Title: Efficient consideration of coordinated water molecules improves computational protein-protein and protein-ligand docking

Authors: Ryan E. Pavlovicz[a,b,c], Hahnbeom Park[a,b], Frank DiMaio[a,b]

Author Affiliation: [a]Department of Biochemistry, [b]Institute for Protein Design, University of Washington, Seattle, Washington 98195, United States, [c]current address: Cyrus Biotechnology, 500 Union St. Suite 320, Seattle, Washington 98101.

Corresponding Author: Frank DiMaio
J575, Health Sciences Building, Box 357370, Seattle, WA 98195-7370
(206) 221-8535
dimaio@u.washington.edu

Classification: Biological Sciences, Biophysics and Computational Biology

Short Title: Modeling coordinated waters in biomolecular docking

Keywords: energy function, force field, implicit water model, explicit water, binding free energy

Abstract:

Highly-coordinated water molecules are frequently an integral part of protein-protein and protein-ligand interfaces. We introduce an updated energy model that efficiently captures the energetic effects of these highly-coordinated water molecules on the surfaces of proteins. A two-stage protocol is developed in which polar groups arranged in geometries suitable for water placement are first identified, then a modified Monte Carlo simulation allows highly coordinated waters to emerge. This "semi-explicit" water model is implemented in Rosetta and allows for simultaneous prediction of side chain conformation and coordinated water geometry; the approach is suitable for structure prediction and protein design. We show that our new approach and energy model yield significant improvements in native structure recovery of protein-protein and protein-ligand docking.

Significance Statement:

Coordinated water molecules, those forming multiple hydrogen bonds with protein polar groups, play an important role in the structure of and interaction between biomolecules, yet the effect of these waters is often not considered in biomolecular computations. In this paper, we describe a method to efficiently consider these water molecules both implicitly and explicitly at the interfaces formed by two polar molecules. In computations related to determining how a protein interacts with binding partners, we show that the use of this new method significantly improves results. Future application of this approach may improve the design of new protein and small molecule drugs.

Introduction:

Water plays a significant role in biomolecular structure. The hydrophobic effect drives the collapse of proteins into their general shape while well-coordinated water molecules (water molecules making multiple water-protein hydrogen bonds) on the surface of a protein may confer specific conformations to nearby polar groups. Furthermore, water plays a key role in biomolecular recognition: when a ligand binds its host in an aqueous environment, it must displace water molecules on the surface and energetically compensate for the lost interactions. Coordinated water molecules may also drive host-ligand recognition by bridging interactions between polar groups on each side of the complex.

Simulations of proteins in explicit solvent have been successful in predicting folded conformations[1] as well as computing binding free energies[2] with high accuracy. This comes at significant computational cost, while the use of implicit solvent[3] greatly expedites such calculations, but at the loss of accuracy achieved through the inclusion of highly-coordinated water molecules[4]. Thus, an approach combining the efficiency of implicit solvation with the ability to recapitulate well-

coordinated water molecules is desired. Several such methods have been developed but tend to be target-specific[5-8] or relatively expensive computationally[9-10].

In this paper, we describe the development of general methods for capturing the energetic effects of explicit solvent, but with the computational efficiency of an implicit solvent model, making the approach suitable protein-protein and protein-ligand docking. The methods include: a) a new energy function capturing the energetics of protein and coordinated-water interactions and b) a conformational sampling approach that efficiently samples protein and water conformations simultaneously. We show this approach yields superior results in predicting coordinated water positions as well as improving the ability to predict native protein-protein and protein-ligand interfaces.

Results:

Our approach for modeling coordinated water molecules using Rosetta, fully described in *Methods*, is briefly presented here. We have developed two complimentary approaches for capturing coordinated-water energetics. First, *Rosetta-ICO (Implicit Consideration of cOordinated water)*, implicitly captures pairs of polar groups arranged such that a theoretical "bridging" water molecule may form favorable hydrogen bonds to stabilize the interaction. This calculation is efficient but ignores multi-body interactions, favoring, for example, waters coordinated by >2 hydrogen bond donors or acceptors. Therefore, we have also developed *Rosetta-ECO* (*Explicit Consideration of cOordinated water*), in which Rosetta's Monte Carlo (MC) simulation is augmented with moves to add or remove explicit solvent molecules from bulk. By sampling water orientations at sites where predicted bridging waters overlap (Figure 1), we properly coordinate water molecules to optimize hydrogen bonding.

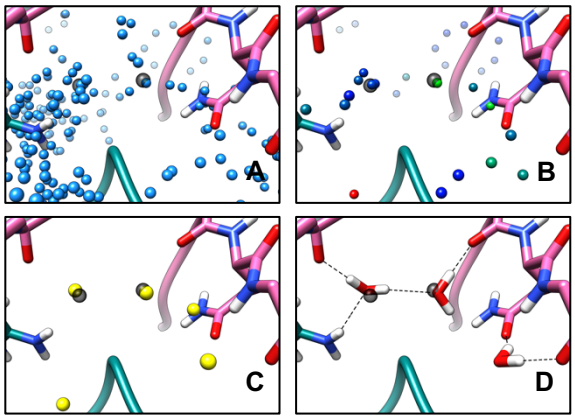

Figure 1. Explicit Protein Solvation with Rosetta. A. Initial possible solvation sites (blue) are built based on statistics of water positions about backbone polar atoms in addition to sites about side chain polar atoms based on all available rotameric positions. This example is using the interface of PDB ID: 1P57, between the N-terminal (pink) and catalytic (teal) domains of hepsin, with crystallographic waters in transparent grey. **B.** After an initial stage of Monte Carlo (MC)

packing with both the possible water sites and surrounding protein side chains, a cutoff is applied based on the dwell time of water sites (color from blue (dwell time = 0%) to green (dwell time = 25%) to red (dwell time = 50%). **C.** Remaining water sites are clustered and a second cumulative dwell time cutoff is applied. **D.** The remaining predicted water sites are converted into full-atom water molecules and re-packed with the surrounding side chains using the full Rosetta score function. Two of the final predicted water molecules in this figure are within 0.50 and 0.18 Å of crystallographic water positions, while another water molecule is well-coordinated by the protein, but does not observed in the crystal structure.

For both approaches, the energy function has been reoptimized using the *dualOptE* framework described by Park et al.[11]. In this optimization, several meta-parameters describing the shape of the *Rosetta-ICO* potential; several terms controlling the strength and shape of protein-water interactions; and ~50 other per-atom polar parameters were optimized to allow for compensating changes to the new energy terms. Energy function parameters for polar groups, including partial atomic charges, were refit using the same training tasks originally used in the parameterization of the *opt-nov15* energy function[11], now called REF2015[12]. The results in this section are shown with these updated energy functions compared to baseline tests run using the REF2015 energy function[11].

Rotamer and Water Recovery at Protein-Protein Interfaces

A set of 153 native protein-protein interfaces from high-resolution X-ray crystal structures was used to test how well the new energy models perform at simultaneously predicting amino acid side chain conformations and coordinated water sites. These tests involved the re-sampling of side chain conformations of interface residues on a fixed backbone in MC simulations, and evaluating resulting predicted sidechains against the deposited density maps. In tests involving semi-explicit water molecules (*Rosetta-ECO*), protein and water simultaneously sample conformational space. A baseline rotamer recovery error of 9.73 ± 0.13% (over three runs) was obtained using the *REF2015* energy function for the 7040 flexible side chains of the test set. A marginal improvement is made with *Rosetta-ICO,* reducing error to 9.52 ± 0.04%. Inclusion of explicit water molecules in this test fails to further decrease the overall rotamer recovery error beyond the improvements observed with *Rosetta-ICO*, with a *Rosetta-ECO* error of 9.59 ± 0.15%, while predicting ~19 explicit water molecules per protein-protein interface.

In addition to measuring sidechain rotamer recovery at the protein-protein interfaces, we also analyzed the recovery of water positions found in the high-resolution X-ray crystal structures when implementing the *Rosetta-ECO* solvation method. For water recovery tests, modeled water positions are considered "correct" if they are placed within 0.5 Å of the native water or if they are coordinated by the same polar atoms. *Rosetta-ECO* is able to recover 17.1% of native water molecules with a precision of 17.7%. Details of *Rosetta-ECO water* recovery are shown in Table 1. These tables show that our approach is most effective at predicting "buried" waters (28.3% recovery) and highly-coordinated waters (31.2% of triply-coordinated waters). Unsurprisingly, *Rosetta-ECO* is also much more

effective at predicted backbone-coordinated waters, correctly predicting 49.4% of backbone-only coordinated waters. An example of correctly predicted water sites is illustrated in Figure 1D.

Table 1. Classification of Predicted Native Waters

Type[1]	Subset Size	% recovered[2]	# precision[3]
All	3226	17.1%	17.7%
Exposed	941	6.0%	5.0%
Partially Buried	1786	19.2%	22.0%
Buried	408	28.3%	28.3%
1 protein coord	892	6.1%	6.6%
2 protein coord	1219	26.5%	26.8%
3 protein coord	458	31.2%	20.0%
BB only	356	49.4%	10.0%
SC only	384	7.0%	11.3%
BB+SC	1070	27.7%	26.1%

[1]Three groups of categorization of type of predicted water molecules. First, waters are classified 'buriedness' based on number of amino acid neighbors (nCβ) with Cβ within 10 Å. Exposed: nCβ <=15; partially buried: 15 < nCβ <= 25; buried: nCβ > 25. Second, classification by 1, 2, or 3 protein coordination partners within 3.2 Å. Finally, by type of coordinating protein atoms with 3.2 Å of the water O atom: at least two backbone only (BB only), side chain only (SC only) or a mix of backbone and side chain coordination (BB+SC).
[2-3]Percent and number of specific types of waters recovered using recovery criteria described in *Methods*, averaged over three runs.

Native Interface Recapitulation

We next tested the ability the new energy model to recapitulate near-native conformations of protein-protein interfaces (PPIs) and protein-ligand interfaces. In these tests, the binding free energies for a number of near-native and incorrect (decoy) docking conformations of each complex are computed with the aim of discriminating the correct binding poses from the decoys. PPI decoys were sampled using a combination of Zdock[13] and RosettaDock[14], while protein-ligand decoys were generated using RosettaLigand[15]. Both datasets were enriched for water-rich interfaces, leading to 53 protein-protein interfaces and 46 protein-ligand interfaces. Then predicted binding free energies, ΔG_{bind} are calculated for all decoys (see *Methods*). We assess the ability to predict the near-native conformations using a "discrimination score,"[11], which computes the Boltzmann weight of near-native structures. The values range from 0 to 1, with higher values showing better discrimination. An overview of the results is shown in Table 2.

Table 2. Performance of Different Solvation Schemes on Protein-Protein and Protein-Small Molecule Docking Discrimination

	REF2105	*Rosetta-ICO[1]*	*Rosetta-ECO[2]*
Protein-small molecule			
discrimination score[3]	0.7412 ± 0.0027	0.7977 ± 0.0021	0.8585 ± 0.0027

percent correct[4]	75.4 ± 2.1	76.1 ± 1.8	92.0 ± 1.0
run time[5]	1.00	1.09	1.52
Protein-protein			
discrimination score	0.6277 ± 0.0138	0.7386 ± 0.0061	0.7860 ± 0.0088
percent correct	63.6 ± 0.9	74.9 ± 0.9	78.6 ± 1.7
normalized run time	1.00	1.25	2.59

[1]Implicit consideration of coordinated water molecules
[2]Inclusion of well-ordered explicit water molecules
[3]Reported are the average Boltzmann-weighted discrimination scores ± 1σ averaged over three independent runs for 46 protein-ligand and 53 protein-protein docking cases.
[4]The percentage of cases in which the lowest scoring model is within 1.0 Å of the native conformation for protein-ligand docking and 2.0 Å for protein-protein docking, averaged over 3 independent runs
[5]Run time, normalized to baseline, is the sum of individual run times to calculate ΔG_{bind} for each near-native and decoy conformation

Protein-Protein Docking

In protein-protein docking tests, significant improvements are observed when comparing *Rosetta-ICO* to the baseline results, with the discrimination score increasing from 0.63 to 0.74. *Rosetta-ECO* further improves this discrimination score to 0.79. We also consider the "success rate," the time the lowest-energy conformation is within 2.0 Å of native: the *ECO* model enables successful prediction of a near-native conformation in 8 additional cases out of the set of 53, a ~15% improvement. This comes at a modest increase in computational cost, with an average 1.25- and 2.59-fold increase in runtime for *ICO* and *ECO*, respectively.

As illustrated in Figure 2A, *Rosetta-ECO* improves the discrimination score for 38 of 53 cases, adding 13.4 water molecules to the average bound state and 15.0 water molecules to the average unbound state. These average improvements remain statistically significant. Looking at one such case (adrenodoxin reductase/adrenodoxin, PDB ID 1E6E), we see that while all three energy models correctly predict a near-native conformation, the "energy gap" between native and non-native conformations is improved under *Rosetta-ECO* (Fig. 2B). Closer investigation of the near-native models shows 21 explicit water molecules added to the binding interface. The combined electrostatic and hydrogen bond energy contributions compose a large proportion of the improved binding energy, 5.2 kcal/mol more favorable than *Rosetta-ICO* for this particular binding configuration.

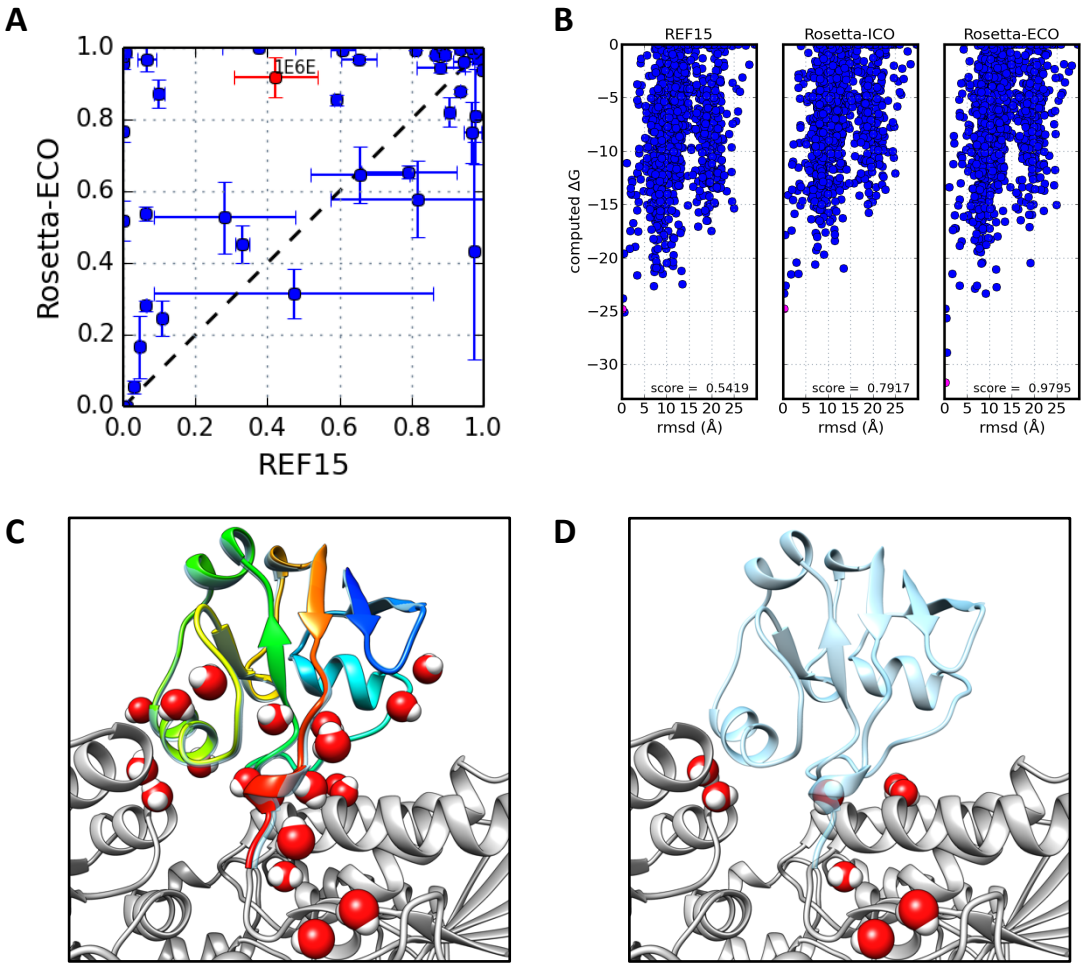

Figure 2. Protein-Protein Docking Results. A. Scatter plot comparing results of 53 cases between *REF2015* and *Rosetta-ECO*. Values are the Boltzmann-weighted score ± 1σ from an average of three independent runs. **B.** Energy funnels for PDB ID: 1E6E, adrenodoxin reductase bound to adrenodoxin (red data point in 2A), plotting computed ΔG_{bind} vs. rmsd from the native binding conformation for three different scoring methods. Boltzmann-weighted scores for each distribution are noted in bottom right of each plot. **C.** Explicitly-solvated near-native docking pose (rmsd=0.14 Å; pink data point in 2B) with the reductase in grey and adrenodoxin in rainbow (N- to C-terminus colored blue to red) overlaying the native structure (transparent blue). **D.** The predicted unbound state in which a number of interface waters return to bulk after recalculation.

Protein-Ligand Docking

For protein-ligand docking tests, *Rosetta-ICO* also provides improvement over *REF2015*, with average discrimination score increasing from 0.74 to 0.80. *Rosetta-ECO* further increases the discrimination score to 0.86. I n terms of "success rate", we see the same trend as with PPIs: *Rosetta-ECO* enables the correct prediction (within 1.0Å of native) in 7 additional cases out of the 46. These results indicate that both *Rosetta-ICO* and *ECO* help discriminate distant decoys from native conformations when compared to the *REF2015* energy model, with the inclusion

of explicit water modeling in *ECO* conferring the largest benefit. This also comes at only a modest increase in run time: about 10% increased time for ICO, and about 52% increased computation time for *ECO*.

The improvements in discrimination score on a case-by-case basis are illustrated in Figure 3A. Here, we see that *Rosetta-ECO* provides a near across-the-board improvement in native discrimination compared to the baseline calculations. The individual energy distributions for PDB ID 1X8X (tyrosyl t-RNA synthase / tyrosine) in Figure 3B show how both *REF2015* and *Rosetta-ICO* incorrectly favor a decoy 6.6 Å from native. *Rosetta-ECO*'s explicit waters dramatically alter the binding energy landscape, improving the discrimination score from 0.27 to 0.89, and energetically favoring a structure only 0.43 Å from native. The *ECO* model predicts two water molecules that bridge the carboxyl group of the tyrosine ligand to interactions with and arginine side chain and a backbone nitrogen group (Fig. 3C): these provide favorable interactions to the native state, with electrostatic and hydrogen bonding interactions a combined 7.5 kcal/mol more favorable when including the explicit interface waters in the *ECO* calculations.

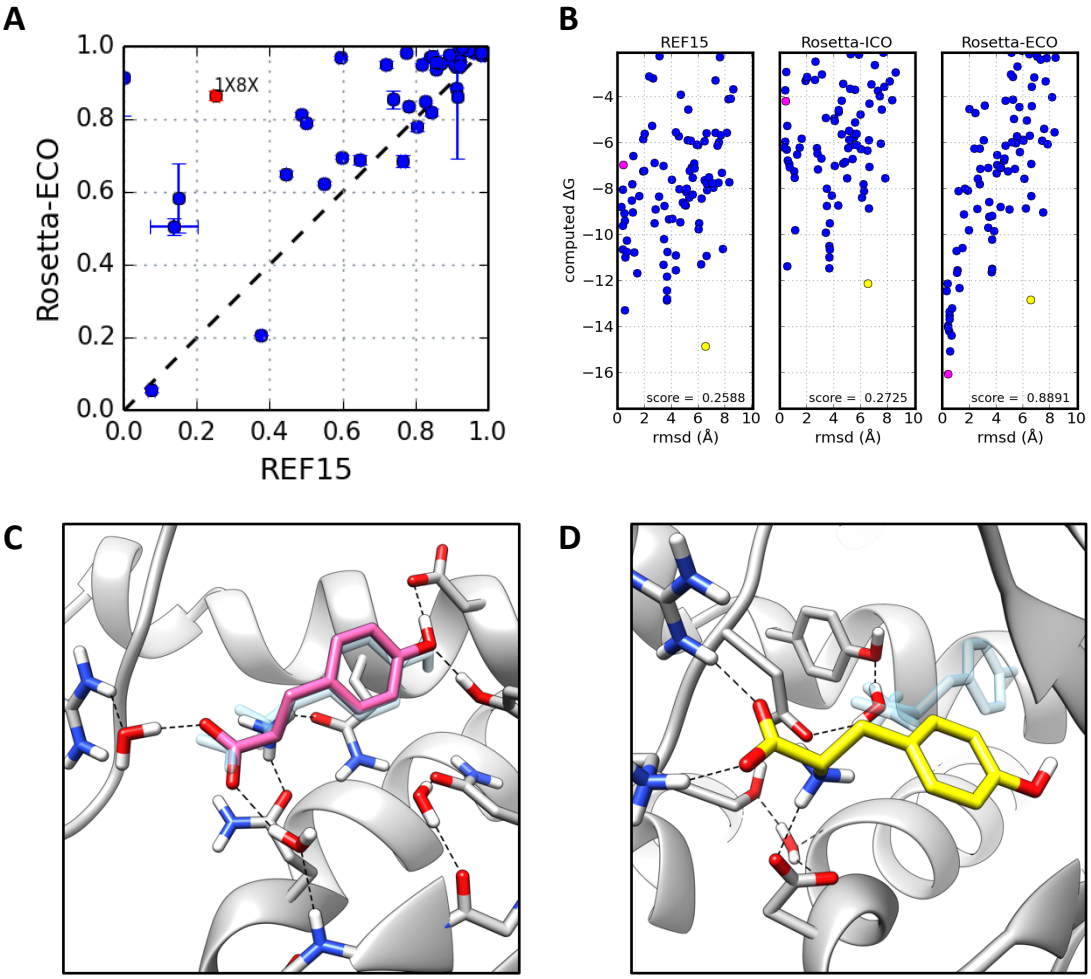

Figure 3. Protein-Ligand Docking Results. A. Scatter plot comparing results of 46 cases between baseline (*REF2015*) and *Rosetta-ECO*. Values are the Boltzmann-weighted score ± 1σ from an average of three independent runs. **B.** Energy funnels, similar to Figure 2, for PDB ID: 1X8X, tyrosyl t-RNA synthase bound to tyrosine (red data point in 3A) **C.** Explicitly-solvated, near-native docking pose in pink (RMSD=0.43 Å; pink data point in 3B) with native ligand in transparent blue. **D.** Explicitly-solvated decoy binding pose (RMSD=6.57 Å; yellow data point in 3B).

Ligand Docking Comparison

Finally, the new energy functions were compared against the results of a state-of-the art docking approach on a standardized dataset. A recent survey[16] of widely-used small molecules docking programs tested for performance against the Astex Diverse Set[17] which includes 85 targets with ligands of pharmaceutical interest. We generated decoys for a 67-target subset, excluding cases in which the ligand is coordinated by an ion, using the top-performing docking software, GOLD[18]. The GOLD-sampled structures were then rescored using the *REF2015*, *ICO*, and *ECO* energy functions of Rosetta. The results, fully presented in Figure S1 and Table S1, show that while the Rosetta-rescored structures are more accurate than GOLD (78.2% versus 67.7% accuracy within a 1 Å RMSD cutoff; 94.6% versus 80.7% accuracy within 2 Å RMSD cutoff), we see little improvement between *REF2015* and *ICO*/*ECO*. While this suggests Rosetta may be a powerful tool for this dataset, the restricted conformational sampling obtained from GOLD (see Figure S2) does not benefit from the water model developments presented here, and prevents a thorough evaluation of the energy functions.

Discussion:

We have presented two approaches for considering coordinated water molecules in the prediction of native protein-protein and protein-ligand interfaces: *Rosetta-ICO*, which very efficiently captures the energetics of bridging waters implicitly, and *Rosetta-ECO*, which allows a small set of waters to emerge from bulk. Both show improvements in protein interface recapitulation tasks, and both represent different levels of efficiency-accuracy tradeoffs, with *Rosetta-ECO* more accurate but 1.5 to 2 times slower than *Rosetta-ICO* depending on interface size.

Furthermore, while this manuscript highlights the results on water prediction and protein interface recapitulation, we might expect the *Rosetta-ICO* energy function to show modest improvements at tasks related to monomeric structure prediction and protein sequence design. Indeed, that seems to be the case: when tested on independent datasets, modest improvements were observed in decoy discrimination with *ICO*. All other metrics were comparable between the two energy functions, leading us to conclude that the ICO model is a reasonable general-purpose energy function.

The improvement in both the protein and ligand docking tests suggests that these new energy functions may prove useful in the design of novel proteins intended to bind a particular ligand or protein. Successful design of protein-protein interfaces is often driven by van der Waals interactions that arise from shape complementarity, however better consideration of ordered solvent molecules may allow for the design of more natural interfaces which include numerous polar residues. Application of these new methods need not be limited to the solvation of interfaces or the description of binding partners. For example, the methods may be applied to more accurately predict the folded state of monomeric proteins in which buried solvent plays an important structural role or for prediction of the stabilizing or destabilizing effect of mutated residues on the surface of a protein. Additionally, the experiments described herein only consider the solvation of proteins and small molecules, however the framework can easily be extended to solvate other biomolecules, such as nucleic acids.

Methods:

Two new biomolecular solvation methods are introduced here. The first builds upon the existing implicit water model used in Rosetta to not only account for desolvation penalties, but energetically reward conformations that are suited to accommodate theoretic bridging waters which are calculated on the fly. The second model places well-coordinated water molecules on the surface or at interfaces of biomolecules based largely on statistics from high-resolution experimental data.

1.) Implicit Solvation (*Rosetta ICO*)

An additional energy term is added to the Rosetta's implicit solvation model that models the energetic costs of highly ordered water molecules coordinated by multiple protein polar groups. The term builds upon our previously developed anisotropic solvation model[11], where for each polar group, one or more virtual water sites are placed in a configuration ideal for hydrogen bonding with the corresponding polar group. An energetic bonus is then given when the water sites of multiple polar groups overlap in such a way that a single water could coordinate, or "bridge", these polar groups:

$$E_{lk-bridge}(r_i, r_j) = \left(E_{lk}^{(i,j)}\right) \cdot G\left(\min_{w_i,w_j}\|w_i - w_j\|\right) + \left(E_{lk}^{(i,j)}\right) \cdot G\left(\|b_i - b_j\| - D_0\right)$$

With:

$$G(x) = \left(1 - \left(\frac{x^2}{S_0}\right)^2\right)^2$$

Here, w_i is the xyz coordinate of a theoretic water corresponding to polar group r_i; b_i is the xyz coordinate of the base heavy atom used to construct the water (e.g.,

the backbone N or O), and D_0 and S_0 are parameters that are optimized during energy function evaluation. The two terms in the equation characterize the overlap and the angle formed between polar groups that potentially coordinate a water.

This energy term was added to the current anisotropic solvation model in Rosetta, and optimization of all polar terms was carried out. While this term does not prevent disallowed coordination geometries (e.g., 3 donors or 3 acceptors coordinating a single water site), in practice, the water sites implicitly identified by this approach are quite reasonable. Because this two-body energy term is only dependent upon the configuration of pairs of protein polar groups, it can be used in all Monte Carlo minimization methods used in Rosetta[19], with negligible computational overhead.

Additionally, to properly handle the geometry of water-protein and water-water hydrogen bonds, we modified the functional form of sp_3-hybridized hydrogen bond acceptors. Previously, the interaction between a hydrogen bond donor and the lone pair electrons of sp_3-hybridized acceptors was described by an angle and torsional term about the base atoms; e.g., for serine, the angle CB-OG· · ·Hdon and the pseudo-torsion HG-CB-OG· · ·Hdon. For water, however, this led to an undesirable property in that the potential treated water asymmetrically. Therefore, the torsional term water replaced with a "softmax" potential between the both atoms bonded to the sp_3-hybridized acceptor:

$$E_{sp3-chi}(a_i, h_j) = -M \cdot \log\left(\sum_{b_k \text{ bound to } a_i} \exp\left(E_{BAH}(b_k, a_i, h_j)/M\right)\right)$$

Above, a_i and h_j are the acceptor heavy-atom and donor hydrogen, respectively; E_{BAH} is the angular potential about the heavy-atom[20]. The summation is carried out over all bound atoms to the acceptor: for water acceptors, this would be over both hydrogens. In the serine example above, the angular potential is applied to both CB-OG- - - Hdon and HG-OG- - - Hdon and the softmax gives a score equal to the worse of the two angular potentials. This ensures the potential is symmetric about both water hydrogens.

2.) Explicit Solvation Model (*Rosetta ECO*)

One key challenge in prior explicit water modelling[21] is the large conformational space a single water molecule can adopt. This is particularly problematic in applications (like those in this manuscript) where it is desirable to simultaneously sample sidechain conformations and water positions. *Rosetta-ECO* makes use of a two-stage approach to get around this problem (Figure 1). In the first stage, rotationally independent "point waters" are sampled using a statistical potential; not considering water rotation lets thousands of putative water positions be sampled efficiently. In the second stage, for the most favorable water positions (typically only several dozen) we consider rotations of these molecules using a physically derived potential.

In both steps of the protocol, Monte Carlo sampling is used to simultaneously sample sidechain rotamer and water "rotamer." In both stages, water molecules may be set to "bulk," gaining an entropy bonus by doing so, or may be set to explicit, gaining favorable interactions but losing the entropy bonus. Rotational sampling of waters using a uniform SO_3 gridding strategy[22] with 30° angular spacing.

2.1) Derivation of the Statistical Point Water Potential

The first step in determining possible water sites involves a low-resolution, statistical water potential to quickly evaluate the interaction between possible water sites and nearby polar groups of biomolecules. This potential, which we are calling the "point water potential", treats water molecules as simple, uncharged, points with attractive and repulsive Lennard-Jones terms.

The point water potential takes the form of:

$$E_{point-water}(W = \{w_i\}) = \sum_{\text{waters } i} \sum_{\substack{\text{polar} \\ \text{atoms } j}} -\log\ P\big(d(w_i, x_j), \theta(w_i, x_j, x_j^{base})\big)$$
$$-K \cdot \sum_{\substack{\text{waters } k \\ i \neq k}} \exp[-(d(w_i, w_k) - 2.7)^2/\sigma^2] + E_{ref}$$

Here, *P* is the statistical point-water distribution, parameterized over distance and angle; *d* gives the distance between a water and polar atom, and θ gives the angle between water, polar atom, and its "base atom." The point water energy term also considers other nearby point water sites, k, as Gaussian distributions with width σ and height *K* (with min energy at a distance of 2.7 Å), which was determined by averaging water-water distances observed in high resolution crystal structures. Finally, an overall energetic cost of bringing the water molecule "out of bulk," E_{ref}, is added for each water. These parameters were fit using crystallographic waters in the Top8000 database (see Supplemental for more details).

2.2) Identifying and Packing Point Waters

A key challenging in building possible water sites is we want to simultaneously sample sidechain formations along with water positions. Thus, the initial placement of water molecules to be optimized by the point water potential come from two sources: a) ideal solvation about protein backbones, and b) *possible* solvation sites from sidechain rotamers. For backbone waters, point generation is straightforward: 10 "ideal" sites are generated from each backbone C=O group (based on clustering waters from crystal structures).

Generation of sidechain-coordinated waters is not so straightforward. Considering all possible polar groups of all sidechain rotamers is computationally intractable. We again build off prior work[23] and consider instead sidechain/sidechain (and sidechain/backbone) "overlaps." That is, we generate all possible sidechain rotamers for every sidechain, and identify all positions where there is overlap (within 0.75 Å) between two different sidechains. A 3D hash table makes this calculation efficient even when there are millions of putative water positions. Finally, to further reduce conformational sampling, during the Monte Carlo "packing" algorithm, when both sidechain and point water positions are sampled, all putative point waters are clustered into sets into which only one site can be occupied.

A modified version of Rosetta's traditional packing algorithm[24] is used when point waters are present. Typically, Rosetta uses simulated annealing to find the discrete rotamer set minimizing system energy, where the temperature of the trajectory is slowly annealed from RT=100 to RT=0.3. With the point water potential, we do not expect the forcefield (which does not consider water rotation) to be perfect, and we want the packer not to optimize total energy but to simply separate reasonable from unreasonable water positions for a more expensive subsequent calculation. Thus we instead used long simulations at low temperatures (RT=0.3) with intervening high-temperature "spikes" (RT=100). Then, instead of taking the lowest energy state sampled, we measured water "occupancy" at each position, taking point water positions with a "dwell time" more than 2%.

The water positions passing this criterion, typically only several dozen to a hundred, are then allowed to rotate and are packed (along with all sidechains) using Rosetta's standard simulated annealing rotamer optimization routine.

3.) Datasets

Four different data sets were used in the testing of the new energy functions described here. The first includes 153 high-resolution crystal structures of protein-protein interfaces (PPIs) that was used for both native water and rotamer recovery at the interfaces. Two docking data sets were used to test the ability of the new energy functions to discriminate near-native from decoy docking conformations, a subsets of those used by Park et al.[11], but selected for water-rich interfaces (and to exclude problematic cases such as PPIs with disulfides across the interface or ions contributing to binding). For protein-protein interactions, a 53-case subset of the ZDock 4 Benchmark set[25] was used, while a 46-case subset of the Binding MOAD database[26] was used for protein-ligand interactions. Finally, another ligand docking set, generated with GOLD on a subset of the Astex Diverse Set[17] was used to compare the new energy functions against an established docking score function. Details on the datasets, including lists of PDB IDs used are included in the Supplemental Materials.

4.) Binding Energy Calculations

The binding free energies, ΔG_{bind}, were calculated for the near-native and incorrect (decoy) docking poses by taking the difference between the computed energies of the bound and unbound states. This is accomplished in Rosetta by first calculating the energy for the bound system, then re-computing the energy when the two binding components are separated to obtain unbound state energies. An important part of interface energetics involves computing the energy cost of water displacement[27], and so treatment of explicit waters of the unbound state was an important consideration. Due to size differences of the average interface, we found slightly different treatment performed better with PPIs versus protein-ligand interfaces. In both PPIs and protein-ligand interfaces, the bound states are solvated, using the two-stage Monte Carlo procedure described above. Then all sidechains are minimized, and – for protein-ligand interfaces only – the rigid-body transformation between subunits is also minimized. Then subunits are separated and re-solvated. The waters from the bound state are saved (and duplicated) following separation and are always considered in the rotatable water calculation.

5.) Training Tasks

The training tasks used for energy function parameterization are the same as detailed in the development of the REF2015 Rosetta energy function[11] and are summarized in the Supplemental Materials.

Availability

This code is fully integrated into the Rosette software suite, with example XML files available in the Supplemental Materials.

Acknowledgements:

This work was facilitated though the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington. Structure visualization and analysis used the UCSF Chimera software[28], while GNU Parallel was used for distributed processing and data analysis[33].

References:

1. Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E., How fast-folding proteins fold. *Science* **2011,** *334* (6055), 517-20.
2. Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A., Predicting absolute ligand binding free energies to a simple model site. *Journal of molecular biology* **2007,** *371* (4), 1118-34.
3. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J Am Chem Soc* **1990,** *112* (16), 6127-6129.

4. Beauchamp, K. A.; Lin, Y. S.; Das, R.; Pande, V. S., Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J Chem Theory Comput* **2012,** *8* (4), 1409-1414.
5. Lemmon, G.; Meiler, J., Towards ligand docking including explicit interface water molecules. *PloS one* **2013,** *8* (6), e67536.
6. Parikh, H. I.; Kellogg, G. E., Intuitive, but not simple: including explicit water molecules in protein-protein docking simulations improves model quality. *Proteins* **2014,** *82* (6), 916-32.
7. Huggins, D. J.; Tidor, B., Systematic placement of structural water molecules for improved scoring of protein-ligand interactions. *Protein engineering, design & selection : PEDS* **2011,** *24* (10), 777-89.
8. van Dijk, A. D.; Bonvin, A. M., Solvated docking: introducing water into the modelling of biomolecular complexes. *Bioinformatics* **2006,** *22* (19), 2340-7.
9. Huang, X.; Yang, J.; Zhu, Y., A solvated ligand rotamer approach and its application in computational protein design. *J Mol Model* **2013,** *19* (3), 1355-67.
10. Jiang, L.; Kuhlman, B.; Kortemme, T. A.; Baker, D., A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* **2005,** *58* (4), 893-904.
11. Park, H.; Bradley, P.; Greisen, P.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F., Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J Chem Theory Comput* **2016,** *12* (12), 6201-6212.
12. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L., Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J., The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **2017,** *13* (6), 3031-3048.
13. Chen, R.; Li, L.; Weng, Z., ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **2003,** *52* (1), 80-7.
14. Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D., Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology* **2003,** *331* (1), 281-99.
15. Meiler, J.; Baker, D., ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* **2006,** *65* (3), 538-48.
16. Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T., Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Phys Chem Chem Phys* **2016,** *18* (18), 12964-75.
17. Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W., Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of medicinal chemistry* **2007,** *50* (4), 726-741.

18. Korb, O.; Stutzle, T.; Exner, T. E., Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model* **2009,** *49* (1), 84-96.
19. Tyka, M. D.; Keedy, D. A.; Andre, I.; Dimaio, F.; Song, Y.; Richardson, D. C.; Richardson, J. S.; Baker, D., Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of molecular biology* **2011,** *405* (2), 607-18.
20. O'Meara, M. J.; Leaver-Fay, A.; Tyka, M. D.; Stein, A.; Houlihan, K.; DiMaio, F.; Bradley, P.; Kortemme, T.; Baker, D.; Snoeyink, J.; Kuhlman, B., Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J Chem Theory Comput* **2015,** *11* (2), 609-22.
21. Li, S.; Bradley, P., Probing the role of interfacial waters in protein-DNA recognition using a hybrid implicit/explicit solvation model. *Proteins* **2013,** *81* (8), 1318-1329.
22. Mitchell, J. C., Sampling Rotation Groups by Successive Orthogonal Images. *Siam J Sci Comput* **2008,** *30* (1), 525-547.
23. Yanover, C.; Bradley, P., Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic acids research* **2011,** *39* (11), 4564-4576.
24. Leaver-Fay, A.; Tyka, M.; Lewis, S. M.; Lange, O. F.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, P. D.; Smith, C. A.; Sheffler, W.; Davis, I. W.; Cooper, S.; Treuille, A.; Mandell, D. J.; Richter, F.; Ban, Y. E.; Fleishman, S. J.; Corn, J. E.; Kim, D. E.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popovic, Z.; Havranek, J. J.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, J. J.; Kuhlman, B.; Baker, D.; Bradley, P., ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **2011,** *487*, 545-74.
25. Pierce, B. G.; Hourai, Y.; Weng, Z., Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one* **2011,** *6* (9), e24657.
26. Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A., Binding MOAD, a high-quality protein-ligand database. *Nucleic acids research* **2008,** *36* (Database issue), D674-8.
27. Li, Z.; Lazaridis, T., The effect of water displacement on binding thermodynamics: concanavalin A. *The journal of physical chemistry. B* **2005,** *109* (1), 662-70.
28. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-12.
29. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J Comput Chem* **2002,** *23* (16), 1623-1641.
30. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* **2006,** *25* (2), 247-60.
31. Liebeschuetz, J. W.; Cole, J. C.; Korb, O., Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *Journal of computer-aided molecular design* **2012,** *26* (6), 737-48.

33. Tange, O., GNU Parallel – The Command-Line Power Tool, ;login: The USENIX Magazine, February **2011**, 42-47.