JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

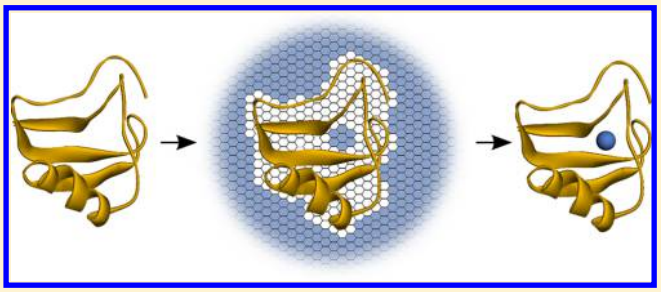
Prediction of Water Binding to Protein Hydration Sites with a Discrete, Semiexplicit Solvent Model

Piotr Setny*

Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland

Supporting Information

ABSTRACT: Buried water molecules are ubiquitous in protein structures and are found at the interface of most protein−ligand complexes. Determining their distribution and thermodynamic effect is a challenging yet important task, of great of practical value for the modeling of biomolecular structures and their interactions. In this study, we present a novel method aimed at the prediction of buried water molecules in protein structures and estimation of their binding free energies. It is based on a semiexplicit, discrete solvation model, which we previously introduced in the context of small molecule hydration. The method is applicable to all macromolecular structures described by a standard all-atom force field, and predicts complete solvent distribution within a single run with modest computational cost. We demonstrate that it indicates positions of buried hydration sites, including those filled by more than one water molecule, and accurately differentiates them from sterically accessible to water but void regions. The obtained estimates of water binding free energies are in fair agreement with reference results determined with the double decoupling method.



1. INTRODUCTION

Biomolecules typically exist in aqueous environments and interaction with water is of key importance for their structure and function.[1,2] Aside from acting as embedding medium responsible for bulk effects such as electrostatic screening or hydrophobic interactions, water is also present inside macromolecular structures. Here, its buried, isolated particles are usually involved in a network of specific hydrogen-bonding interactions, playing a variety of important roles. They contribute to the stability and dynamics of macromolecules,[3−7] facilitate receptor−ligand recognition,[8] or the formation of macromolecular assemblies.[9,10] The thermodynamic effect of tying up or releasing such confined water is often of comparable magnitude to the overall folding or binding free energies.[11]

Accordingly, the presence of buried water molecules needs to be taken into account in theoretical description and modeling of biomolecular systems. It has been demonstrated that accurate placement of isolated water molecules is vital for obtaining quantitative agreement with experimental data in free energy perturbation calculations,[12,13] as well as for improving the accuracy of receptor−ligand docking results.[14−16] In the area of drug design, the knowledge of binding site hydration is valuable for the optimization of lead compounds, as one of the strategies relies on the addition of polar groups displacing bound water.[17,18] Moreover, the detection of potentially hydrated cavities within a protein and filling them with water is an important prerequisite of explicit solvent MD simulations. Natural solvent permeation to such sites may be too slow for typical simulation times, while in some cases a single missing water molecule may be enough to destabilize protein−ligand complex[12] or disrupt protein structure.[19]

Valuable information about the location of bound water molecules comes from experimental methods, primary X-ray crystallography,[20] but also neutron diffraction or nuclear magnetic resonance.[21] In many cases, however, such information is uncertain (for example, due to difficulties in resolving water molecules by crystallography[20]) or missing (for example, in a docking problem, where each ligand requires its own hydration pattern[22]); hence, computational approaches allowing efficient prediction of putative hydration sites are of great value.

To date, numerous methods have been developed to deal with the problem of localized water molecules. Some of them focus on crystallographically resolved hydration sites and evaluate a set of descriptors in order to distinguish important, potentially conserved water molecules, from those easily displaceable.[23−25] Other approaches aim at *ab initio* prediction of hydration sites and rely on scanning the system of interest with a solvent probe in order to determine favorably hydrated locations based on force-field energy calculations,[26−28] which can be augmented by continuum electrostatics,[29] or using empirical scoring functions,[30−32] or docking techniques.[33] Separate methods have been developed to specifically tackle the problem of localized water molecules in protein−ligand docking.[34−39]

While those methods are fast and efficient, they neglect possible interactions between neighboring confined water

Received: September 2, 2015
Published: October 28, 2015

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

molecules (roughly 40% of bound water molecules are grouped in clusters containing 2 or more particles stabilized by mutual hydrogen bonds[40]) and do not take into account the entropic cost of confinement. A notable exception here is the SZMAP method,[29] for which predictions are based on the rotational partition function of a water probe and hence include orientation dependent entropic effects. Moreover, approaches parametrized in a knowledge-based manner, typically in the context of proteins, often lack transferability to arbitrary systems of interest. On a broader view, they do not provide a holistic description of solvation; a separate model is needed to account for bulk effects.

More accurate, yet computationally expensive, modeling of buried water requires simulations with explicit solvent, augmented by algorithms allowing efficient water permeation to internal cavities. A general, rigorous framework for obtaining insight into solvent distribution and energetics is based on simulations in a grand canonical ensemble, such as Grand Canonical Monte Carlo (GCMC),[41,42] but more specialized methods utilizing a combination of Monte Carlo sampling and free energy perturbation for switchable water particles have also been proposed.[43] In turn, precise calculations of binding free energies to already known hydration sites can be performed with the use of the double decoupling method (DDM).[44]

Computational effort and technical difficulties of the above methods can be partially avoided by assuming that equilibrium solvent distribution can be reached within unperturbed molecular dynamics (MD) simulations, either in the case of easily accessible hydration sites, or following initial placement of solvent in reasonable configuration. The analysis of water structure obtained during such simulations (density distribution and sometimes higher order correlations) in the framework of inhomogeneous solvation theory[45,46] gives insight into locations of hydration sites and their thermodynamic parameters.[47−51] An alternative route is offered by cell theory, in which water molecules are treated as moving in individual wells of effective harmonic potentials, whose force constants are determined by the analysis of average forces and torques, allowing subsequent analytical evaluation of partition function and resulting thermodynamics.[52] Relative ease of use combined with sound theoretical basis of such approaches spurred the advent of several modified methods commonly founded around the idea of postprocessing information gathered from explicit solvent MD.[53−57]

A notable, separate branch in the modeling of solvation is based on integral equation theory of solutions, generalized to the 3D reference interaction site model (3D-RISM)[58−60] that is applicable to large biomolecules. The use of the 3D-RISM approach to study protein hydration, including the localization of bound water molecules, has been demonstrated,[61−63] along with examples of hydration free energy calculations.[64]

The apparent wide assortment of diverse methods described above, on the one hand, confirms the importance of water but, on the other hand, reflects significant difficulties in capturing the complex nature of its interactions with biomolecules. In particular, combining the description of bulk hydration effects with proper treatment of buried hydration sites within a single, practically usable, and quantitatively correct model is still a largely unresolved issue. Implicit solvent models, increasingly successful in reproducing experimental hydration free energies,[65] can neither predict nor account for the presence of confined solvent regions. In turn, empirical methods dealing with buried water molecules have typically little to offer with respect to bulk hydration.

In our previous work, we introduced a solvation model based on discrete solvent representation and mean field approach.[66,67] We demonstrated that the model reproduces well experimental hydration free energies for a diverse set of neutral and charged organic solutes. Its important feature is that, unlike in most implicit solvent approaches, spatial solvent distribution in the presence of a solute does not follow from any kind of geometrically defined solvent accessible surface but rather is the result of calculations. As such, it is sensitive to solute topography and local physicochemical properties, reflecting for instance the appearance of isolated hydration sites within proteins or the existence of sterically accessible to water but void regions.

In this work, we make use of this feature of our model in order to obtain predictions regarding the positions of buried water molecules and their affinities to respective hydration sites. Importantly, aside from introducing a single additional adjustable parameter, related to bulk excess chemical potential of water (necessary to obtain protein−water binding free energies in relation to bulk water), we did not interfere with the original model parametrization; hence, its capabilities with respect to providing hydration free energy estimates remain unchanged. In the following, we briefly review the key model assumptions, describe an algorithm introduced to partition confined solvent regions into hydration sites, and present the obtained results in comparison to extensive analysis of water binding to protein cavities based on free energy calculations with the use of DDM.

2. METHODS

2.1. Discrete Solvent Model. *2.1.1. Basic Model Assumptions.* The model is based on atomistic treatment of solute−water interactions, combined with mean field description of intrasolvent interactions.[66,67] Solvent molecules, described with the three-site SPC/E water model,[68] sample space discretized into a body centered cubic (BCC) lattice in such a way that with an oxygen atom occupying a given lattice point, two hydrogen atoms are located on two of its eight nearest neighbors lying along the diagonal of the unit cell face (Figure 1). The SPC/E water model has a 109° HOH angle and thus fits perfectly into the tetrahedral arrangement of nearest neighbors in the BCC lattice once their distance is adjusted to match the OH distance of 1 Å. An estimate of local water excess chemical potential at point **r** is evaluated as a position dependent effective Hamiltonian, representing an ensemble average over 12 possible orientations, θ, of the SPC/E water probe allowed by unique placements of hydrogens around the oxygen atom located at **r**:

$$H_{eff}(\mathbf{r}, \{n\}) = -k_B T \ln\left(\sum_{\theta} e^{-\beta H_{UV}(\mathbf{r},\theta) - \beta H_{VV}(\mathbf{r},\theta,\{n\})}\right) \quad (1)$$

Here, $\{n\}$ denotes instantaneous distribution of occupied and empty lattice points, $H_{UV}(\mathbf{r},\theta)$ is a solute−solvent interaction energy including electrostatic and Lennard-Jones (LJ) contributions (SPC/E parameters for water and standard force field mixing rules are used), $H_{VV}(\mathbf{r},\theta,\{n\})$ is a mean-field solvent−solvent term, and k_BT and β are the Boltzmann constant times temperature (T = 300 K is assumed) and its inverse, respectively.

The solvent−solvent term assumes only hydrogen bond-like interactions between lattice points. The number of bonds (0 to 4) in which a water probe participates at its given position and

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

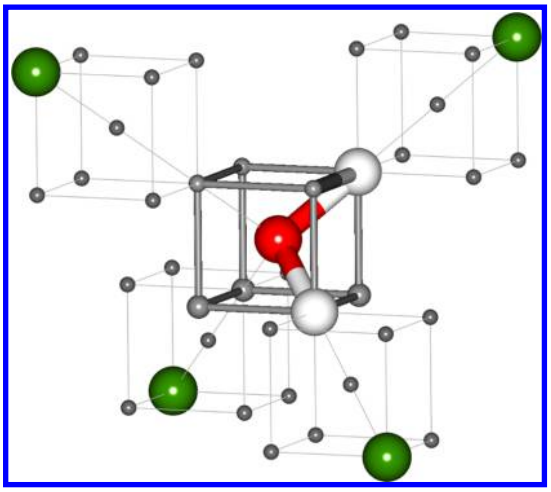


Figure 1. Fragment of BCC lattice with a water probe built into an elementary cell (sticks). Four green spheres indicate the lattice points considered for hydrogen bonding between the water probe at its current orientation and the rest of the solvent (see text for description).

orientation depends on the occupancies of four nearby lattice points (Figure 1, green spheres): two that are three grid spacings away in the direction of hydrogen atoms (which gives the distance of 3 Å from oxygen center and remains within the length limit for water−water hydrogen bond[69]), and two that are at the direction of oxygen lone pairs at the same distance. A bond is created when solvent density at its respective lattice point is greater than zero. The assumed energy contributed by a single hydrogen bond, ϵ_{HB}, is constant and is one of five adjustable parameters of the model.

Solvent distribution, that is, the configuration of occupied and vacated lattice points in the presence of a solute, is determined in an iterative, self-consistent manner. The process starts with uniform solvent density of bulk water, ρ_b, assigned to all lattice points. The state of each lattice point is then determined by its effective Hamiltonian:

$$\rho(\mathbf{r}) = f(H_{eff}(\mathbf{r}, \{n\})) = \begin{cases} \rho_b(1 + a(H_b - H_{eff})), & \text{if } 1 + a(H_b - H_{eff}) \geq \eta, \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

H_b in the above, corresponds to the assumed bulk excess chemical potential, $a > 0$ is a density scaling factor, and η is a "drying threshold," a fraction of bulk density, below which a given cell is considered as no longer occupied. All those three quantities are adjustable parameters whose values were determined in our previous work directed at optimization of model performance with respect to hydration free energy predictions.[67] Calculations of updated $H_{eff}(\mathbf{r},\{n\})$ for the entire lattice and density adjustments according to eq 2 are repeated until stationary solvent distribution is reached, that is, until no more lattice points become vacated due to their H_{eff} being too unfavorable.

2.1.2. Solvent Partitioning. The above procedure results in predicted $H_{eff}(\mathbf{r})$ distribution in space (H_{eff} dependence on solvent lattice configuration, $\{n\}$, is omitted in the following notation for brevity). In the context of protein hydration, of particular interest are regions of occupied lattice with favorable H_{eff}, buried within macromolecular structures; their partitioning into distinct hydration sites and subsequent evaluation of respective water binding free energies is the main goal of the current work. In this respect, however, the results obtained upon considering just a single placement of a protein on the solvent lattice are of limited use. Low spatial resolution of the BCC lattice used to represent solvent degrees of freedom (1 Å distance between the nearest neighbors is imposed by OH bond length in the SPC/E water model), together with only 12 orientations of a water probe relative to protein environment that are considered for each lattice point, provide for limited sampling.

In order to circumvent this issue, we perform multiple, independent calculations for random orientations of the BCC lattice with respect to protein structure and accumulate the results in a reference Cartesian grid of 0.5 Å resolution. Once solvent evolution is completed, H_{eff} value for each occupied point of the BCC lattice is mapped to the closest node of the Cartesian grid and stored there. Having performed desired number of N independent calculations, we use the following algorithm to determine the positions of putative hydration sites and corresponding water binding free energies:

(1) We create a map of preliminary estimates of excess chemical potential, $\tilde{\mu}(\mathbf{r}_0)$, on the Cartesian grid (Figure 2A) by

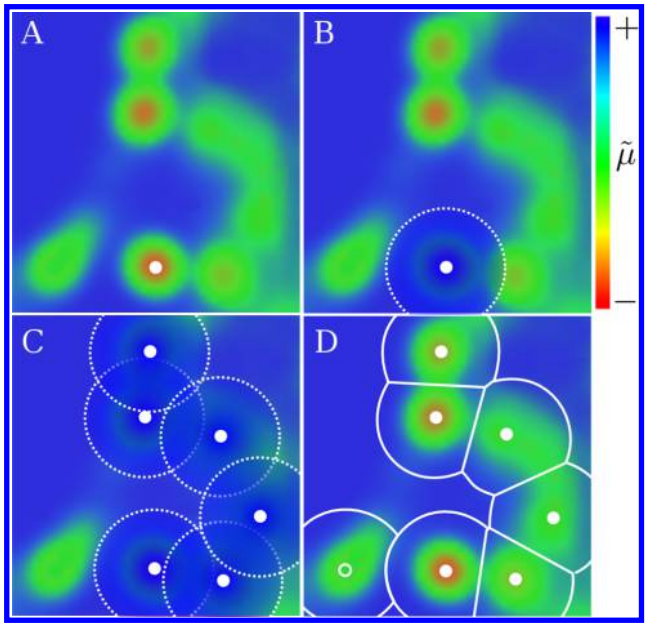


Figure 2. Schematic, 2D map of preliminary estimates of excess chemical potential, $\tilde{\mu}$, illustrating subsequent steps of the solvent partitioning procedure. Blue, unfavorable; red, favorable $\tilde{\mu}$ values.

Boltzmann averaging H_{eff} values accumulated during N runs over a sphere of 1.4 Å radius, centered at $\mathbf{r}_0$, $S(\mathbf{r}_0)$:

$$\tilde{\mu}(\mathbf{r}_0) = -k_B T \ln \sum_{\mathbf{r}\in S(\mathbf{r}_0)} \sum_{N} e^{-\beta H_{eff}^{N}(\mathbf{r})} \quad (3)$$

The summation over N includes only $H_{eff}(\mathbf{r})$ values actually accumulated at point $\mathbf{r}$ of the Cartesian grid; if no $H_{eff}^{i}(\mathbf{r})$ value was mapped following the i-th calculation, then the respective Boltzmann term is taken as zero.

(2) We select the point with minimal $\tilde{\mu}$ value as a hydration site and add a repulsive, Weeks−Chandler−Andersen (WCA) potential, V_{WCA},[70] centered around it to the remaining $\tilde{\mu}$ values

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

(Figure 2B); the details of the WCA potential parametrization are described in Supporting Information.

(3) We repeat the procedure described in point 2, seeking this time for the minimal $\tilde{\mu} + V_{WCA}$ value, until any node $\mathbf{r}$ on the Cartesian grid can be found such that $\tilde{\mu}(\mathbf{r}) + V_{WCA}(\mathbf{r}) < \mu_b$ (Figure 2C). Here, the μ_b parameter plays a role of bulk excess chemical potential, and its derivation and more detailed interpretation will be provided in the next paragraph.

(4) Having determined a set of $\{\mathbf{r}_i\}$ hydration site locations, we partition points of the Cartesian grid into hydration sites volumes V_i, by assigning each point of the grid to its closest hydration site location, whenever such a location can be found within 2.8 Å distance (Figure 2D).

(5) Finally, we partition all unassigned points into cavities (empty sites), using a procedure analogous to the one described in points 2−4, with the exception that iteration of point 3 is continued until no grid node with $V_{WCA}(\mathbf{r}) = 0$ can be found, that is, until each node is within the WCA potential range of at least one hydration site.

The last point is introduced to determine water binding free energy to sites that should not be hydrated according to the model but still show some residual solvent presence (i.e., include Cartesian grid points with some accumulated H_{eff} values).

2.1.3. Water Binding Free Energy. Once the positions of putative hydration sites are determined and the remaining points on the Cartesian grid are partitioned into their corresponding volumes, water binding free energy at each site i is obtained as a difference between the estimate of water excess chemical potential, $\mu(\mathbf{r}_i)$, and the reference, bulk water excess chemical potential, μ_b. The former one is evaluated as

$$\mu(\mathbf{r}_i) = -k_B T \ln \sum_{\mathbf{r}\in V_i} \sum_{N} e^{-\beta H_{eff}^{N}(\mathbf{r})} + V_{WCA}(\mathbf{r}_i) \quad (4)$$

where the summation over $\mathbf{r}$ includes all points belonging to hydration site volume V_i, and $V_{WCA}(\mathbf{r}_i)$ corresponds to the WCA potential accumulated at $\mathbf{r}_i$ during solvent partitioning.

The estimation of μ_b requires assumptions regarding the configuration volume per water molecule in the bulk solvent. In order to obtain μ_b values based on equivalent amount of sampling as entered into eq 4, we assume that the bulk water molecule occupies a volume of Ω Cartesian lattice points and we express μ_b as

$$\mu_b = -k_B T \ln\left(\Omega N \frac{v_c}{v_b} e^{-\beta H_b}\right) \quad (5)$$

Here, v_c/v_b is the ratio of volumes per lattice point for Cartesian (c) and BCC (b) lattices, which multiplied by N, the number of considered BCC lattice orientations, gives the expected number of samples per Cartesian lattice point for uniform, bulk-like solvent distribution on the BCC lattice. Finally, water binding free energy to the i-th hydration site is given as

$$\Delta F_i = \mu(\mathbf{r}_i) - \mu_b \quad (6)$$

Positions of buried water molecules and their binding free energies reported in the following were determined using model calculations with $N = 10^4$ lattice orientations.

By adjusting Ω, the volume per water molecule in the bulk (which is *a priori* not known under model assumptions), the values of μ_b, and consequently ΔF, can be arbitrarily shifted along the energy scale. Thus, the value of Ω was fitted to best reproduce reference water binding free energies obtained with explicit solvent simulations, making it an additional, sixth adjustable parameter of the model. We note that Ω was the only parameter introduced and adjusted in this work. All of the remaining 5 parameters were set previously to provide optimal hydration free energies for small solutes[67] and remained unchanged: $\epsilon_{HB} = -2.65$ kcal/mol, $H_b = -9.94$ kcal/mol, $\eta = 0.28$, $a = 0.112$ (kcal/mol)$^{-1}$, and $\epsilon_g = 7.0$ (this last parameter represents an outer grid dielectric constant and plays a role only in hydration free energy calculations).

The fitting of Ω parameter followed a single round of model calculations for all considered protein structures and was done by determining the μ_b value that provided the best agreement of ΔF estimates with DDM calculations for all sites within identical, rigid protein structures (see below). The obtained Ω value was included into the model, and all results presented in the following were recalculated with this final model version.

2.2. Proteins and Hydration Sites. In order to select protein structures for the evaluation of model performance, we searched Protein Data Bank (PDB; http://ww.rcsb.org/pdb) with following criteria: macromolecule type−protein, number of chains − 1, X-ray resolution <2 Å, no ligands, and no modified residues. For the resulting 2263 structures, the following descriptors were calculated with the VMD program:[71] the number of buried water molecules (defined as crystallographic water molecules with no solvent accessible surface area, SASA, evaluated with the standard VMD SASA method, using a 1.4 Å solvent probe), the number of protein atoms (N_A), radius of gyration (R_g), and compactness (N_A/R_g^3). The results were then inspected in order to possibly find small and compact, globular proteins with at least a single buried water molecule, leading to the selection of 8 structures with chain lengths ranging from 65 to 167 amino acids.

All selected protein structures were than subjected to a common preparation procedure with the use of a protein preparation tool of the Schrödinger software suite[72] which involved: the removal of all crystallographic water molecules, addition of hydrogen atoms in configuration corresponding to pH 7.0 (PropKa module was used for this purpose), and selection of energetically favorable conformers of amide side chains. Finally, the Amber ff99SB*ILDN force field[73] was assigned, and hydrogen positions were optimized using *in-vacuo* energy minimization in Gromacs,[74] while keeping heavy atoms constrained in crystallographic positions.

In order to find all potentially hydratable cavities in the considered proteins, we scanned their force-field parametrized structures with a spherical probe representing LJ properties of the SPC/E water model. The probe was placed at points of the BCC lattice with 0.2 Å distance between the nearest neighbors, spanning the entire protein structure (the BCC lattice was selected as it provides the optimal sampling efficiency in 3 dimensions for a given number of points[75]). Cavities were defined as clusters of points with LJ energy <5 kcal/mol with no path leading through the nearest neighbors to protein exterior.

Detected cavities served as a basis for discrimination between hydrated and empty sites. Crystallographic water molecules were superimposed on protein structures with the obtained cavity coordinates, and the initial set of buried water molecules based on SASA classification was updated, by merging it with all water molecules closer than 1.4 Å to any cavity point. This allowed the inclusion of water molecules remaining in cavities large enough to produce a "solvent accessible surface" inside of protein. The positions of buried water molecules were regarded as centers of "hydration sites." Finally, all cavity points closer

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

than 2.8 Å to any hydration site water oxygen were discarded, and the remaining points were classified as "empty cavities."

A predicted water molecule was categorized as belonging to a hydration site or empty cavity if crystallographic water or *any* cavity point was found within 1.4 Å distance, respectively. If no predicted solvent was found within this distance, a given site or cavity was considered as undetected. We also assured that (a) there were no situations when more than one predicted water molecule was assigned to a single crystallographic water and that (b) there were no unassigned predictions.

2.3. Explicit Solvent Simulations. Binding free energies of water into all hydratable sites (actual hydration sites and cavities) were evaluated with the use of a DDM[44] based on explicit solvent MD simulations. We considered two scenarios: (a) binding into rigid protein structure constrained at crystallographic geometry, which allowed direct comparison with model predictions that are per definition based on a single protein conformation, and (b) binding into freely moving protein, giving insight into more realistic hydration properties, which were compared with model predictions averaged over a set of MD structures.

In the case of binding to rigid proteins, their structures (including hydrogen atoms) were fixed throughout simulations, and absolute harmonic restraints were applied to the oxygen atom of the decoupled water molecule. In the case of flexible proteins, the oxygen atom of the water molecule under study was restrained relative to three protein atoms (typically, three Cα atoms of the neighboring main chain region) using a combination of dihedral, angle, and distance restraints in the form introduced by Boresch.[76] In both cases, rotational freedom of the decoupled water molecule was unrestrained.

Prior to DDM simulations for each site, free MD runs of 500 ps and 5 ns length were executed for fixed and flexible proteins, respectively, with the unrestrained water molecule manually placed into the site of interest in order to determine equilibrium positions/lengths of the restraining potential terms and their harmonic constants. Equilibrium values and force constants of the subsequent restraints were based on the average and variance, respectively, of unrestrained water position during simulation, resulting in parameter values that provide optimal sampling of the binding site volume during the decoupling stage.[77]

Free energies for restraining water in a protein environment (ΔF^{in}_{restr}) and subsequent decoupling (ΔF_{ub}) (Figure 3) were calculated using together 25 windows, each of 2 ns total length, during which the restraining potential was applied (4 windows), followed by switching off electrostatic interactions (10 windows) and LJ potential (11 windows) of the decoupled water molecule. The removal of LJ interactions was performed with the use of soft core potential, with standard Gromacs settings. During decoupling simulations for flexible proteins, an additional particle with the mass of a water molecule was added to the system, which was restrained with identical set of potentials as the decoupled water molecule and interacted only through repulsive WCA potential with the rest of unperturbed solvent. Its role was to repel solvent molecules that could bias the results by penetrating into the hydration site from the bulk during final decoupling stages of hydration site water. The range of WCA potential was equivalent to water−water separation giving 1 k_BT LJ energy for SPC/E water potential.

The free energy effect of liberating a water molecule from the restraining potential (ΔF^{out}_{restr}) was calculated analytically as

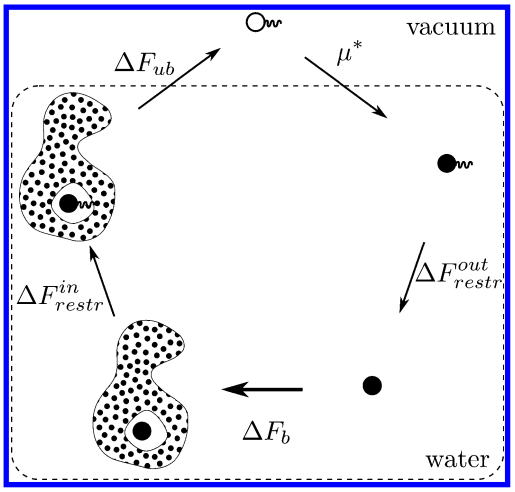


Figure 3. Thermodynamic cycle illustrating calculations of binding free energy (ΔF_b) with double decoupling method (see text for the description of subsequent steps).

$$\Delta F^{out}_{restr} = k_B T \ln\left[C^0 \left(\frac{2\pi k_B T}{k} \right)^{3/2} \right] \tag{7}$$

in the case of absolute restraints with isotropic harmonic constant k used in calculations for rigid protein structures, and as

$$\Delta F^{out}_{restr} = k_B T \ln\left[r_0^2 \sin\theta_0 C^0 \frac{(2\pi k_B T)^{3/2}}{(k_r k_\theta k_\phi)^{1/2}} \right] \tag{8}$$

in the case of Boresch restraints, where r_0 and θ_0 are equilibrium values of distance and angle potentials, respectively, and k_r, k_θ, and k_ϕ are distance, angle, and dihedral harmonic constants, respectively. In both cases, C^0 = 55 M is a standard liquid water concentration.

Free energy for introducing a decoupled water molecule back to pure solvent, μ^*, was calculated using reverse transformation (i.e., decoupling from bulk water) with the same decoupling scheme as that in the case of protein cavities but without the first 4 restraining windows, using steady restraints with a force constant of 2.4 kcal/mol/Å^2. The obtained μ^* = −7.0 ± 0.1 kcal/mol does not depend on the restraining potential and needed to be calculated only once.

Finally, based on the thermodynamic cycle depicted in Figure 3, water binding free energy to the protein hydration site was evaluated as

$$\Delta F_b = -\Delta F^{in}_{restr} - \Delta F_{ub} - \mu^* - \Delta F^{out}_{restr} \tag{9}$$

In cases of hydration sites occupied by more than one water molecule, binding free energy was estimated by stepwise removal of hydration site water during which any remaining water molecules were free to expand to the vacated volume. For flexible protein structures, water decoupling from such multiply hydrated sites did not involve using a repelling WCA sphere, but it was manually checked that no water molecule from the surrounding solvent entered the site. In the case of rigid structures for consistency check, additional simulations with simultaneous decoupling of all molecules were conducted.

Free energy calculations were performed using the Bennet acceptance ratio method as implemented in Gromacs. Their

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

uncertainty was estimated as the standard error of the mean value, obtained by block averaging over 5 simulation blocks.

All of the above simulations were carried out with the use of the Gromacs program. Protein structures were parametrized with the Amber ff99SB*ILDN force field, and the SPC/E water model was used for solvent in all cases. The choice of SPC/E instead of TIP3P water potential, along which the Amber force field family was designed, was dictated by the fact that the developed semiexplicit hydration model relies on SPC/E water geometry and parameters (TIP3P model, which assumes a 104.5° HOH angle, would not fit into the BCC lattice). We note, however, that the combination of Amber force field with both SPC/E and TIP3P water models was shown to provide consistent hydration properties.[78,79] The proteins under study were solvated in dodecahedral boxes with 10 Å solvent margins, with periodic boundary conditions. Equations of motion were propagated with a time step of 2 fs at a constant temperature of 310 K, maintained with the use of a leapfrog stochastic dynamics integrator, electrostatic interactions were treated using the particle mesh Ewald method, and LJ potential was handled with the Verlet cutoff scheme, with a cutoff distance of 10 Å. Simulations with rigid protein structures were conducted in the NVT ensemble, while for flexible proteins the NpT ensemble with pressure of 1 bar, maintained by a Parrinello−Rahman barostat, was used.

3. RESULTS

3.1. Protein Cavities and Bound Water. The analysis of 8 considered crystallographic structures revealed 48 buried cavities, holding together 42 water molecules. Twenty-five sites were occupied by a single water molecule, 8 contained two or more, and 15 were empty. Volume distributions were similar in the case of hydrated and empty sites (Figure 4, left), with

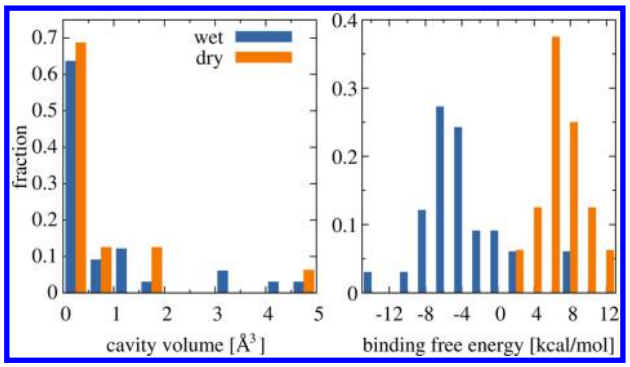


Figure 4. Distributions of cavity volumes (left) and water binding free energies into rigid protein structures (right) for sites containing crystallographic water (wet) and empty cavities (dry).

clear maxima for small volumes below 0.5 Å^3 (note that as a cavity volume, we consider space available to the center of the water oxygen atom; see the Methods section) and tails reaching 5 Å^3. As expected, water binding free energies calculated with the DDM method are generally favorable for hydrated sites (Figure 4, right), though positive values were obtained for 9 crystallographic water molecules when rigid protein structures were considered. Once protein flexibility is allowed, 4 out of those 9 water molecules regain favorable binding free energies; however, the remaining 5 either escape from their binding sites during 5 ns of free MD simulations or have clearly positive ΔF_b, indicating possible crystallographic artifacts. On the contrary, no favorable binding ($\Delta F_b < 0$) was observed toward cavities showing no X-ray water, both for rigid and flexible proteins.

3.2. Model Predictions: Rigid Structures. *3.2.1. General Results.* The comparison of model predictions with the results of DDM calculations for rigid protein structures is shown in Figure 5, and detailed free energy values are presented in Tables

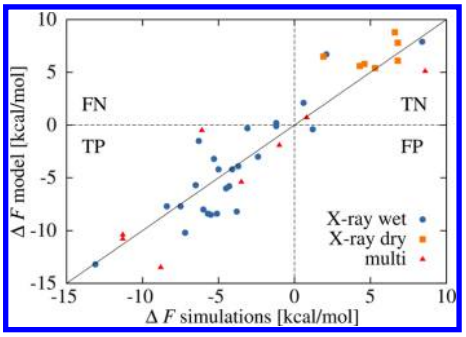


Figure 5. DDM results vs model predictions for water binding free energies into rigid protein cavities with crystallographically detected water molecules (X-ray wet), empty (X-ray dry), or containing more than one water molecule (multi). Dashed lines divide the plot into: TP, true positive; TN, true negative; FP, false positive; and FN, false negative sections.

1 to 3. In general, qualitative agreement is fairly good: the model clearly captures the distinction between wet and dry cavities and correctly detects the set of suspicious crystallographic waters with unfavorable binding free energies. If we take $\Delta F = 0$ as a borderline between the presence and absence of water and simplistically assume that the total number of sites to consider is 57 (the number of crystallographic water molecules plus the number of empty cavities), the model correctly identifies 29 of 33 wet sites, while leaving 23 of 24 empty sites as unoccupied.

As can be observed in Table 2, in the case of some particularly hydrophobic cavities (typically with $\Delta F_b > 6$ kcal/mol according to DDM simulations) no binding free energy values were obtained at all. Such results are expected and indicate that no occupied lattice points were found in a given region upon completion of the solvent evolution phase, thus not allowing for the estimation of local excess chemical potential (eq 4).

Binding free energies obtained with the model are also in fair quantitative agreement with DDM results. If we focus on single hydrated sites and dry cavities, for which numerical estimates of binding free energy were obtained, Pearson's correlation between the two sets of predictions is 0.94, with the mean square error (RMSE) of free energy values of 2.3 kcal/mol, and mean signed error of 0.06 kcal/mol. In 4 of 32 cases, absolute errors are particularly large (~4.5 kcal/mol). Intriguingly, aside from the fact that all those hydration sites are relatively small, with volumes <0.2 Å^3, our analysis of local protein environment did not reveal anything specific. In order to resolve this issue, given the fact that the model is tested against reference calculations based on exactly the same force field, we focused on two possible sources of errors: model limitations due to the simplification of physical description and sampling problems.

3.2.2. Model Limitations. The major simplification upon which the model is founded is the introduction of a mean-field term reducing solvent−solvent interactions to five energy levels. They are based on the number of "hydrogen bonds" in which a solvent probe can participate, depending on its position and orientation. While justified in the context of small molecules hydration, where thermodynamic contributions are to a large

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

Table 1. Summary of Predictions for Water Binding to Crystallographic Hydration Sites[a]

PDB	water	vol.	ΔF_b^{rigid}	ΔF_{model}^{rigid}	ΔF_b^{flex}	X/R/M/F
1aho	108	0.3	−5.3 ± 0.1	−3.2	−2.1 ± 0.2	1/1/1/1
1snb	172	0.2	−8.4 ± 0.1	−7.7	−1.8 ± 0.4	1/1/1/1
1snb	196	0.9	−7.2 ± 0.1	−10.2	−2.5 ± 0.2	1/1/1/1
1bas	201	0.3	−7.5 ± 0.1	−7.7	−6.0 ± 0.3	1/1/1/1
1bas	210	0.1	−6.3 ± 0.1	−1.5	−6.5 ± 0.2	1/1/1/1
1bas	212	1.1	−3.1 ± 0.1	−0.3	−3.2 ± 0.3	1/1/1/1
1bas	219	0.1	−6.5 ± 0.1	−5.7	−2.1 ± 0.3	1/1/1/1
1w8v	2046	1.5	−4.1 ± 0.1	−4.2	−2.4 ± 0.1	1/1/1/1
1w8v	2047	1.4	−1.2 ± 0.1	0.2	−2.2 ± 0.3	1/1/0/1
1w8v	2066	0.3	−3.7 ± 0.1	−3.9	−2.2 ± 0.3	1/1/1/1
1w8v	2122	0.3	−1.2 ± 0.1	−0.1	−2.8 ± 0.2	1/1/1/1
1w8v	2127	0.3	−6.0 ± 0.1	−8.0	−4.7 ± 0.5	1/1/1/1
1w8v	2133	0.2	−13.1 ± 0.1	−13.2	−5.6 ± 0.6	1/1/1/1
2ygs	140	0.0	+8.4 ± 0.1	+7.9	escapes	1/0/0/0
3hvv	304	0.5	−4.3 ± 0.1	−5.8	−3.3 ± 0.1	1/1/1/1
3hvv	305	0.0	−5.0 ± 0.1	−4.2	−1.1 ± 0.4	1/1/1/1
3hvv	307	0.4	−2.4 ± 0.1	−3.0	−0.8 ± 0.3	1/1/1/1
3hvv	342	0.0	−4.5 ± 0.1	−6.0	−3.4 ± 0.4	1/1/1/1
1uoy	2078	0.1	−3.8 ± 0.1	−8.2	−2.4 ± 0.4	1/1/1/1
1uoy	2115	0.2	+2.1 ± 0.1	+6.7	+1.4 ± 0.1	1/0/0/0
3q7y	1	0.1	−5.5 ± 0.1	−8.5	−2.4 ± 0.5	1/1/1/1
3q7y	2	0.2	−5.1 ± 0.1	−8.4	−2.2 ± 0.6	1/1/1/1
3q7y	3	0.2	−5.7 ± 0.1	−8.4	−2.8 ± 0.6	1/1/1/1
3q7y	144	0.6	+1.2 ± 0.1	−0.4	−3.3 ± 0.6	1/0/1/1
3q7y	225	0.2	+0.6 ± 0.1	+2.1	−2.5 ± 0.8	1/0/0/1

[a]vol., site volume in Å^3; ΔF_b^{rigid} and ΔF_b^{model}, DDM and model binding free energies to rigid X-ray structures, respectively; ΔF_b^{flex}, DDM predictions for fully flexible protein structures; all free energies are in kcal/mol; X/R/M/F, occupancy prediction (0−dry, 1−wet) by X-ray, rigid DDM, model, and flexible DDM, respectively.

Table 2. Summary of Predictions for Water Binding to Empty Cavities in Crystallographic Structures[a]

PDB	cavity	vol.	ΔF_b^{rigid}	ΔF_{model}^{rigid}	ΔF_b^{flex}	X/R/M/F
1bas	V117	0.7	8.2 ± 0.2		4.7 ± 0.1	0/0/0/0
1bas	L83	0.6	6.2 ± 0.1		escapes	0/0/0/0
1bas	M77	0.2	8.0 ± 0.1		4.2 ± 0.2	0/0/0/0
1bas	C26	0.1	1.9 ± 0.1	6.5	2.8 ± 0.3	0/0/0/0
1w8v	V20	1.9	4.6 ± 0.1	5.8	escapes	0/0/0/0
1w8v	C62	0.4	6.6 ± 0.1	8.8	0.4 ± 0.2	0/0/0/0
1w8v	L98	0.1	9.5 ± 0.1		escapes	0/0/0/0
2ygs	L16	0.4	6.8 ± 0.1	7.8	escapes	0/0/0/0
3hvv	V65	4.5	4.3 ± 0.1	5.6	escapes	0/0/0/0
3hvv	F68	1.9	5.3 ± 0.1	5.4	escapes	0/0/0/0
3hvv	V120	0.1	6.4 ± 0.1		escapes	0/0/0/0
3hvv	V137	0.1	9.0 ± 0.1		escapes	0/0/0/0
3q7y	L14	0.1	12.3 ± 0.1		escapes	0/0/0/0
3q7y	I25	0.4	6.8 ± 0.1	6.1	escapes	0/0/0/0
3q7y	L56	0.1	8.7 ± 0.1		escapes	0/0/0/0

[a]Cavities are designated by neighboring protein residues. All columns are as described in Table 1.

extent determined by local properties of the first hydration shell, such an approach may lead to inaccuracies when hydration sites buried inside protein structures are considered: due to the short-range of the mean field term (the assumed length of a hydrogen bond is 3 Å), such isolated lattice regions are decoupled from the reaction field generated by the solvent surrounding the protein.

In order to verify the importance of this effect, we evaluated water binding free energies into rigid protein structures kept in vacuum, using the same DDM scheme as that for hydrated protein structures. Strikingly, differences in ΔF_b caused by the solvent reaction field are often beyond 3 kcal/mol, both in favor or against binding (see Supporting Information for detailed results). Interestingly, the magnitude of this effect seems not to correlate with the distance between the buried water molecule and protein hydration shell (measured as a minimal oxygen−oxygen distance between the water molecule of interest and the rest of solvent, averaged over simulation frames of the unperturbed system): shifts in ΔF_b in the order of 4 kcal/mol were observed for some hydration sites separated by more than 6 Å from the rest of solvent, while little or no effect was sometimes found for distances below 4 Å (however, no hydration site was found closer than 3 Å to the first hydration shell). A typical scenario when a pronounced shift in binding free energy was observed included cases in which several water molecules were well ordered by local interactions at the protein surface in configurations producing strong electrostatic effect in the remote buried binding site.

Remarkably, the comparison of model predictions with binding free energies toward "dry" proteins (Figure 6) results in RMSE of 1.1 kcal/mol, which is further reduced to 0.3 kcal/mol once a systematic shift in hydration free energies is accounted for (ΔF_b values for "dry" proteins are on average by 0.9 kcal/mol more favorable than those for their hydrated counterparts). Such an excellent agreement suggests that the effect of solvent reaction field, which is not captured by the model, is indeed an important contribution to discrepancies between model predictions and DDM results for hydrated proteins.

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

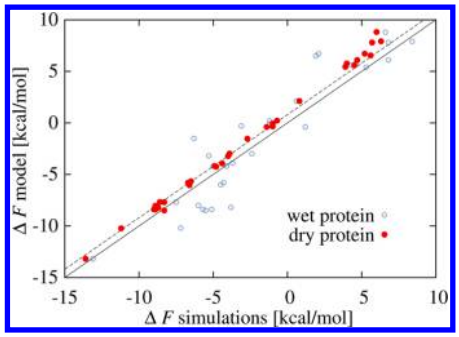


Figure 6. Comparison of model predictions with binding free energies estimated using DDM for proteins in solvent (wet protein) and in vacuum (dry protein). The dashed line shows an average shift of −0.9 kcal/mol for dry protein results.

3.2.3. Sampling Convergence. The amount of sampling for ΔF calculations depends on the number of considered random solvent lattice orientations with respect to protein structure. For each such orientation, a solvent evolution run is performed to produce a coarse estimate of excess chemical potential distribution in space. The obtained values are accumulated during subsequent runs and Boltzmann-averaged to evaluate binding free energies according to eqs 4−6. In order to determine conditions necessary to provide converged results, we analyzed ΔF estimates for all single-occupied or empty hydration sites in three protein structures (pdb: 1BAS, 1W8V, and 1UOY) based on increasing amount of sampling. We considered 5 sampling schemes with the number of random lattice orientations ranging from 10 to 10^4, and for each such scheme, we performed a series of 50 independent ΔF evaluations (Figure 7).

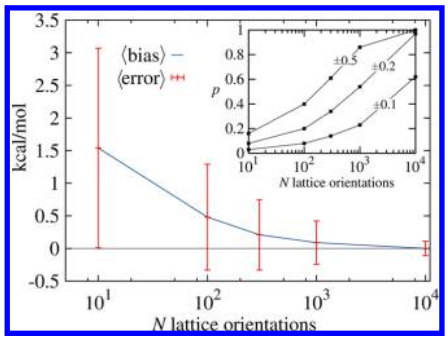


Figure 7. Convergence of free energy estimates for increasing number of considered lattice orientations, N. An average bias corresponds to the difference of average results for all sites; for 50 independent runs with respect to averages for $N = 10^4$, an average error corresponds to an average standard deviation obtained over 50 runs. Inset: probability of obtaining a result within ± energy range (in kcal/mol) from the converged value, estimated with error function, based on the assumption of the normal distribution of independent free energy predictions.

As can be expected for free energy estimates, ΔF values become more negative as the amount of sampling increases. At the same time, their uncertainty, measured for each sampling scheme as an average standard deviation for all sites, for each of them based on 50 independent results, decreases. If we assume that averages of the set involving 10^4 lattice orientations represent converged results, useful ΔF predictions, with small positive bias for an average hydration site of 0.09 ± 0.11 kcal/mol and an average standard deviation of 0.3 ± 0.1 kcal/mol, can be obtained for 1000 protein orientations. Assuming normal distribution of ΔF values, they fall within the ±0.5 kcal/mol from the converged estimates with a probability of 0.86. Both positive bias and uncertainties are larger in the group of hydration sites with small volumes (<0.2 Å^3); however, as the number of considered protein orientations increases the differences with respect to values observed for larger sites monotonically tend toward zero, indicating no systematic problems related to sampling.

3.2.4. Transferability. The results discussed here involve the fitting of one parameter, Ω, which specifies the number of Cartesian lattice points that contribute to the volume used to estimate bulk water excess chemical potential μ_b (eq 5). The μ_b value serves as a reference for binding free energy estimates; therefore, its changes uniformly shift all ΔF predictions on the energy scale (eq 6). Accordingly, we tuned the Ω value to obtain μ_b, which provides zero mean error of ΔF with respect to DDM results for 32 cases corresponding to single hydrated sites and dry cavities, for which numerical estimates of ΔF were obtained. We found Ω = 35, which amounts to 4.4 Å^3 per water molecule in the bulk, to provide such an optimal μ_b. In order to evaluate how sensitive the μ_b value is to training data, we performed 8-fold cross-validation. The set of 32 binding sites was randomly split into 8 subsets, and model performance for each of them was evaluated after using the remaining 7 for Ω fitting. μ_b values obtained in 8 cross-validation runs were consistent, showing a standard deviation of 0.19 kcal/mol. The RMSE for ΔF values used for validation (i.e., not involved in fitting) was 2.33 kcal/mol, which is only slightly higher than the 2.25 kcal/mol obtained for the Ω value based on the entire set, as reported above. Overall, these results seem to indicate model transferability to different protein structures.

3.2.5. Multiple Sites. According to X-ray data, there are 8 multiple hydrated sites in the considered protein set, hosting together 17 water molecules. In order to evaluate reference binding free energies, we performed two kinds of calculations: one in which water molecules were subsequently decoupled from the binding site and another in which all involved water molecules were decoupled simultaneously. In the first scheme, harmonic restraints were applied only to the decoupled water molecule allowing its neighbors to expand freely into vacated space. Such stepwise calculations give insight into the optimal hydration number of a given site; however, they may be more error-prone than single-step ΔF_b estimates. Accordingly, the latter were included as a consistency check (Table 3).

Somewhat surprisingly, according to DDM estimates of binding free energies toward rigid protein structures, as many as 5 of 17 water molecules inhabiting multiple hydrated sites have $\Delta F_b > 0$. The results suggest that two sites may be completely dry and that one may be filled by only a single water molecule. The model reproduces those findings quite well, correctly predicting occupancies in 6 out of 8 cases. It detects two dry sites, also indicating that one of them (in the 1SNB structure) is very hydrophobic and that the second one (in the 3Q7Y structure) is only moderately hydrophobic for the first water molecule. In the case of an apparently single occupied site in the 3HVV structure, the model also places there only one water molecule and predicts its binding free energy with moderate, 0.9 kcal/mol difference with respect to DDM calculations. For one of the two mispredicted sites (waters 224 and 238 in the 1BAS structure), quantitative error is relatively small: the model indicates that binding of a second water is slightly unfavorable

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

Table 3. Summary of Predictions for Multiply Hydrated Sites[a]

PDB	water	vol.	ΔF_b^{step}	ΔF_b^{simult} ΔF_b^{steps}	ΔF_{model}^{site}	ΔF_{model}^{TOT}	X/R/M/F
1snb	137	0.2	+8.6 ± 0.1	+8.6 ± 0.1	+5.1	+5.1	1/0/0/0
	150		escapes	*+8.6* ± 0.1			1/0/0/0
1bas	224	4.6	−3.3 ± 0.2	−3.5 ± 0.2	−5.4	−5.4	1/1/1/1
	238		−0.5 ± 0.1	−3.8 ± 0.2	0.2		1/1/0/1
1bas	202	4.3	−6.1 ± 0.3	−11.3 ± 0.2	−5.0	−10.8	1/1/1/1
	203		−5.6 ± 0.1	−11.7 ± 0.3	−5.8		1/1/1/1
1w8v	2113	3.2	−6.3 ± 0.1	−11.3 ± 0.2	−7.3	−10.4	1/1/1/1
	2119		−4.9 ± 0.4	−11.2 ± 0.4	−3.1		1/1/1/1
1w8v	2076	1.3	−2.3 ± 0.1	−6.1 ± 0.2	−0.5	−0.5	1/1/1/1
	2077		−3.0 ± 0.1	−5.7 ± 0.4	+0.4		1/1/0/1
	2094		−0.4 ± 0.3		+2.0		1/1/0/1
3hvv	296	3.1	−8.0 ± 0.1	−8.8 ± 0.2	−5.4	−13.5	1/1/1/1
	297		−0.9 ± 0.1	−8.9 ± 0.2	−8.1		1/1/1/1
3hvv	300	1.6	−1.0 ± 0.1	+0.4 ± 0.2	−1.92	−1.92	1/1/1/1
	306		+1.3 ± 0.1	−1.0 ± 0.1			1/0/0/1
3q7y	165	0.4	+0.8 ± 0.1	+3.6 ± 0.2		> 0.7	1/0/0/1
	170		+2.6 ± 0.2	*+0.8* ± 0.1	0.7		1/0/0/0

[a]vol., hydration site volumes in Å^3; ΔF_b^{step} and ΔF_b^{simult}, DDM free energies for stepwise and simultaneous water binding, respectively; ΔF_b^{steps}, total DDM free energy for the hydration site, accounting for the actual number of bound water molecules; ΔF_{model}^{site} and ΔF_{model}^{TOT}, model estimates for each water and entire hydration site, respectively; all free energies are in kcal/mol; X/R/M/F, occupancy prediction (0−dry, 1−wet) by X-ray, rigid DDM, model, and flexible DDM, respectively.

(with ΔF = 0.2 kcal/mol), while DDM estimates its binding free energy as only −0.5 kcal/mol. Definitely, a wrong prediction is obtained for one site hosting 3 water molecules (in 1W8 V structure): according to the model, the site should only weakly bind a single water molecule.

In general, if we take adjusted results, that is, assume contributions to ΔF per site only from actually bound water molecules, the RMSE between DDM calculations and model predictions is 3.0 kcal/mol with a mean error of −0.5 kcal/mol. This is certainly worse than that in the case of single-hydrated cavities, but one has to take into account the fact that predictions of binding to multiple-hydrated sites are significantly more difficult. First, obviously, ΔF values include contributions from more than one water, each with its own uncertainty. Second, more importantly, binding free energy depends here not only on water−protein interactions but also on the effect of mutual interactions between buried water molecules, whose capture is not trivial for a simplified model. In order to estimate the strength of this effect, we used the model to perform additional calculations for doubly hydrated sites in 1BAS and 1W8V structures. In each case, we estimated binding free energies to both positions in the binding site by water molecules that could not feel their partners (the respective part of a site was forced to remain dry by placing there a hard-sphere object of 2.8 Å diameter). The sum of such binding energies was typically larger than ΔF for the complete site, indicating a stabilizing effect of water−water interactions of ∼2 kcal/mol, which qualitatively agrees with the expected energy of a single hydrogen bond.

3.3. Predictive Value for Flexible Structures. A practically important purpose of the proposed model is to obtain predictions for the occupancy of buried cavities in protein structures exhibiting full flexibility, for instance during unrestrained MD simulations. The relevant question here is what particular protein conformation should be used for such predictions. Crystallographic structures typically already contain some information regarding water placement; however, one may wish to verify or complement such information, also wondering what the probability is that the state of hydration sites indicated by crystallography or by the model is valid for flexible protein. In turn, if no experimentally validated protein structure is at hand, the predictions based on one arbitrary conformation may be of limited accuracy.

In order to assess whether meaningful information for flexible structures can be obtained in such scenarios, we considered (A) predictions based on crystallographic geometry and (B) predictions based on an ensemble of frames extracted from an MD simulation of protein structure from which all internal water was purposefully removed beforehand. We performed respective calculations for all 8 protein structures considered in this work and compared the results with the "true" distribution of buried water estimated by evaluating binding free energies, ΔF_b, to all sites in freely moving proteins using the DDM method (see the Methods section). As previously done, we consider any given site as "wet" if ΔF_b < 0 and "dry" otherwise, or if a water molecule spontaneously abandons it during 5 ns of free MD simulation (in such case, we did not calculate ΔF_b).

If we take just the distribution of crystallographic water molecules as a predictor for cavity occupancy, we find that it achieves the true positive rate of 1.0 (i.e., all truly wet cavities in the considered set are detected by X-ray crystallography), however, with a relatively high false positive ratio of 0.25, which

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

means that crystallographic water is found in 25% of cavities that should rather be dry. This translates to an overall accuracy (the probability of proper distinction between wet and hydrates site) of 0.91.

Model calculations for rigid crystallographic structures (scenario A) achieve similar accuracy, however, with fewer false positive predictions at the expense of smaller true positive rates (Table 4). It indicates that sites for which just slightly positive free energies were predicted can be considered as most likely hydrated. Indeed, shifting, model predictions just by −0.5 kcal/mol, while still maintaining $\Delta F = 0$ as a limiting value between wet and dry states, increases the true positive rate and model accuracy to 0.92 and 0.95, respectively. We refrain, however, from tuning model performance (by adjusting the Ω_0 value) toward optimizing these descriptors, as we believe that seeking the best agreement with DDM results for identical protein conformations is a more consistent method.

Table 4. Efficiency of Predictions for Water Placement in Flexible Protein Structures Based on Rigid Crystallographic Conformation[a]

	X-ray	ΔF_b^{rigid}	ΔF_{model}^{rigid}
TPR	1.00	0.89	0.81
TNR	0.75	0.95	1.00
FPR	0.25	0.05	0.00
FNR	0.00	0.11	0.19
ACC	0.91	0.93	0.88

[a]X-ray, crystallographic water; ΔF_b^{rigid} and ΔF_{model}^{rigid}, free energy estimates based on DDM and model calculations, respectively. TPR, true positive; TNR, true negative rate; FPR, false positive rate; FNR, false negative rate; ACC, accuracy.

Interestingly, DDM results for rigid protein structures show only marginally better performance than the model in predictions of "truly" wet sites, as they remain in only moderate quantitative agreement with DDM calculations for flexible structures. Most likely, it is due to the fact that calculations based on crystallographic structures represent water insertion to preorganized cavities, thus not accounting for the work against the protein environment necessary to stabilize hydration sites in configurations suitable for water binding. Apparently, such work varies from site to site, thus not enabling simple translation between free energy values obtained for rigid and flexible proteins.

Rearrangement of the protein environment is of particular importance when one considers water binding to structures whose models were created without any knowledge regarding internal solvent distribution. As buried water molecules typically gain most of the binding energies from directional hydrogen bonds, even a small alteration of local protein structure may lead to substantial change in binding energy. This problem is emulated in scenario B, where internal, crystallographic water was removed prior to simulations that served to generate representative protein conformations.

We assume here that water binding to its site would be best described by a Widom insertion-like process,[80] with many unsuccessful and some successful trials, with the latter having dominant contributions to binding free energy value. As a large number of protein conformations would be necessary to provide converged sampling, we do not aim at obtaining numerical estimates of binding free energies but rather just at determining whether a given cavity is hydrated or not. To this end, we consider that a site is hydrated when $\Delta F < 0$ was observed in at least in one of the evaluated conformations.

In order to evaluate the model performance, we proceeded as follows. On the basis of the application of the model to 50 protein conformations extracted every 100 ps from 5 ns of unperturbed MD runs, we estimated for each site i the probability, p_i, of finding favorable binding free energy in a single, random conformation (see Supporting Information for detailed values). The probability of finding a hydrated site in at least one of the k independent conformations is then $P_i^k = 1 - (1 - p_i)^k$. We assessed the quality of model predictions based on k protein conformations by considering fictitious computational experiments in which the state of each site was assigned randomly based on its P_i^k value and compared with flexible DDM results. For each k, such experiments were repeated 1000 times, allowing the estimation of expected true positive rate and accuracy, and their uncertainties (Figure 8).

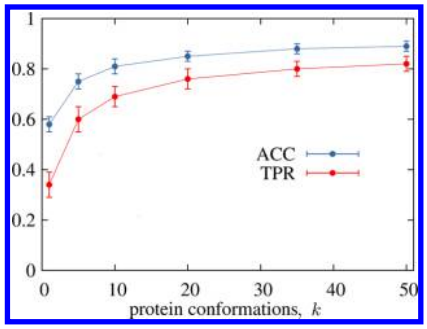


Figure 8. ACC, accuracy, and TPR, true positive rate, for predictions obtained using increasing numbers of protein conformations, k. Error bars correspond to one standard deviation obtained for 1000 trial experiments for each k.

As expected, predictions based on single, random protein conformation have low true positive rate (0.34 ± 0.05) indicating that many hydration sites fluctuate between accessible and inaccessible states when no water molecule is present inside. The true positive rate doubles for $k = 10$ conformations, and for $k = 50$, it reaches 0.82 ± 0.03, which is the value observed for crystallographic structures. Accuracy is systematically higher, owing to a high number of true negative predictions that result from the fact that hydrophobic cavities remain unfavorably hydrated regardless of their conformation, hence not producing false positive results. Interestingly, for $k \gtrsim 35$ the accuracy reaches the value of 0.89, also almost exactly the one observed for crystallographic structures.

A general observation arising from those results is that predictions based on a single protein conformation that does not include preorganized hydration sites may easily lead to wrong conclusions. We believe that this result applies as well to other methods which do not rely on sufficiently flexible protein structures. Perhaps heuristic approaches based on statistical descriptors may have an advantage here, being less sensitive to details of protein conformation. In any case, performing a systematic analysis in this respect may be of interest.

4. CONCLUSIONS

We presented an extension of our previously introduced semiexplicit discrete solvent model that allows the detection of buried hydration sites within macromolecular structures and estimation of water binding free energies. We tested its

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

performance on 8 protein structures with validated solvent distribution and demonstrated that the obtained predictions accurately discriminate between actually hydrated sites and sterically accessible to water but dry cavities.

Numerical estimates of water binding free energies were in satisfactory agreement with the results of much more time-consuming calculations employing the double decoupling method. Notably, the model is capable of detecting multiply hydrated sites, accounting as well for the stabilizing effect of mutual interactions between cobound water molecules. At the current stage, root-mean-square error between the two methods is in the order of 2 kcal/mol. It is probably too much to provide quantitatively meaningful estimates for contributions of buried water to receptor−ligand binding; however, the results should serve well for discrimination between tightly bound and easily displaceable water. We believe that the inclusion of solvent reaction field, whose influence on buried water free energies was shown to be substantial, may lead to significant improvement in the accuracy of free energy estimates. A possibility to do so with little numerical cost is currently under study.

To predict the locations of buried water molecules in protein structures the semiexplicit hydration model can be applied as a standalone method that does not rely on any preassumption regarding solvent distribution. All necessary input is provided just in the form of a force-field parametrized structure, and a single run delivers information regarding *all* hydration sites within minutes of computational time (the results refer to the execution of nonoptimized, development program version on a single processor core, for ∼100 amino-acids protein, with several hundreds of considered lattice orientations; more detailed benchmarks are given in Supporting Information). This is significantly faster in comparison to more rigorous methods based on explicit solvent simulations, especially those utilizing the free energy perturbation protocol. For instance, simulations necessary to evaluate water binding to *a single* hydration site performed in this study take an order of magnitude more time for identical protein structure, using an optimized and accelerated computer code. It is also worth noting that the methods based on free energy parturbation require prior indication of a putative hydration site, while the methods based on postprocessing of standard MD simulations typically rely on the assumption that solvent distribution can equilibrate from starting configuration in a nanoseconds time scale, which may not always be the case.

One of the application areas for the proposed model may be prediction or validation of internal solvent distribution during refinement of protein models or in preparation of protein structures for further computational studies, such as MD simulations or receptor ligand docking. In this respect, we demonstrated that the quality of predictions based on a single, rigid protein conformation to a large extent depends on its origin. Crystallographic structures, which inherently contain information on water's presence even if not explicitly resolved, provide for much better results, as solvent distribution readily fits into cavities already configured for water binding. In order to reach similar accuracy for still reasonable, extracted from MD protein structures, but for which information concerning internal water was erased, several tens of conformations need to be considered. This latter scenario may apply to protein models obtained by *de novo* or homology modeling.

In the present work, the model was tested against a set of buried protein sites; however, a question may arise whether its predictions are also valid for wide open cavities such as ligand binding sites or tightly bound, specific water molecules at the protein surface. Such cases belong to the particularly challenging borderline between localized and bulk hydration effects. Here, the task of partitioning increasingly large, continuous solvent regions into discrete hydration sites and determining their binding free energies is increasingly error prone. Closed cavities within protein structures provide natural borders of unfavorable potential enclosing the region of local solvent partition function, with meaningful, favorable contributions to binding gathered at the center. No such borders exist in extended, open cavities or at the protein surface, where the results may become especially sensitive to the performance of a partitioning scheme.

Such problems are of particular relevance for protein−ligand systems, where accounting for contribution of hydration effects to binding requires considering both holo and apo (typically open) structures. Whereas predictions for buried, bridging water molecules in protein−ligand complexes should be readily available as long as their force-field parametrized structures are provided, proper evaluation of semistructured solvent within wide, empty binding cavities is challenging. The efforts to meet this challenge and fully utilize the capabilities of the proposed discrete solvation model for combining bulk and localized hydration effects are currently under way.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00839.

Notes on the parametrization of WCA potential, detailed data for hydration sites considered in the study, and discussion of the computational efficiency of the method (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: p.setny@cent.uw.edu.pl.

Funding

This work was supported by the Foundation for Polish Science grant Homing-Plus 2012-6/13.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Levy, Y.; Onuchic, J. N. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 389−415.
(2) Ball, P. *Chem. Rev.* **2008**, *108*, 74−108.
(3) Park, S.; Saven, J. G. *Proteins: Struct., Funct., Genet.* **2005**, *60*, 450−63.
(4) Rhodes, M. M.; Réblová, K.; Sponer, J.; Walter, N. G. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 13380−5.
(5) Angel, T. E.; Gupta, S.; Jastrzebska, B.; Palczewski, K.; Chance, M. R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 14367−14372.
(6) Wallnoefer, H. G.; Handschuh, S.; Liedl, K. R.; Fox, T. *J. Phys. Chem. B* **2010**, *114*, 7405−7412.
(7) Prakash, P.; Sayyed-Ahmad, A.; Gorfe, A. A. *PLoS Comput. Biol.* **2012**, *8*, e1002394.
(8) Lu, Y.; Wang, R.; Yang, C.-Y.; Wang, S. *J. Chem. Inf. Model.* **2007**, *47*, 668−75.
(9) Rodier, F.; Bahadur, R. P.; Chakrabarti, P.; Janin, J. *Proteins: Struct., Funct., Genet.* **2005**, *60*, 36−45.
(10) Ahmad, M.; Gu, W.; Geyer, T.; Helms, V. *Nat. Commun.* **2011**, *2*, 261.

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972

(11) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. *J. Am. Chem. Soc.* **2007**, *129*, 2577−87.
(12) Setny, P.; Geller, M. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 511−7.
(13) Luccarelli, J.; Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2010**, *6*, 3850−3856.
(14) Kellogg, G. E.; Fornabaio, M.; Spyrakis, F.; Lodola, A.; Cozzini, P.; Mozzarelli, A.; Abraham, D. J. *J. Mol. Graphics Modell.* **2004**, *22*, 479−86.
(15) de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E. *J. Med. Chem.* **2005**, *48*, 2308−18.
(16) Roberts, B. C.; Mancera, R. L. *J. Chem. Inf. Model.* **2008**, *48*, 397−408.
(17) Mancera, R. L. *Curr. Opin. Drug Discovery Devel.* **2007**, *10*, 275−80.
(18) García-Sosa, A. T. *J. Chem. Inf. Model.* **2013**, *53*, 1388−1405.
(19) Szep, S.; Park, S.; Boder, E. T.; Van Duyne, G. D.; Saven, J. G. *Proteins: Struct., Funct., Genet.* **2009**, *74*, 603−11.
(20) Davis, A. M.; Teague, S. J.; Kleywegt, G. J. *Angew. Chem., Int. Ed.* **2003**, *42*, 2718−36.
(21) Otting, G.; Liepinsh, E. *Acc. Chem. Res.* **1995**, *28*, 171−177.
(22) Boström, J.; Hogner, A.; Schmitt, S. *J. Med. Chem.* **2006**, *49*, 6716−25.
(23) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. a. *J. Mol. Biol.* **1997**, *265*, 445−64.
(24) García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. *J. Mol. Model.* **2003**, *9*, 172−82.
(25) Amadasi, A.; Surface, J. A.; Spyrakis, F.; Cozzini, P.; Mozzarelli, A.; Kellogg, G. E. *J. Med. Chem.* **2008**, *51*, 1063−7.
(26) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849−857.
(27) Miranker, A.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 29−34.
(28) Zhang, L.; Hermans, J. *Proteins: Struct., Funct., Genet.* **1996**, *24*, 433−8.
(29) *SZMAP*, version 1.1.1; OpenEye Scientifc Software, Inc.: Santa Fe, NM, 2013.
(30) Schymkowitz, J. W. H.; Rousseau, F.; Martins, I. C.; Ferkinghoff-Borg, J.; Stricher, F.; Serrano, L. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10147−52.
(31) Jiang, L.; Kuhlman, B.; Kortemme, T.; Baker, D. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 893−904.
(32) Rossato, G.; Ernst, B.; Vedani, A.; Smiesko, M. *J. Chem. Inf. Model.* **2011**, *51*, 1867−81.
(33) Ross, G. a.; Morris, G. M.; Biggin, P. C. *PLoS One* **2012**, *7*, e32036.
(34) Rarey, M.; Kramer, B.; Lengauer, T. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 17−28.
(35) Schnecke, V.; Kuhn, L. A. *Perspect. Drug Discovery Des.* **2000**, *20*, 171−190.
(36) Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. *J. Med. Chem.* **2005**, *48*, 6504−15.
(37) Österberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 34−40.
(38) Corbeil, C. R.; Englebienne, P.; Moitessier, N. J. *J. Chem. Inf. Model.* **2007**, *47*, 435−49.
(39) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. a.; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006**, *49*, 6177−96.
(40) Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. *Protein Sci.* **1994**, *3*, 1224−35.
(41) Adams, J. D. *Mol. Phys.* **1975**, *29*, 307−311.
(42) Woo, H.-J.; Dinner, A. R.; Roux, B. *J. Chem. Phys.* **2004**, *121*, 6392−400.
(43) Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2009**, *113*, 13337−46.
(44) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047−69.
(45) Lazaridis, T. *J. Phys. Chem. B* **1998**, *102*, 3531−3541.
(46) Lazaridis, T. *J. Phys. Chem. B* **1998**, *102*, 3542−3550.
(47) Li, Z.; Lazaridis, T. *J. Phys. Chem. B* **2006**, *110*, 1464−75.
(48) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 808−13.
(49) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817−31.
(50) Abel, R.; Wang, L.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory Comput.* **2010**, *6*, 2924−2934.
(51) Snyder, P. W.; Mecinovic, J.; Moustakas, D. T.; Thomas, S. W.; Harder, M.; Mack, E. T.; Lockett, M. R.; Héroux, A.; Sherman, W.; Whitesides, G. M. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 17889−94.
(52) Henchman, R. H. *J. Chem. Phys.* **2007**, *126*, 064504.
(53) Li, Z.; Lazaridis, T. *Methods Mol. Biol.* **2012**, *819*, 393−404.
(54) Nguyen, C. N.; Young, T. K.; Gilson, M. K. *J. Chem. Phys.* **2012**, *137*, 044101.
(55) Cui, G.; Swails, J. M.; Manas, E. S. *J. Chem. Theory Comput.* **2013**, *9*, 5539.
(56) Michel, J.; Henchman, R. H.; Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Law, R. J. *J. Chem. Theory Comput.* **2014**, *10*, 4055−4068.
(57) Gerogiokas, G.; Southey, M. W. Y.; Mazanetz, M. P.; Hefeitz, a.; Bodkin, M.; Law, R. J.; Michel, J. *Phys. Chem. Chem. Phys.* **2015**, *17*, 8416−8426.
(58) Chandler, D. *J. Chem. Phys.* **1972**, *57*, 1930.
(59) Beglov, D.; Roux, B. *J. Chem. Phys.* **1996**, *104*, 8678−8689.
(60) Kovalenko, A.; Hirata, F. *Chem. Phys. Lett.* **1998**, *290*, 237−244.
(61) Imai, T.; Hiraoka, R.; Kovalenko, A.; Hirata, F. *Proteins: Struct., Funct., Genet.* **2007**, *66*, 804−813.
(62) Yoshida, N.; Imai, T.; Phongphanphanee, S.; Kovalenko, A.; Hirata, F. *J. Phys. Chem. B* **2009**, *113*, 873−86.
(63) Kovalenko, A.; Kobryn, A. E.; Gusarov, S.; Lyubimova, O.; Liu, X.; Blinov, N.; Yoshida, M. *Soft Matter* **2012**, *8*, 1508.
(64) Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C. L.; Case, D. a.; Tuszynski, J.; Kovalenko, A. *J. Chem. Theory Comput.* **2010**, *6*, 607−624.
(65) Mobley, D. L.; Wymer, K. L.; Lim, N. M.; Guthrie, J. P. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 135−50.
(66) Setny, P.; Zacharias, M. *J. Phys. Chem. B* **2010**, *114*, 8667−75.
(67) Setny, P. *J. Phys. Chem. B* **2015**, *119*, 5970−5978.
(68) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(69) Kumar, R.; Schmidt, J. R.; Skinner, J. J. *J. Chem. Phys.* **2007**, *126*, 204107.
(70) Weeks, J. D.; Chandler, D.; Andersen, H. *J. Chem. Phys.* **1971**, *54*, 5237−5247.
(71) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.
(72) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221−234.
(73) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47.
(74) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; Van Der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845−854.
(75) Petersen, D. P.; Middleton, D. *Information and Control* **1962**, *5*, 279.
(76) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M.; Biophysique, L. D. C.; Uni, V.; Pasteur, L. *J. Phys. Chem. B* **2003**, *107*, 9535−9551.
(77) Roux, B.; Nina, M.; Pomès, R.; Smith, J. C. *Biophys. J.* **1996**, *71*, 670−81.
(78) Hess, B.; van der Vegt, N. F. a. *J. Phys. Chem. B* **2006**, *110*, 17616−26.
(79) Beauchamp, K. a.; Lin, Y.-S.; Das, R.; Pande, V. S. *J. Chem. Theory Comput.* **2012**, *8*, 1409−1414.
(80) Widom, B. *J. Chem. Phys.* **1963**, *39*, 2808.

DOI: 10.1021/acs.jctc.5b00839
J. Chem. Theory Comput. 2015, 11, 5961−5972