 252:106–109, 1991.

[116] Kim A. Sharp, Anthony Nicholls, Richard Friedman, and Barry Honig. Extracting hydrophobic free energies from experimental data: Relationship to protein folding and theoretical models. *Biochem.*, 30(40):9686–9697, 1991.

[117] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.*, 119(11):5740–5761, 2003.

[118] Michael R. Shirts and Vijay S. Pande. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.*, 122(13):134508, 2005.

[119] Devleena Shivakumar, Yuqing Deng, and Benoît Roux. Computations of absolute solvation free energies of small molecules using explicit and implicit solvent model. *J. Chem. Theory Comput.*, 5(4):919–930, 2009.

[120] Devleena Shivakumar, Joshua Williams, Yujie Wu, Wolfgang Damm, John Shelley, and Woody Sherman. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the opls force field. *Journal of Chemical Theory and Computation*, 6(5):1509–1519, 2010.

[121] Doree Sitkoff, Kim A. Sharp, and Barry Honig. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, 98(7):1978–1988, 1994.

[122] Eric J. Sorin and Vijay S. Pande. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical Journal*, 88(4):2472–2493, Apr 2005.

[123] N. T. Southall and K. A. Dill. The mechanism of hydrophobic solvation depends on solute radius. *J. Phys. Chem. B*, 104(6):1326–1331, 2000.

[124] Thomas Steinbrecher, David L. Mobley, and David A. Case. Nonlinear scaling schemes for lennard-jones interactions in free energy calculations. *J. Chem. Phys.*, 127(21):214108, 2007.

[125] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112(16):6127–6129, 1990.

[126] Frank H. Stillinger. Structure in aqueous solutions of nonpolar solutes from the standpoint of scaled-particle theory. *J. Solution Chem.*, 2(2/3):141–158, 1973.

[127] T. P. Straatsma, H. J. C. Berendsen, and J. P. M. Postma. Free energy of hydrophobic hydration. a molecular dynamics study of noble gases in water. *J. Chem. Phys.*, 85(11):6720–6727, 1986.

[128] C. Tan, Y.-H. Tan, and R. Luo. Implicit nonpolar solvent models. *J. Phys. Chem. B*, 111:12263–12274, 2007.

[129] Chunhu Tan, Lijiang Yang, and Ray Luo. How well does poisson-boltzmann implicit solvent agree with explicit solvent? a quantitative analysis. *J. Phys. Chem. B*, 110(37):18680–18687, 2006.

[130] C Tanford. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.*, 84(22):4240–4247, 1962.

[131] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum mechanical continuum solvation models. *Chem. Rev.*, 105(8):2999–3093, 2005.

[132] Yaoquan Tu and Aatto Laaksonen. The electronic properties of water molecules in water clusters and liquid water. *Chem. Phys. Lett.*, 329(3-4):283–288, October 2000.

[133] Marc van der Kamp and Valerie Daggett. Molecular dynamics as an approach to study prion protein misfolding and the effect of pathogenic mutations. In Jrg Tatzelt, editor, *Prion Proteins*, volume 305 of *Topics in Current Chemistry*, pages 169–197. Springer Berlin / Heidelberg, 2011.

[134] Alessandra Villa and Alan E. Mark. Calculation of the free energy of solvation for neutral analogs of amino acid side chains. *J. Comput. Chem.*, 23(5):548–553, 2002.

[135] Vincent A. Voelz, Gregory R. Bowman, Kyle Beauchamp, and Vijay S. Pande. Molecular simulation of ab initio protein folding for a millisecond folder ntl9(139). *Journal of the American Chemical Society*, 132(5):1526–1528, 2010. PMID: 20070076.

[136] J. A. Wagoner and N. A. Baker. Assessing implicit models for nonpolar mean solvation forces: The importance of dispersion and volume terms. *Proc. Natl. Acad. Sci. USA*, 103(22):8331–8336, 2006.

[137] A. Wallqvist and B. J. Berne. Molecular dynamics study of the dependence of water solvation free energy on solute curvature and surface area. *J. Phys. Chem.*, 99(9):2885–2892, 1995.

[138] J Wang, W Wang, P Kollman, and D Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25:247–260, October 2006.

[139] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[140] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Mod.*, 25:247–260, 2006.

[141] A. Warshel and G. King. Polarization constraints in molecular dynamics simulation of aqueous solutions: The surface constraint all atom solvent (SCAAS) model. *Chem. Phys. Lett.*, 121(1-2):124–129, 1985.

[142] Dongqing Wei and D. R. Salahub. A combined density functional and molecular dynamics simulation of a quantum water molecule in aqueous solution. *Chem. Phys. Lett.*, 224(3-4):291–296, July 1994.

[143] R. V. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate. Affinities of amino acid side chains for solvent water. *Biochem.*, 20(4):849–855, 1981.

148

[144] R. V. Wolfenden, P. M. Cullis, and C. C. B. Southgate. Water, protein folding, and the genetic code. *Science*, 206(4418):575–577, 1979.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Charles W. Kehoe
_____
Author Signature

6/13/12
_____
Date