Rapid estimation of hydration thermodynamics of macromolecular regions

E. Prabhu Raman^{a)} and Alexander D. MacKerell, Jr.^{b)}

*Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy,
20 Penn Street HSF II, Baltimore, Maryland 21201, USA*

(Received 13 May 2013; accepted 16 July 2013; published online 7 August 2013)

This work describes a novel protocol to efficiently calculate the local free energy of hydration of specific regions in macromolecules. The method employs Monte Carlo simulations in the grand canonical ensemble to generate water configurations in a selected spherical region in the macromolecule. Excess energy and entropy of hydration are calculated by analyzing the water configurational distributions following the recently published grid inhomogeneous solvation theory method [C. N. Nguyen, T. K. Young, and M. K. Gilson, *J. Chem. Phys.* **137**, 044101 (2012)]. Our method involves the approximations of treating the macromolecule and distant solvent as rigid and performing calculations on multiple such conformations to account for conformational diversity. These approximations are tested against water configurations obtained from a molecular dynamics simulation. The method is validated by predicting the number and location of water molecules in 5 pockets in the protein Interleukin-1 β for which experimental water occupancy data are available. Free energy values are validated against decoupling free energy perturbation calculations. The results indicate that the approximations used in the method enable efficient prediction of free energies of water displacement.

© 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4817344>]

I. INTRODUCTION

Water plays a fundamental role in molecular recognition. Therefore, computational methods seeking to accurately predict macromolecule-ligand binding affinities need to account for contributions arising from water displacement and reorganization upon binding. The importance of water molecules was highlighted by an analysis of protein-ligand complex crystal structures that revealed an average of 4.6 interfacial water molecules per complex.² To understand the contribution of such waters in protein-ligand interactions, it is necessary to use methods that include explicit water molecules that can capture the molecular nature of water and the directionality of hydrogen bonds.

Physically rigorous alchemical transformation methods such as free energy perturbation (FEP) and thermodynamic integration (TI), when applied to calculating ligand-macromolecule affinities, include contributions from explicit water reorganization. However, it is a challenge to efficiently capture the change in water configurational ensemble as a result of the changing coupling parameter in these calculations, especially in regions of the macromolecule that get sequestered from bulk solvent as a result of a buried pocket or binding of a ligand.³ Second, alchemical transformation methods are computationally very expensive due to the simulation of the intermediate states and therefore cannot be employed in a high-throughput setting, at least with the current state of computational capabilities. FEP methods can also be used for calculating the hydration free energies of ligand

free (apo) macromolecule by decoupling the water molecules themselves. But this approach is problematic due to, for example, the diffusion of the perturbed water molecule(s) and long convergence times.

Inhomogeneous fluid solvation theory (IST)⁴ based methods have been used to calculate the free energy of solvent displacement and reorganization. The method relies on generating a configurational distribution of water, typically by performing molecular dynamics (MD) simulations, and analyzing the resultant ensemble to calculate the energy and entropy associated with water molecules. IST based methods have previously been employed to provide valuable insights into the thermodynamic contributions of water molecules in protein pockets and protein-ligand interfaces^{5,6} and used to semi-quantitatively guide drug design.⁷ The recently developed grid inhomogeneous solvation theory (GIST¹) method discretizes the IST equations, which enables the calculation of hydration free energy of arbitrary volumes. However, this approach has not been validated quantitatively against free energy perturbation calculations. Second, if the solvent displacement free energy is to be calculated for a large set of ligands that could potentially bind to a given site on a macromolecule, each with a different equilibrium configurational distribution of water, an MD sampling is likely to require long simulation times.

Keeping these issues in mind, in this article we introduce a grid-GCMC (grand canonical Monte Carlo) protocol for computing water displacement and reorganization free energies at selected sites of an arbitrary macromolecule. The configurational ensemble of water molecules in the subregion of the macromolecule is generated using GCMC simulations, as done in many applications.^{8–11} The method

^{a)}Electronic mail: prabhu@outerbanks.umaryland.edu

^{b)}Electronic mail: alex@outerbanks.umaryland.edu

uses several approximations in the system description: (1) The macromolecule is treated as rigid and represented by a pre-computed potential grid (hence, “grid”-GCMC). Macromolecule flexibility is only modeled by using multiple rigid conformations in this initial implementation. (2) Solvent molecules distant from the binding site (>12 Å) are not moved during the simulations. The resultant ensemble is analyzed to calculate the hydration energy of the protein pocket in a straightforward manner as a trajectory average, and hydration entropy is calculated using the GIST¹ formulation.

To test the method, we calculated the hydration free energy of five buried pockets in the protein Interleukin-1 β (IL-1 β). Crystallographically resolved water molecules in a high resolution structure of this protein helped validate the generation of water configurational distribution using our grid-GCMC method. To validate the approximations of rigid protein and distant rigid solvent representation, an alternate water configurational ensemble was obtained using an MD simulation of the protein in explicit water and analyzed using the same approach. To validate the free energy values we performed decoupling FEP calculations and compared the resultant free energies to those obtained by GIST based analysis of the grid-GCMC and MD ensemble. Such comparisons can be problematic due to the diffusion of bulk water into the region of decoupling. For this reason, occluded pockets in the protein were chosen so as to minimize this effect. Sections II–IV below describe the methodology, the results obtained using different calculation schemes, and a discussion of limitations, utility, and future directions.

II. METHODS

In this section we first describe the details of the MD simulation used to obtain the water conformations. We then describe our GCMC simulation setup and protocol to obtain another set of water configurational ensemble, which is compared structurally and thermodynamically to the MD ensemble. All molecular mechanics calculations used the CHARMM22¹² force-field with CMAP backbone correction¹³ for the protein and the TIP3P model¹⁴ to represent water, which was treated as a rigid body. The crystal structure of IL-1 β solved by Quillin *et al.* (Protein Data Bank (PDB) 2NVH¹⁵) was used for the calculations after removing the sulphate ion.

A. Molecular dynamics simulations

The GROMACS¹⁶ package (version 4.5) was used for all MD simulations. The crystal structure of the protein was solvated in a box of dimensions 52 Å $\times$ 55 Å $\times$ 58 Å. All 149 crystallographic water molecules were retained and non-crystallographic waters with overlapping van der Waals radii with protein atoms were deleted. One sodium ion was added to make the system charge neutral. The five protein pockets were manually inspected to ensure the absence of non-crystallographic water molecules. All protein heavy atoms were restrained with a force constant of 0.5 kcal/mol Å⁻² to facilitate convergence and to allow comparison of results

between the MD simulations and FEP calculations (see below). The equilibration and production MD simulations were performed under periodic boundary conditions. van der Waals (vdW) interactions were switched off smoothly in the range of 8 – 9 Å and the particle mesh Ewald method was used to treat long range electrostatics with a real space cut off of 10 Å. Long-range dispersion correction to the energy and pressure was applied. MD simulations were performed with the leap-frog stochastic dynamics integrator (GROMACS integrator “sd”) with an inverse friction constant of 1 ps. Pressure was maintained using the Parrinello-Rahman barostat. The LINCS algorithm was used to constrain all bonds involving a hydrogen atom. Following a short minimization, the system was equilibrated to 298 K temperature and 1 atm pressure over 1.1 ns, following which a 20 ns production simulation was performed.

B. System setup for GCMC simulations

The GROMACS¹⁶ package was used for structure preparation and a short energy minimization to remove any bad contacts, following which all water molecules were removed. The hydration thermodynamics computations were performed using an in-house suite of programs written in Fortran. Following minimization, the protein structure was treated as rigid and was represented by a grid, similar to docking approaches.¹⁷ In addition to the crystal conformation, 10 additional conformations were chosen for the GCMC simulations (one conformation output every 2 ns) from the 20 ns MD simulation described above. System preparation involved the following steps for each pocket: (i) A pre-equilibrated water sphere of radius 20 Å centered at the pocket center was overlaid on the protein structure with waters overlapping with the protein deleted. (ii) All water molecules added that lie within 12 Å of the pocket center were deleted. This resulted in an 8 Å shell of water between 12 and 20 Å from the pocket center. (iii) A grid centered at the pocket center with cubic dimensions and 20 Å edge length was constructed. The grid represents the potential energy at discrete points in space. One grid for electrostatic energy and two grids for water oxygen and hydrogen vdW types were calculated with a resolution of 0.2 Å. All atoms (protein and the static water shell) were considered for the electrostatic grid energy computation. Only atoms within 14 Å of a grid point were considered for the vdW grid energy computation as the potential is much shorter in range. The grid based interaction energy evaluation requires a one time investment to compute the energy grid and provided a nearly 70 -fold speedup compared to explicit pairwise energy evaluation.

During the Monte Carlo (MC) simulations, the non-bonded interaction energy of a water atom with the protein environment was computed by trilinear interpolation of energy values of the surrounding 8 grid points. The grid energy included the contributions from protein atoms and the static water shell. Figure 1 shows a schematic two-dimensional diagram of the system setup. The pocket center coincides with the center of the concentric spheres. The outermost layer consists of constrained water molecules (white filled circles).

FIG. 1. A two-dimensional schematic of the system setup used in the grid-GCMC calculations. The three regions Site, Buffer, and Static are demarcated as circles, and the protein is depicted in green. The dark blue circles occupying the Site and Buffer represent water molecules sampled during the GCMC simulations. The white colored circles represent constrained water molecules, which are not moved during the simulations.

Water molecules in the middle “Buffer” layer and the innermost “Site” region (dark blue circles) are treated indistinguishably during the GCMC simulations. However, statistics are collected only for the Site waters and therefore the free energy computed is the hydration free energy of the Site only. In this study, we have set the radius of the Static, Buffer, and Site regions to be 20, 10, and 5 Å, respectively, unless indicated otherwise.

C. Water GCMC sampling in the macromolecule grid

The simulation setup described above results in a Site+Buffer region centered on the pocket in which no water molecules are present initially. GCMC simulations were employed to sample the water configurations in this region. The GCMC probabilities for insertion and deletion acceptance have been derived in many publications.^{18,19} For completeness, we provide them here:

$$P_{n \rightarrow n+1} = \min \left\{ 1, \frac{e^{B-\beta\Delta E}}{n+1} \right\}, \quad (1)$$

$$P_{n \rightarrow n-1} = \min \{ 1, n e^{-B-\beta\Delta E} \}, \quad (2)$$

where ΔE is the energy change upon insertion or deletion of a water molecule. β is the inverse of the product of the Boltzmann constant and the absolute temperature (298 K). n is the number of GCMC water molecules interacting with the macromolecular system. $B(\equiv \beta\mu_{ex} + \ln(\rho_o V_{SB}))$ is related to the excess chemical potential of water μ_{ex} ($=-6$ kcal/mol) and its bulk density $\rho_o(=0.0334 \text{ Å}^{-3})$. V_{SB} is the volume of Site+Buffer region in which insertions and deletions are attempted.

For the translational and rotational moves of water molecules, the standard Metropolis criterion provides the acceptance probability:

$$P_{i \rightarrow j} = \min \{ 1, e^{-\beta\Delta E} \}. \quad (3)$$

The probabilities of water insertion and deletion moves were 0.2 and 0.2, respectively. The probability of a translation or rotation move was 0.6. For translation attempts, the selected water molecule was translated by a distance randomly chosen in the interval $(-0.2 \text{ Å}, 0.2 \text{ Å})$ along one of the 3 randomly chosen orthogonal axes. For rotation, the selected water molecule was rotated by an angle randomly selected in the interval $(-180^\circ, 180^\circ)$ along one of the 3 randomly selected Euler angles. During the simulations, water configurations in the Site were stored every 1000 steps for free energy calculations. The corresponding energetic terms were also saved, which included: (i) interaction energies of Site waters with the Buffer and the Static region, and (ii) interaction energy of water molecules within the Site.

D. Water binding free energy

Water configurations and energies for the site of interest from the GCMC or MD simulations were analyzed using the following approach to compute the thermodynamic properties. The free energy of hydrating the Site is decomposed into the excess energetic and entropic terms as follows:

$$\Delta G_{Site}^{hyd} = \Delta E_{Site} - T(\Delta S_{Site}^{trans} + \Delta S_{Site}^{orie}), \quad (4)$$

where, ΔS_{Site}^{trans} and ΔS_{Site}^{orie} refer to the excess translational and orientational entropies of the water molecules in the Site. The excess energy is a sum of pairwise electrostatic and vdW energy terms involving Site waters:

$$\Delta E_{Site} = \left\langle \sum_{\text{waters } i} E(i, env) + \frac{1}{2} \sum_{\text{waters } i} \sum_{\text{waters } j \neq i} E(i, j) - \sum_{\text{waters } i} E_{bulk} \right\rangle, \quad (5)$$

where, the average is over the simulation snapshots and the sum is only over the Site water molecules. $E(i, env)$ is the non-bonded interaction energy between water i and the environment (entire system, except the Site), and $E(i, j)$ is the interaction energy between waters i and j in the Site. The second term in Eq. (5) is the self energy of the waters in the Site and the prefactor corrects for the double counting of interactions. The last term discounts the energy of bulk water so as to make ΔE_{Site} an excess energy. A value of -9.9 kcal/mol was used for E_{bulk} . This was computed as half of the average interaction energy of a randomly chosen water molecule obtained from a 5 ns simulation of a cubic box consisting of 2483 water molecules. To be consistent with the grid-GCMC protocol, a real-space cutoff of 20 Å was used for non-bonded interactions in the energy evaluation. This value is also very close to previously reported ones in the literature.^{14,20} For the analysis of grid-GCMC simulations, the energy is output during the simulation. For MD trajectories, we calculated the energy

using a post-processing routine including all image transformations up to the neighboring unit cell without a cutoff.

The excess translational entropy is integrated over the grid spanning the system, per the GIST formulation.¹ The grid was composed of cubic volume elements of edge length 0.5 Å. Relative probabilities $p(v)$, as a function of voxel v , are obtained by discretizing the water oxygen coordinates output from the simulations and dividing the time averaged occupancy by the analogous value in bulk, ρ_o ($=0.0334 \text{ Å}^{-3}$). The following summation describes the translational entropy computation:

$$\Delta S_{Site}^{trans} = -k_B \rho_o V_{vox} \sum_{v \in Site} p(v) \ln p(v), \quad (6)$$

where, V_{vox} is the volume of each voxel element, and the sum is over the voxels that belong to the Site. Similarly, the orientational entropy is integrated over the grid, but weighted by the voxel probability. The excess orientational entropy, ΔS_{Site}^{orie} , is given by the following:¹

$$\Delta S_{Site}^{orie} = \rho_o V_{vox} \sum_{v \in Site} p(v) S^{orie}(v), \quad (7)$$

where, the orientational entropy $S^{orie}(v)$ associated with a voxel v can be computed either through the histogram method ($S_h^{orie}(v)$) or the nearest neighbor method ($S_{nn}^{orie}(v)$) as shown below:¹

$$S_h^{orie}(v) = -k_B \frac{\Delta\omega}{8\pi^2} \sum_{\text{Euler } \omega} p(\omega|v) \ln p(\omega|v), \quad (8)$$

where the sum is over uniformly spaced bins of the 3 Euler angles and ω is the three-dimensional Euler angle. The constant $\Delta\omega$ is the 3D Euler angle bin width and $p(\omega|v)$ is the conditional probability of observing the Euler angle ω for a water molecule located in voxel v divided by $\frac{\Delta\omega}{8\pi^2}$. Orientational entropy via the nearest neighbor method is calculated using

$$S_{nn}^{orie}(v) = -\frac{k_B}{N_v} \left[-\gamma + \sum_{i=1}^{N_v} \ln \frac{6\pi}{N_v (\Delta\omega_i)^3} \right], \quad (9)$$

where N_v is the number of observed water molecules in voxel v , γ is the Euler constant, and $\Delta\omega_i$ is the Euler angular distance of a water molecule i in the current voxel with its nearest neighbor conformation in the same voxel.¹

For the GCMC simulations in the grid environment, where multiple simulations were performed with $N_{conf}(=11)$ conformations, $\Delta G_{Site}^{hyd}(i)$ was computed separately for each conformation as per Eq. (4), and a Boltzmann average was performed to obtain $\langle \Delta G_{Site}^{hyd} \rangle$ as the final hydration free energy of each Site:

$$\langle \Delta G_{Site}^{hyd} \rangle = -\frac{1}{\beta} \ln \frac{1}{N_{conf}} \sum_{i=1}^{N_{conf}} e^{-\beta \Delta G_{Site}^{hyd}(i)}. \quad (10)$$

A Boltzmann averaged number of water molecules in each Site was obtained as follows:

$$\langle n_{calc} \rangle = \frac{\sum_{i=1}^{N_{conf}} n_i e^{-\beta \Delta G_{Site}^{hyd}(i)}}{\sum_{i=1}^{N_{conf}} e^{-\beta \Delta G_{Site}^{hyd}(i)}}. \quad (11)$$

E. Free energy perturbation calculations

Decoupling free energy perturbation calculations were performed to validate the free energies computed using the GIST approach. Following the method formulated by Gilson *et al.*²¹ and used by Hamelberg and McCammon,²² the standard free energy of a stable water molecule binding to a site on the protein is decomposed into the negative of the decoupling free energy, the translational contribution, and the hydration free energy of water, as shown in the following equation. A contribution ($=RT \ln(2)$) arising due to the symmetry number of water was not included.²³

$$\Delta G_{Site}^{FEP} = -\Delta G^{decouple} - n_w RT \ln(C^o V_1) - n_w \Delta G^{hyd}, \quad (12)$$

where, n_w is the number of waters being decoupled simultaneously in the given Site. Instead of using a standard concentration C^o of 1 M as done by Hamelberg and McCammon,²² we used a value of 55 M so as to make comparison with our GCMC approach where excess free energy is calculated with respect to bulk water under standard conditions. The water molecules in the decoupling FEP simulations were harmonically restrained with a force-constant k of $0.5 \text{ kcal mol}^{-1} \text{ Å}^{-2}$ and the effective volume V_1 , which the water molecule is restrained to occupy was calculated as shown in Eq. (13).²² The value of k is nearly 10 times lower than $\frac{3RT}{\langle \delta r^2 \rangle}$ for all water molecules in the pockets and therefore does not significantly perturb the natural dynamics expected in absence of the restraint:²²

$$V_1 = \left(\frac{2\pi RT}{k} \right)^{3/2}. \quad (13)$$

The decoupling FEP simulations performed to compute $\Delta G^{decouple}$ involved the same dynamics parameters as used in the MD simulations, and were performed separately to decouple the water molecules in each of the four occupied pockets of the protein. LINCS constraints were not applied to the water molecules being perturbed due to technical reasons. For pockets 1 and 2, the two water molecules were decoupled simultaneously during the simulations. In the fully decoupled state, the two water molecules do not feel each other. The simulations involved 11 λ -states to turn-off the charges in which the vdW interactions were fully active. The vdW interactions were decoupled with “soft-core” scaling using 21 λ -states with zero charge on the water molecules. Each independent simulation involved a short minimization, 100 ps equilibration and 2 ns production stages. The results were analyzed using the Bennett acceptance ratio method (GROMACS utility `g_bar`) to yield the $\Delta G^{decouple}$ values.

III. RESULTS

Localized hydration thermodynamics calculations were performed in the five pockets of the protein IL-1 β . In this section, we will first demonstrate the ability of the grid-GCMC sampling to reproduce the crystal locations of water molecules in the five pockets.¹⁵ All the pockets are occluded in the crystal structure (Figure S1 in the supplementary material).²⁷ The sampling will then be compared to that

FIG. 2. Water probability contours in the four pockets of IL-1 β at 10 times bulk density. Panels (a), (c), (e), and (g) show contours obtained from the grid-GCMC simulations with each trajectory in a different color. Panels (b), (d), (f), and (h) show the corresponding contour obtained from the analysis of the MD simulation. Crystal conformation of waters is depicted by the red spheres.

obtained using MD simulations. While the MD simulation was performed in the NPT ensemble, where water insertions or deletions were not permitted, the locations of water molecules within the Site serve as a benchmark. Second, we computed the thermodynamic properties from the sampling obtained from the grid-GCMC and MD simulations. The resultant free energy values were then compared with decoupling FEP calculations.

A. Reproducing crystal locations of water molecules

GCMC simulations were performed to obtain the configurational ensemble of water molecules in the site of interest. For each of the 5 buried pockets in IL-1 β which have been analyzed crystallographically,¹⁵ the center of the site was chosen as the center of mass of the water molecules in the crystal conformation. For Site 5 in which no water molecules were observed experimentally, the Site center was chosen as the center of mass of the side-chains lining the pocket. A sphere of 5 Å radius centered on the pocket center was demarcated as the “Site” region for which the excess hydration free energy was computed (Fig. 1). An exception to this rule was pocket 2, where the Site radius was set to 4 Å instead. This was necessary because a radius of 5 Å included an additional water molecule outside the pocket making it difficult to compare with the FEP method.

To model conformational flexibility of the protein, we performed the grid-GCMC calculations on multiple conformations. In addition to the crystal conformation of IL-1 β , 10 conformations were chosen from the 20 ns explicit water MD simulation, equally spaced in time. In addition to providing the 10 additional protein conformations for the grid-GCMC calculations, the water distribution obtained from the MD simulation was analyzed using the GIST approach and compared to the analysis of the grid-GCMC ensemble. During the MD simulation, all protein heavy atoms were harmonically restrained using a force constant of 0.5 kcal mol⁻¹ Å⁻². In the grid-GCMC approach, for each of the 11 conformations, the system was prepared for performing the GCMC

simulations as detailed in Sec. II. 500 million Monte Carlo steps (MCS) resulted in water configurations that filled the Site and Buffer regions. The panels (a), (c), (e), and (g) in Figure 2 show the water probability contours obtained from the grid-GCMC approach applied to the 11 conformations for the pockets 1, 2, 3, and 4. The panels (b), (d), (f), and (h) show the corresponding contours obtained from the 20 ns MD simulation. The contours in all cases were prepared at 10 times the bulk density of water. The crystal conformations of the waters are depicted by red spheres. The figure first shows that the grid-GCMC approach correctly locates the numbers and positions of the crystal waters. Second, the densities obtained from the grid-GCMC approach are in agreement with the MD densities. These observations validate the approximations in the grid-GCMC method for water ensemble generation.

B. Binding free energies

Using the configurational distribution of water molecules obtained from the grid-GCMC approach, excess free energies of water molecules in the Site region were computed using Eq. (4). For each of the 11 protein conformations used in the grid-GCMC calculations, Eq. (4) was applied separately to calculate 11 hydration free energy values for each of the 5 pockets. A Boltzmann average of the 11 values as per Eq. (10) was performed to obtain the excess hydration free energy of each pocket. The excess energy of the water molecules in the Site, ΔE_{Site} , was computed by subtracting the bulk energy of the corresponding number of water molecules from the energy in the Site and obtaining a trajectory average, as per Eq. (5). Following the GIST approach,¹ the excess translational entropy of water molecules was computed according to Eq. (6) by discretizing the water configurations into a 3D grid composed of voxels of 0.5 Å edge length. The excess orientational entropy (Eq. (7)) was calculated using both the histogram method (Eq. (8)) and the nearest-neighbor method (Eq. (9)). For the histogram method, two different Euler angle bin sizes were tested: 36° and 18°. Both showed good

TABLE I. Thermodynamic properties computed using the conformations obtained using (a) grid-GCMC and (b) MD sampling. (a) displays the lower and upper bounds of thermodynamic properties of water computed for 11 protein conformations in the 5 cavities of IL-1 β . (b) displays the properties obtained from the single MD trajectory. All energies are in kcal/mol.

Site	n_{Xtal}	n_{calc}		ΔE		$-T\Delta S^{trans}$		$-T\Delta S_{h36}^{orie}$		$-T\Delta S_{h18}^{orie}$		$-T\Delta S_{nn}^{orie}$	
(a) Grid-GCMC													
1	2	2.00	2.00	-18.33	-13.04	4.31	5.28	4.69	5.13	5.25	5.65	6.25	6.72
2	2	2.00	2.05	-16.13	-11.21	4.39	5.15	4.57	5.29	5.14	6.42	6.12	13.86
3	1	1.00	1.00	-10.68	-7.30	2.02	2.46	2.29	2.49	2.58	2.82	3.09	3.39
4	1	1.00	1.00	-11.67	-3.59	2.04	2.53	1.55	2.29	1.76	2.54	2.18	3.00
5	0	0.04	0.63	0.16	1.10	-0.07	0.47	0.04	0.67	0.07	0.81	0.03	0.91
(b) MD													
1	2	2.00		-16.42		3.71		4.51		5.20		5.87	
2	2	2.01		-15.10		3.51		4.53		5.29		5.94	
3	1	1.00		-8.89		1.80		2.36		2.76		3.16	
4	1	1.00		-11.70		1.69		2.13		2.48		2.73	

convergence properties, but the latter is shown to provide higher accuracy, as expected.

To validate the free energy values quantitatively, we performed double decoupling free energy perturbation simulations of water molecules in pockets 1, 2, 3, and 4. The calculations were performed independently for each pocket and used the same dynamics parameters as in the MD simulation including the protein restraints. Each λ window involved 100 ps of equilibration and 2 ns of sampling. The convergence of the decoupling FEP calculations is monitored as a function of simulation time and plotted in Figure S2 in the supplementary material.²⁷ As the simulation time is increased from 1 ns to 2 ns, ΔG^{FEP} changes by less than 0.5 kcal/mol for pockets 1, 3, and 4, whereas the difference for pocket 2 is 0.7 kcal/mol. With pocket 2, during the FEP simulations, another water molecule diffused into the cavity at $\lambda = 1$ when the crystal waters were fully decoupled. Even though this pocket is occluded, it is much closer to the surface than the other 3 pockets. To prevent other water molecules from occupying the empty cavity at $\lambda = 1$, a repulsive particle was placed and held fixed at the center of pocket 2 with vdW parameters $R_{min} = 5.25$ Å and $\epsilon = 0.025$ kcal/mol. These parameters lead to a repulsive energy of 0.3 kcal/mol at a distance of 4 Å from the pocket center and zero at a distance of 4.25 Å. This choice was motivated by the fact that the Site was demarcated in the previous analysis to a sphere of radius 4 Å. Figure S3 in the supplementary material²⁷ shows an analysis of the FEP trajectories where the introduction of the repulsive restraint removes the density of unperturbed water molecules from the pocket and yet does not alter their density in the surrounding region. The inclusion of this additional restraint makes comparison with FEP calculations problematic because they do not account for the free energy of cavity formation. This point is explained in Sec. IV.

Table I(a) shows results from the analysis of the grid-GCMC simulations performed using the 11 protein conformations in the 5 pockets. The lower and upper bounds, among the 11 conformations, of the number of water molecules in the Site, the excess energy, the excess translational entropies, and the excess orientational entropies computed using the three approaches are shown. The numbers of calculated water

molecules using the grid-GCMC approach n_{calc} for pockets 1, 2, 3, and 4 closely match the experimentally observed number n_{Xtal} . For pocket 5, the maximum occupancy is 0.63, but the associated free energy is unfavorable. More importantly, the thermodynamically averaged number of waters $\langle n_{calc} \rangle$ (Table II(a)) is 0.09. The negative free energy of the first four pockets is consistent with the presence of water molecules in the pockets in the crystal structure. These data show that the grid-GCMC approach correctly predicts the water occupancy. As the angular bin width is decreased from 36° to 18° , the orientational entropy loss in the protein pocket becomes more pronounced. This trend is further observed in the ΔS_{nn}^{orie} results, which does not use a constant bin width. For pocket 2, for one of the conformations it reaches a high value of 13.86 kcal/mol. Such an unusually high value suggests a limitation in calculations performed with a constrained macromolecule conformation as this was not observed for the MD ensemble. This is explained in Sec. IV.

Table II(a) shows the thermodynamically averaged number of water molecules observed in each Site and three different calculations of excess hydration free energies that differ in the orientational entropy method. The average number of

TABLE II. Excess free energy values of the pockets in IL-1 β computed by analyzing the (a) grid-GCMC and (b) MD ensemble. Three values corresponding to the different orientational entropy calculations are reported. All energies are in kcal/mol.

(a) Grid-GCMC						
Site	n_{Xtal}	$\langle n_{calc} \rangle$	$\langle \Delta G_{h36} \rangle$	$\langle \Delta G_{h18} \rangle$	$\langle \Delta G_{nn} \rangle$	ΔG^{FEP}
1	2	2.00	-7.90	-7.29	-6.26	-6.01
2	2	2.01	-5.17	-4.32	-3.18	-5.20
3	1	1.00	-4.91	-4.58	-4.04	-3.94
4	1	1.00	-6.35	-6.10	-5.64	-6.23
5	0	0.09	0.48	0.53	0.48	NA
(b) MD						
Site	n_{Xtal}	n_{MD}	ΔG_{h36}	ΔG_{h18}	ΔG_{nn}	ΔG^{FEP}
1	2	2.00	-8.20	-7.52	-6.84	-6.01
2	2	2.01	-7.06	-6.31	-5.65	-5.20
3	1	1.00	-4.73	-4.33	-3.93	-3.94
4	1	1.00	-7.89	-7.53	-7.28	-6.23

water molecules was calculated as per Eq. (11) using ΔG_{nn} as the thermodynamic weight. Using the other two free energy values associated with the 36° or 18° bin widths as weight did not change $\langle n_{calc} \rangle$ significantly and therefore those values are not reported. The free energies become progressively less favorable in the order $\langle \Delta G_{h36} \rangle$, $\langle \Delta G_{h18} \rangle$, $\langle \Delta G_{nn} \rangle$. This is due to the increasing entropy loss discussed above. Figures S4 and S5 in the supplementary material²⁷ show the convergence of ΔG_{h18} and ΔG_{nn} , respectively. For most trajectories the values converge within 10^8 MCS suggesting that shorter trajectories might be sufficient. Two trajectories for pocket 2 in Figure S5 in the supplementary material²⁷ overestimate the orientational entropy loss calculated using the nearest neighbor method as judged by the MD estimate.

Table II(a) also shows the hydration free energy computed using decoupling FEP method. Most of the grid-GCMC $\langle \Delta G \rangle$ values are within 1 kcal/mol absolute deviation from the ΔG^{FEP} values. For pocket 2, the deviations are much larger, with the largest one being that of $\langle \Delta G_{nn} \rangle$ at 2.02 kcal/mol. This points to a limitation in the grid-GCMC method as discussed below.

To test our approximation of using multiple protein conformations to model protein flexibility, we calculated the thermodynamic properties using the same analysis method, but with the water configurational distribution obtained from the 20 ns explicit water MD simulation, which was initiated from the crystal conformation. In some protein pockets, during the simulation, water molecules exchanged with the bulk. Since our analysis approach is volume-centric and not molecule-centric (such as in FEP calculations), this posed no problems. Table I(b) shows the energetic and entropic contributions to the free energies. Since Site 5 did not contain any water molecules, the analysis is not applicable. n_{calc} being close to the experimental value is trivial in this case because exchange with bulk is rare for these occluded pockets as in the GCMC simulations described above. The slight deviation from exact integer number in pocket 2 is due to a nearby water molecule entering the 4 Å sphere transiently. The excess energy values are in general agreement with the ranges seen in the grid-GCMC calculations shown in panel (a). A similar comparison with the translational entropy contribution shows that the entropy loss is less pronounced in the MD ensemble, which is expected owing to the flexibility. However, there is not as significant a difference in the orientational entropic contributions (except for pocket 2). Figures S6 and S7 in the supplementary material²⁷ show the convergence behavior of the thermodynamic properties computed using the histogram and nearest neighbor methods, respectively. For all pockets other than 2, the histogram method shows good convergence. For pocket 2, Figure S7 in the supplementary material²⁷ shows that the nearest neighbor estimator improves convergence.

Table II(b) shows the three values of free energies for each pocket. ΔG_{h18} and ΔG_{nn} are in much better agreement with ΔG^{FEP} . ΔG_{h36} is consistently too favorable owing to the underestimation of the orientational entropy loss which is due to the large bin width. Data from Tables II(a) and II(b) summarize the findings and are plotted in Figure 3. For the nearest neighbor based calculation, the average unsigned error (AUE) with respect to FEP for MD and grid-GCMC approaches are

FIG. 3. Excess binding free energy of water molecules in the 5 pockets of IL-1 β computed using the three approaches. GIST calculations were applied to calculate ΔG from the water configurational ensemble obtained from the grid-GCMC and MD simulations. Decoupling free energy perturbation calculations (FEP) results closely match the other two methods.

0.59 and 0.74 kcal/mol, respectively. For the histogram (h_{18}) approach, the AUE are 0.73 and 1.08 kcal/mol, respectively. The results thus suggest that approximating protein flexibility using multiple conformations still provides a reasonable agreement with a restrained protein simulation for the purposes of protein site hydration calculation.

IV. DISCUSSION

The presented grid-GCMC method generates a local configurational ensemble of water molecules within an arbitrarily chosen region in a macromolecule. The thermodynamic properties from the ensemble were calculated following the recently published GIST method.¹ Inhomogeneous solvation theory⁴ based methods have previously been employed to provide valuable insights into the thermodynamic contributions of water molecules in protein pockets and protein-ligand interfaces^{5,6} and used to semi-quantitatively guide drug design.⁷ The GIST method provided an approach to calculate the free energy associated with hydration of an arbitrary volume of space by discretizing the IST equations. The approach allows calculation of the energetic and entropic contributions made by all water molecules irrespective of their occupancy sampled in a given volume as opposed to previous IST based implementations which dealt with high occupancy sites of single water molecules independently. This study presents the combination of the grid-based evaluation of interaction energy, GCMC simulations for configurational sampling and the GIST method for the calculation of thermodynamics. The grid-GCMC simulation system was designed for efficiency while still capturing the dominant contributions to site hydration. GCMC simulations sampled the site of

interest and an enveloping buffer region. The configurational distribution of water in the buffer region is important due to the direct interaction with the Site and is reflected in the enthalpic term.

The GIST method includes only the solute-water entropic contributions and does not consider water-water correlations. Despite this approximation a good reproduction of FEP based free energies was obtained. The conformational distribution and thermodynamic properties computed using the grid-GCMC method were checked against values computed using MD sampling of the restrained protein in explicit water under periodic boundary conditions. Protein restraints were set to be $0.5 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ on heavy atoms, for both the MD simulation and the decoupling FEP simulations. This was mainly done to unambiguously compare the three calculations. The protein restraints in the FEP simulations limit the contributions from pocket relaxation upon the deletion of water molecules to the free energy. Since our analysis is performed only on the ensemble obtained from a hydrated pocket and does not include protein conformational contributions, the restraints facilitate a comparison of methodologies. By the same token, comparisons with other published computational studies of IL-1 β hydration involving a flexible treatment of the protein²⁴ cannot be made. The protein restraints help restrict the alchemical free energetic contributions to the interaction of the water molecules with the environment and the entropic terms associated with the waters being decoupled. For pocket 2, the inclusion of the restraint makes it no longer possible to compare directly the decoupling calculations with the GIST calculations due to the cavitation term not being captured in the former. In other pockets, this term is simply absent. The cavitation free energy of a TIP3P water molecule in bulk is approximately 3.9 kcal/mol ($\Delta G^{\text{hyd}} - \Delta E^{\text{interaction}} = -6 - (-9.9)$). However, in the protein cavity, it is expected to be much lesser due to the presence of a preformed cavity and for this reason, we still include this comparison in our results.

The good agreement in water conformational distribution and thermodynamic properties between the two approaches validates the approximations in the grid-GCMC sampling. These approximations include the use of constrained protein conformations, constrained distant solute conformations, and representing the static atoms as a potential energy grid of resolution $0.2 \text{ \AA}$. The use of multiple conformations was found to be crucial to get agreement with the FEP results. Table SI in the supplementary material²⁷ reports $\Delta G_{\text{Site}}^{\text{hyd}}$ computed for each of the 11 conformations separately. The variation in the calculated values across different conformations highlights the importance of using multiple conformations. The underlying assumption in using $\Delta G_{\text{Site}}^{\text{hyd}}$ as the Boltzmann weight is that the probabilities of the different protein conformations are very similar. In reality this will not be the case for an unrestrained simulation of a flexible pocket. Further development is needed for such cases that will need to account for the protein reorganization free energy as well.

Expectedly the free energies computed using the MD based sampling were in much better agreement with FEP results due to the inclusion of explicit protein flexibility (still limited by protein restraints). For MD results, the nearest-

neighbor orientational entropy estimator showed better agreement with FEP than the histogram based methods. For grid-GCMC, the agreement with FEP results was reasonable except for pocket 2. Two of the three calculation methods applied in this pocket underestimated the free energy. This was most pronounced in the nearest neighbor estimate. Two out of the 11 grid-GCMC trajectories in pocket 2 showed high orientational entropy contribution (Figure S5 in the supplementary material²⁷), suggesting orientationally restrained water geometries. Perhaps due to the large bin width in the histogram method, this feature gets masked (Figure S4 in the supplementary material²⁷). Since this is not observed in the MD estimate, the constrained protein geometries used in the grid-GCMC calculations are likely to be responsible for this difference. The use of rigid protein geometries has been of some utility in previous computational studies to detect fragment binding hot-spots²⁵ and calculate fragment binding free energies.²⁶ It is perhaps due to the small size of the ligands used as fragments, which results in predictive results despite this serious approximation. The small size of the water molecule likely allows for predictive results obtained in our case. However, problems can be encountered, as seen for pocket 2. Introducing protein flexibility through MC moves only on the local atoms and retaining the grid based representation of distant atoms are expected to alleviate this problem while at the same time maintaining efficiency of the calculation.

To our knowledge, GIST based free energy values have not been validated quantitatively against free energy perturbation calculations. Thus, this study represents a first such investigation. In future validation studies, test systems should include larger number of water molecules to check the validity of the one body water entropy computation and also investigate solvent exposed protein pockets. Second, in the current study, we do not take into account the effect of differing buffer response to water displacement. Ideally, one would perform a calculation of the hydration free energy of the Site+Buffer region and then the Buffer region alone, and take the difference to obtain the Site water displacement free energy to account for this effect. In future studies, especially dealing with solvent exposed pockets, such a strategy must be followed.

The grid-GCMC sampling consumed about 2–3 central processing unit (CPU) hours per 100 million MCS, which appears to be an adequate amount of sampling to obtain convergence as is apparent from the time evolution of calculated thermodynamic quantities. Investigation into the effect of snapshot output frequency and preferential treatment of Site waters (as opposed to Site+Buffer) for coordinate translation and rotation moves could result in further efficiency gains. Validation shown in this study suggests that the grid-GCMC method in combination with GIST can be of utility to rapidly identify thermodynamically important water molecules. The method may find use in the rapid identification of high affinity water molecules to be included in docking calculations. Perhaps, much shorter simulation times could prove adequate if only high occupancy waters are to be identified specific for each ligand, and precise free energy values are not desired. Importantly, the ability to rapidly calculate thermodynamics of hydration suggests that this method may find use in

calculating free energies of solvent displacement and reorganization for macromolecule-ligand complexes. Compared to MD based ensemble generation, the grid-GCMC approach is rapid, not problematic for buried pockets and therefore can be applied in high throughput settings involving macromolecule-ligand interaction interfaces.

ACKNOWLEDGMENTS

This work was supported by NIH Grant Nos. CA107331 and AI080968. The authors acknowledge computer time and resources from the Computer Aided Drug Design (CADD) Center at the University of Maryland, Baltimore.

- ¹C. N. Nguyen, T. K. Young, and M. K. Gilson, *J. Chem. Phys.* **137**, 044101 (2012).
- ²Y. Lu, R. Wang, C.-Y. Yang, and S. Wang, *J. Chem. Inf. Model.* **47**, 668 (2007).
- ³J. Michel, J. Tirado-Rives, and W. L. Jorgensen, *J. Am. Chem. Soc.* **131**, 15403 (2009).
- ⁴T. Lazaridis, *J. Phys. Chem. B* **102**, 3531 (1998).
- ⁵T. K. Young, R. Abel, B. Kim, B. J. Berne, and R. A. Friesner, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 808 (2007).
- ⁶D. J. Huggins, M. Marsh, and M. C. Payne, *J. Chem. Theory Comput.* **7**, 3514 (2011).
- ⁷R. Abel, T. K. Young, R. Farid, B. J. Berne, and R. A. Friesner, *J. Am. Chem. Soc.* **130**, 2817 (2008).
- ⁸M. Mezei, *Mol. Phys.* **61**, 565 (1987).
- ⁹H. Resat and M. Mezei, *J. Am. Chem. Soc.* **116**, 7451 (1994).
- ¹⁰H. Resat and M. Mezei, *Biophys. J.* **71**, 1179 (1996).
- ¹¹H. Resat, J. Tami, and J. A. McCammon, *Biophys. J.* **72**, 522 (1997).
- ¹²A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha *et al.*, *J. Phys. Chem. B* **102**, 3586 (1998).
- ¹³A. D. MacKerell, Jr., M. Feig, and C. L. Brooks, *J. Comput. Chem.* **25**, 1400 (2004).
- ¹⁴W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).
- ¹⁵M. L. Quillin, P. T. Wingfield, and B. W. Matthews, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19749 (2006).
- ¹⁶B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, *J. Chem. Theory Comput.* **4**, 435 (2008).
- ¹⁷M. Cui, M. Mezei, and R. Osman, *Protein. Eng. Des. Sel.* **21**, 729 (2008).
- ¹⁸D. J. Adams, *Mol. Phys.* **29**, 307 (1975).
- ¹⁹H.-J. Woo, A. R. Dinner, and B. Roux, *J. Chem. Phys.* **121**, 6392 (2004).
- ²⁰P. Mark and L. Nilsson, *J. Phys. Chem. A* **105**, 9954 (2001).
- ²¹M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, *Biophys. J.* **72**, 1047 (1997).
- ²²D. Hamelberg and J. A. McCammon, *J. Am. Chem. Soc.* **126**, 7683 (2004).
- ²³M. K. Gilson and K. K. Irikura, *J. Phys. Chem. B* **117**, 3061 (2013).
- ²⁴H. Yin, G. Feng, G. M. Clore, G. Hummer, and J. C. Rasaiah, *J. Phys. Chem. B* **114**, 16290 (2010).
- ²⁵J. L. Kulp III, J. L. Kulp, Jr., D. L. Pompliano, and F. Guarnieri, *J. Am. Chem. Soc.* **133**, 10740 (2011).
- ²⁶M. Clark, S. Meshkat, G. T. Talbot, P. Carnevali, and J. S. Wiseman, *J. Chem. Inf. Model.* **49**, 1901 (2009).
- ²⁷See supplementary material at <http://dx.doi.org/10.1063/1.4817344> for the figures and tables referred to in the paper.